Subgradient Regularization: A Descent-Oriented Subgradient Method for Nonsmooth Optimization

Hanyang Li [*]   Ying Cui [†]

May 11, 2025

Abstract

In nonsmooth optimization, a negative subgradient is not necessarily a descent direction, making the design of convergent descent methods based on zeroth-order and first-order information a challenging task. The well-studied bundle methods and gradient sampling algorithms construct descent directions by aggregating subgradients at nearby points in seemingly different ways, and are often complicated or lack deterministic guarantees. In this work, we identify a unifying principle behind these approaches, and develop a general framework of descent methods under the abstract principle that provably converge to stationary points. Within this framework, we introduce a simple yet effective technique, called subgradient regularization, to generate stable descent directions for a broad class of nonsmooth marginal functions, including finite maxima or minima of smooth functions. When applied to the composition of a convex function with a smooth map, the method naturally recovers the prox-linear method and, as a byproduct, provides a new dual interpretation of this classical algorithm. Numerical experiments demonstrate the effectiveness of our methods on several challenging classes of nonsmooth optimization problems, including the minimization of Nesterov's nonsmooth Chebyshev-Rosenbrock function.

Keywords: Nonsmooth, Subdifferential, Stable Descent Directions, Linear Convergence.

1 Introduction

We consider the minimization of a nonconvex, nonsmooth and locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$. The Clarke subdifferential of f, denoted by ∂f, generalizes the concept of the subdifferential in convex analysis and reduces to the gradient when f is smooth. A point $\bar{x}$ is said to be a stationary point of f if $0 \in \partial f(\bar{x})$. In nonsmooth optimization, it is well known that an arbitrary negative subgradient $v \in \partial f(x)$ may fail to yield a descent direction at x. While a descent direction can be obtained by using the minimal norm subgradient

$$g_x \triangleq - \operatorname*{argmin}_{v \in \partial f(x)} \|v\|, \tag{1}$$

---

[*]Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA (hanyang_li@berkeley.edu)

[†]Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA (yingcui@berkeley.edu)

the resulting steepest descent method often suffers from zigzagging behavior and may even converge to non-stationary points; see, e.g., [26, Section 2.2]. These issues arise from the discontinuity of the mapping $x \mapsto g_x$, posing a fundamental challenge in the design of efficient descent methods for nonsmooth minimization.

To develop fast descent methods with convergence guarantees, significant efforts have been devoted to constructing stable descent directions using enlarged subdifferentials, defined in an extended space $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$. The auxiliary parameter ϵ plays a crucial role in ensuring the continuity of the subdifferential mapping. Existing algorithms can be broadly classified into two categories based on how the enlarged subdifferentials are constructed: Goldstein-type methods and bundle-type methods.

Goldstein-type methods. The seminal work of Goldstein [22] introduced the concept of the *Goldstein ϵ-subdifferential*:

$$\partial_\epsilon^G f(x) \triangleq \operatorname{conv}\left(\bigcup \{ \partial f(z) \mid z \in \overline{\mathbb{B}}_\epsilon(x) \} \right), \quad \epsilon > 0,$$

where $\overline{\mathbb{B}}_\epsilon(x) \triangleq \{ y \in \mathbb{R}^n \mid \|y - x\| \leq \epsilon \}$ and $\operatorname{conv}(S)$ represents the convex hull of a set S. Goldstein's subgradient method updates the iterates as

$$x^{k+1} = x^k - \epsilon_k \frac{g^k}{\|g^k\|} \quad \text{with} \quad g^k = \operatorname*{argmin}_{v \in \partial_{\epsilon_k}^G f(x^k)} \|v\| (\neq 0), \qquad \text{(Goldstein)}$$

which guarantees descent in the objective: $f(x^{k+1}) \leq f(x^k) - \epsilon_k \|g^k\|$. With properly chosen stepsizes $\{\epsilon_k\}$, the sequence $\{x^k\}$ converges subsequentially to stationary points of f.

However, computing the direction in (Goldstein) is often challenging in practice, as it requires access to all subgradients within a neighborhood of x^k. The gradient sampling algorithm [7, 8, 5] circumvents this by approximating the Goldstein ϵ-subdifferential using randomly sampled points in $\overline{\mathbb{B}}_\epsilon(x)$. Recently, several variants of Goldstein-type methods have been proposed in [47, 13, 44], focusing on efficient random approximation of the direction in (Goldstein) to find a point x satisfying $\operatorname{dist}(0, \partial_\epsilon^G f(x)) \leq \delta$ with complexity guarantees. The recent work [14] proposed a randomized scheme of the Goldstein method that is locally nearly linearly convergent by exploring underlying smooth substructures. Without considering the complexity inherent in the minimal norm oracle for finding Goldstein directions, [28] identifies abstract properties that facilitate the near-linear convergence in Goldstein-type methods. Those properties involve a new modulus measuring the growth of a function through the Goldstein ϵ-subdifferential.

Bundle-type methods. Bundle methods, originating from the works [46, 29], are based on a different notion of enlarged subdifferentials, particularly suited for convex nonsmooth functions. Specifically, for a convex function f, the *ϵ-subdifferential* at x is defined as

$$\partial_\epsilon f(x) \triangleq \left\{ v \mid f(z) \geq f(x) + v^\top (z - x) - \epsilon \text{ for all } z \right\}, \qquad \epsilon > 0.$$

This definition allows subgradients computed at one point to be effectively reused at another. In particular, if $g \in \partial f(\hat{x})$ at some $\hat{x}$, then g can serve as an inexact subgradient at another point x, satisfying $g \in \partial_\epsilon f(x)$ with $\epsilon = f(x) - f(\hat{x}) - g^\top(x - \hat{x})$. This property, known as the transportation formula [26, Chapter XI, Section 4.2], allows one to iteratively build up a convex polyhedron $S_k \subset \partial_{\epsilon_k} f(x^k)$ by aggregating objective values and subgradients from past iterations. In the dual

2

interpretation of bundle methods [30, Lemma 3.1 and equation (3.20)], the so-called serious step with the stepsize $\mu_k > 0$ corresponds to the update

$$x^{k+1} = x^k - \mu_k \tilde{g}^k \quad \text{with} \quad \tilde{g}^k = \operatorname*{argmin}_{v \in S_k} \|v\| (\neq 0), \qquad \text{(Bundle)}$$

ensuring a decrease in the objective value if S_k sufficiently approximates $\partial_{\epsilon_k} f(x^k)$. Between serious steps, null steps are used to refine the approximation S_k by incorporating more objective values and subgradient information at trial points. Bundle methods have also been extended to certain classes of nonconvex functions; see, for example, [36, 24, 15].

In terms of the convergence rate, the pioneering work [41] shows that serious steps converge at a linear rate for convex nonsmooth functions whose subdifferential is metrically regular. This analysis has been recently extended to weakly convex functions [2]. For convex functions admitting $\mathcal{VU}$-structures [37], a variant of bundle methods that incorporates second-order information relative to a smooth subspace enjoys superlinear convergence in terms of serious steps [38]. Accounting for both serious and null steps, [16] shows that the bundle method achieves linear convergence under the more restrictive sharp growth condition, alongside convergence rate analysis under various other smoothness and growth conditions.

Observe that the frameworks in (Goldstein) and (Bundle) differ primarily in how they enlarge the Clarke subdifferential $\partial f(x)$ to stabilize the steepest descent direction in (1): The Goldstein ϵ-subdifferential $\partial_\epsilon^G f(x)$ expands the set by incorporating subgradients from a neighborhood of x, whereas the ϵ-subdifferential $\partial_\epsilon f(x)$ is defined based on inexact first-order expansions at x. If f is convex and L-Lipschitz continuous, one can show that $\partial_\epsilon^G f(x) \subset \partial_{2\epsilon L} f(x)$, though the reverse inclusion does not generally hold.

1.1 Contributions

The aim of this paper is two-fold: to establish an abstract guiding principle for the design of convergent descent methods in nonsmooth optimization, and to construct descent directions for a broad subclass of nonsmooth functions from the angle of subgradient regularization.

Descent-oriented subdifferentials. Despite the success of Goldstein-type and Bundle-type methods, little research has been devoted to identifying abstract properties of enlarged subdifferentials, defined jointly over (x, ϵ), that can be systematically employed to design descent algorithms with convergence guarantees. Both (Goldstein) and (Bundle) follow a common two-step procedure to construct directions: first (approximately) expand subdifferentials via an auxiliary parameter ϵ, then select the minimal norm element. This process yields a descent direction while avoiding the discontinuity issue of the steepest descent direction (1).

In the first part of our work, we formalize this idea through the concept of *descent-oriented subdifferentials*, a guiding principle for constructing stable descent directions in nonsmooth optimization. Notably, directions in (Goldstein) and (Bundle) can be interpreted as particular realizations within our framework. Building on this concept, we further develop a general descent method (Algorithm 1) that ensures subsequential convergence to stationary points.

Beyond offering a principled theoretical framework, descent-oriented subdifferentials also open the door to develop new algorithms for structured nonsmooth problems, where additional problem structure can be leveraged. For example, it is known that to minimize a nonsmooth composite

3

function $f(x) \triangleq h(c(x))$, where $c : \mathbb{R}^n \to \mathbb{R}^m$ is a C^1-smooth mapping and $h : \mathbb{R}^m \to \mathbb{R}$ is a convex function, the prox-linear method [20, 4, 6, 33, 17] guarantees descent via linearizing only the inner smooth map c at each iteration while keeping the outer convex function h unchanged. This naturally motivates our next contribution.

Descent directions via subgradient regularization. For a broad class of nonsmooth marginal functions, including finite maxima or minima of smooth functions, we consider a simple yet powerful technique for constructing descent-oriented subdifferentials via *subgradient regularization*. When Algorithm 1 is equipped with this construction, we refer to the resulting method as *Subgradient-Regularized Descent* (SRDescent). To illustrate the idea, consider $f(x) = \max\{f_1(x), f_2(x)\}$ for some smooth functions f_1 and f_2. One has

$$\partial f(x) = \left\{ \bar{y}_1 \nabla f_1(x) + \bar{y}_2 \nabla f_2(x) \,\middle|\, \bar{y} \in \operatorname*{argmax}_{y \in \Delta^2} \left[y_1 f_1(x) + y_2 f_2(x) \right] \right\},$$

where $\Delta^d \triangleq \left\{ y \in \mathbb{R}^d \mid \sum_{i=1}^d y_i = 1, y \geq 0 \right\}$. Our subgradient regularization approach constructs a descent-oriented subdifferential as follows:

$$\left\{ \bar{y}_1^\epsilon \nabla f_1(x) + \bar{y}_2^\epsilon \nabla f_2(x) \,\middle|\, \bar{y}^\epsilon \in \operatorname*{argmax}_{y \in \Delta^2} \left[\begin{array}{c} y_1 f_1(x) + y_2 f_2(x) \\ -\frac{\epsilon}{2} \underbrace{\|y_1 \nabla f_1(x) + y_2 \nabla f_2(x)\|^2}_{\text{subgradient regularization}} \end{array} \right] \right\},$$

where the regularization term penalizes the norm of a perturbed subgradient. When $\epsilon > 0$, the above set is outer semicontinuous in x; and when $\epsilon \downarrow 0$, it converges to the minimal norm subgradient at x. SRDescent then proceeds by moving along the negative of the regularized subgradient. Unlike the directions used in (Goldstein) and (Bundle) that require information from a neighborhood of the current iterate, this approach generates descent directions using objective values and subgradients at a single reference point. Figure 1 visualizes the directions generated by gradient sampling, a bundle-type method, and SRDescent. Figure 2 previews the empirical performance of these approaches.

Interestingly, when f is a composition of a convex function with a smooth map, the subgradient regularization reduces to the prox-linear update (see Proposition 7 in Section 5.1). As a byproduct, this connection offers a new dual perspective on the prox-linear method, previously understood primarily from a model-based viewpoint [17].

Motivated by this relationship, we further refine Algorithm 1 by introducing a dynamic adjustment of the regularization parameter ϵ. We show that this adaptive variant preserves subsequential convergence to stationary points for locally Lipschitz continuous functions. Moreover, similar to the prox-linear method, under regularity conditions, the algorithm achieves local linear convergence when applied to the composition of a convex function with a smooth map.

1.2 Notation and preliminaries

For any vector $x \in \mathbb{R}^n$, we use $\|x\|$ to denote its Euclidean norm. The open ball in $\mathbb{R}^n$ centered at $\bar{x} \in \mathbb{R}^n$ with radius $\delta > 0$ is denoted by $\mathbb{B}_\delta(\bar{x}) \triangleq \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| < \delta\}$, and its closure is $\overline{\mathbb{B}}_\delta(\bar{x})$. For nonempty closed sets $A, B \subset \mathbb{R}^n$, the one-sided deviation of A from B is defined as $\mathbb{D}(A, B) \triangleq \sup_{x \in A} \operatorname{dist}(x, B)$, where $\operatorname{dist}(x, B) \triangleq \inf_{y \in B} \|y - x\|$. The Hausdorff distance is defined as $\mathbb{H}(A, B) \triangleq \max\{\mathbb{D}(A, B), \mathbb{D}(B, A)\}$. We denote $\operatorname{cl}(A)$ and $\operatorname{conv}(A)$ as the closure and convex hull
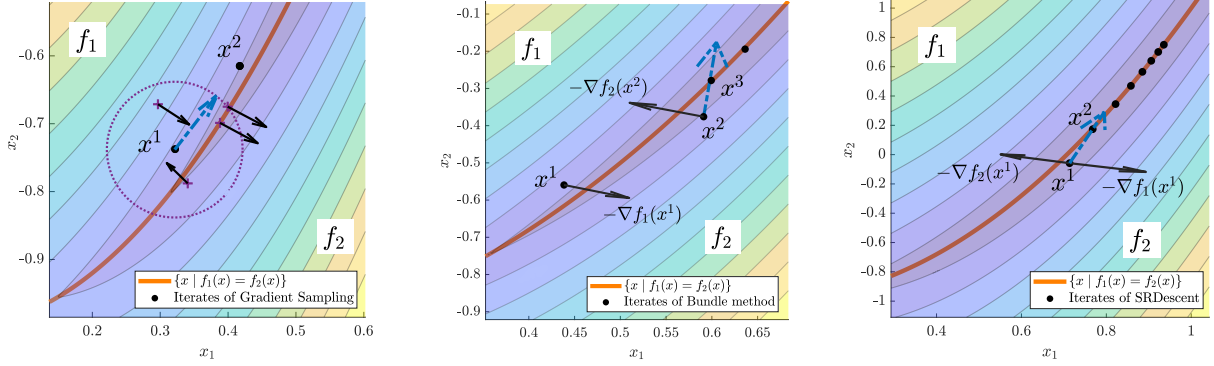
Figure 1: Contour plots and descent directions of Nesterov's nonsmooth Chebyshev-Rosenbrock function $f(x) = \frac{1}{4}(x_1 - 1)^2 + |x_2 - 2(x_1)^2 + 1|$, which can be expressed as the maximum of two smooth functions f_1 and f_2. In each plot, dashed arrows show the descent direction computed by the respective methods, while different markers indicate the points used to construct these directions. **Left**: In gradient sampling algorithm, four nearby sample points (crosses) around x^1 are used to compute a minimal norm convex combination of gradients. **Middle**: In bundle method, gradients from previous iterates $\nabla f_1(x^1)$ and $\nabla f_2(x^2)$ are aggregated to form a descent direction at x^2. **Right**: SRDescent directly combines gradients $\nabla f_1(x^1)$ and $\nabla f_2(x^1)$ and their function values at the current iterate x^1.
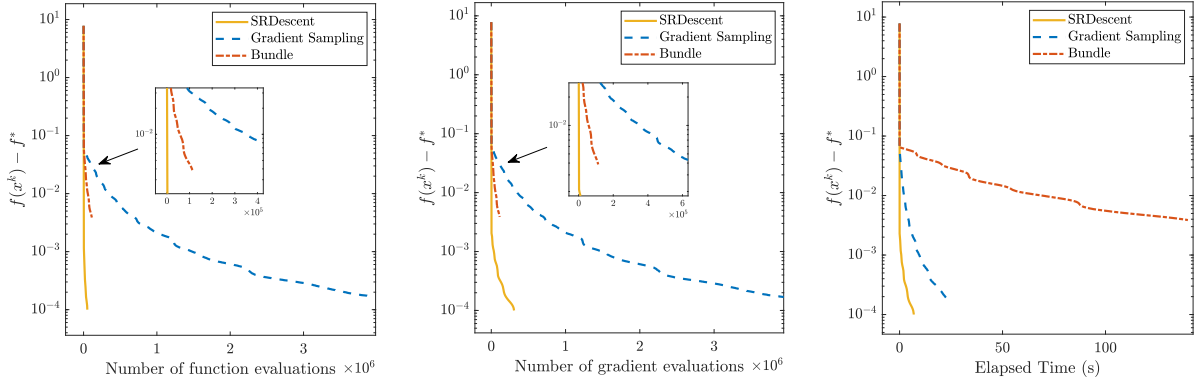

Figure 2: Performance of SRDescent, gradient sampling [8], and a bundle method specially designed for difference-of-convex functions [15] on Nesterov's nonsmooth Chebyshev-Rosenbrock function $f(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{6} |x_{i+1} - 2(x_i)^2 + 1|$. The gradient sampling algorithm terminates due to the failure of line search.

of a set A, respectively. The set of nonnegative integers is $\mathbb{N}$. We write $\mathbb{N}_\infty^\sharp \triangleq \{N \subset \mathbb{N} \mid N \text{ infinite}\}$ and $\mathbb{N}_\infty \triangleq \{N \subset \mathbb{N} \mid \mathbb{N} \setminus N \text{ finite}\}$. Notation $\{t^k\}$ is used to simplify the expression of any sequence $\{t^k\}_{k \in \mathbb{N}}$, where the elements can be points or sets. The convergence of a sequence $\{t^k\}$ and a subsequence $\{t^k\}_{k \in N}$ to t are written as $t^k \to t$ and $t^k \to_N t$. The Clarke directional derivative of a locally Lipschitz function f at x in the direction d is $f^\circ(x; d) = \limsup_{y \to x,\, t \downarrow 0} \frac{f(y + td) - f(y)}{t}$, and recall [11, Proposition 2.1.2] that $f^\circ(x; d) = \text{maximum}\{d^\top v \mid v \in \partial f(x)\}$. For a sequence of sets $\{C^k\}$, the outer limit and inner limit are defined as

$$
\begin{aligned}
\limsup_{k \to +\infty} C^k &\triangleq \quad \{u \mid \exists N \in \mathbb{N}_\infty^\sharp,\, u^k \to_N u \text{ with } u^k \in C^k\}, \\
\liminf_{k \to +\infty} C^k &\triangleq \quad \{u \mid \exists N \in \mathbb{N}_\infty,\, u^k \to_N u \text{ with } u^k \in C^k\}.
\end{aligned}
$$

5

When the outer and inner limits coincide, the sequence $\{C^k\}$ is said to converge, and the set limit is denoted by $\lim_{k\to+\infty} C^k$. For a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, the outer and inner limits of S at $\bar{x}$ are

$$\limsup_{x\to\bar{x}} S(x) \triangleq \bigcup_{x^k\to\bar{x}} \limsup_{k\to+\infty} S(x^k), \qquad \liminf_{x\to\bar{x}} S(x) \triangleq \bigcap_{x^k\to\bar{x}} \liminf_{k\to+\infty} S(x^k).$$

The mapping S is *outer semicontinuous (osc)* at $\bar{x} \in \mathbb{R}^n$ if $\limsup_{x\to\bar{x}} S(x) \subset S(\bar{x})$, equivalently, if $\limsup_{x\to\bar{x}} S(x) = S(\bar{x})$. It is *continuous* at $\bar{x}$ if $S(\bar{x}) = \limsup_{x\to\bar{x}} S(x) = \liminf_{x\to\bar{x}} S(x)$. The mapping S is *locally bounded* at $\bar{x} \in \mathbb{R}^n$ if there exists a neighborhood V of $\bar{x}$ such that $\bigcup\{S(x) \mid x \in V\}$ is bounded.

The trace of a square matrix X is denoted by $\mathrm{Tr}(X)$. Let $\mathcal{S}_m$ be the space of $m \times m$ symmetric matrices and define the Frobenius scalar product of $X, Y \in \mathcal{S}_m$ as $\langle X, Y \rangle = \mathrm{Tr}(XY)$.

1.3 Outline of the paper

The rest of the paper is organized as follows. Section 2 introduces the abstract concept of descent-oriented subdifferentials and establishes its key properties. Section 3 develops a general framework of descent methods based on the proposed descent-oriented subdifferentials and proves its asymptotic convergence. In Section 4, we present a new construction of descent-oriented subdifferentials for a class of marginal functions using subgradient regularization. In Section 5, we propose an adaptive variant of the previous algorithm, which reduces to the prox-linear method when minimizing the composition of a convex function with a smooth map, and establish its local linear convergence under regularity conditions. Section 6 evaluates the performance of the proposed algorithms on several challenging nonsmooth optimization problems. The paper ends with a concluding section.

2 Descent-Oriented Subdifferentials

2.1 Definition and existence of descent directions

We begin by introducing the concept of descent-oriented subdifferentials, which not only unifies the ideas behind Goldstein-type and bundle-type methods, but also serves as a guiding principle for designing new descent algorithms with convergence guarantees in nonsmooth optimization.

Definition 1 (Descent-oriented subdifferentials). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. A set-valued map $G : \mathbb{R}^n \times (0, +\infty) \rightrightarrows \mathbb{R}^n$ is called a *descent-oriented subdifferential* for f if, for each fixed $\epsilon > 0$, the mapping $G(\cdot, \epsilon)$ is closed-valued and locally bounded, and the following two properties hold for all $\bar{x}$:

$$\limsup_{\epsilon\downarrow 0,\, x\to\bar{x}} G(x, \epsilon) \subset \partial f(\bar{x}), \tag{G1}$$

$$\lim_{\epsilon\downarrow 0} \left(\limsup_{x\to\bar{x}} G(x, \epsilon) \right) = \operatorname*{argmin}_{v\in\partial f(\bar{x})} \|v\|. \tag{G2}$$

Remark 1 (Scaling). Suppose G is a descent-oriented subdifferential for f. Then, for any scalar α, the mapping αG is a descent-oriented subdifferential for αf. This result follows directly from the fact that $\partial(\alpha f)(x) = \alpha \, \partial f(x)$ (see [10]). In particular, $(-G)$ is a descent-oriented subdifferential for $(-f)$.

Below we discuss the two required conditions of descent-oriented subdifferentials. Property (G1) ensures that under joint perturbation in (x, ϵ), the set $G(x, \epsilon)$ does not asymptotically introduce information beyond the Clarke subdifferential. This requirement is similar to the one-sided "gradient consistency" of the smoothing method [9], in which $G(x, \epsilon)$ corresponds to the gradient of a smooth approximation for f and plays an important role in establishing subsequential convergence to stationary points of smoothing methods. On the other hand, property (G2) restricts the sequential limit of $G(x, \epsilon)$ in x and $\epsilon \downarrow 0$ to be the minimal norm subgradient. In particular, it implies that $\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$, which guarantees the existence of descent directions for sufficiently small ϵ, as we will show in Proposition 2. Furthermore, property (G2) is also closely related to the continuity of $G(x, \epsilon)$ with respect to x for a fixed $\epsilon > 0$. A sufficient condition for (G2) is that, for all $\bar{x}$, we have:

$$\left[\limsup_{x \to \bar{x}} G(x, \epsilon) = G(\bar{x}, \epsilon), \quad \forall \epsilon > 0 \right] \quad \text{and} \quad \lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname*{argmin}_{v \in \partial f(\bar{x})} \|v\|. \tag{G2$'$}$$

The former limit requires $G(\cdot, \epsilon)$ to be osc for every fixed $\epsilon > 0$. When $G(\cdot, \epsilon)$ is single-valued, outer semicontinuity together with local boundedness of $G(\cdot, \epsilon)$, as specified in Definition 1, is equivalent to continuity of $G(\cdot, \epsilon)$ [43, Corollary 5.20].

Remark 2. The minimal norm subgradient may not be a descent-oriented subdifferential, since property (G2) is not satisfied. Consider for example $f(x) = \max\{-100, 2x_1+3x_2, -2x_1+3x_2, 5x_1+ 2x_2, -5x_1 + 2x_2\}$, which was studied in [26, Section 2.2] to demonstrate why the steepest descent method may fail to minimize nonsmooth functions. At $\bar{x} = (0, 0)$, we have $\operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\| = \{(0, 2)\}$. Along the path $x(t) = (t, 3t)$ with $t > 0$, $\operatorname{argmin}_{v \in \partial f(x(t))} \|v\| = \{(2, 3)\}$. Thus, the minimal norm subgradient violates (G2) because

$$\limsup_{t \to 0} \operatorname*{argmin}_{v \in \partial f(x(t))} \|v\| = \{(2, 3)\} \neq \{(0, 2)\} = \operatorname*{argmin}_{v \in \partial f(\bar{x})} \|v\|.$$

We now characterize stationary points using descent-oriented subdifferentials.

Proposition 1 (Characterization of stationary points). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function and G be a descent-oriented subdifferential for f. Then a point $\bar{x}$ is a stationary point of f if and only if there exist sequences $\{\epsilon_k\} \downarrow 0$, $\{v^k\} \to 0$, and $\{x^k\} \to \bar{x}$ such that $v^k \in G(x^k, \epsilon_k)$.

Proof. The "if" direction follows immediately from property (G1). For the "only if" direction, since $\bar{x}$ is a stationary point, by property (G2), we have

$$\lim_{\epsilon \downarrow 0} \left(\limsup_{x \to \bar{x}} G(x, \epsilon) \right) = \operatorname*{argmin}_{v \in \partial f(\bar{x})} \|v\| = \{0\}.$$

Hence, we can find $\{\epsilon_k\} \downarrow 0$ and $\{x^k\} \to \bar{x}$ such that $0 \in \limsup_{k \to +\infty} G(x^k, \epsilon_k)$, completing the proof. $\square$

Finally, we establish the existence of descent directions at non-stationary points using (G2). This proposition will be central to the convergence analysis of our proposed algorithms.

7

Proposition 2 (Existence of descent directions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function and G be a descent-oriented subdifferential for f. For any point $\bar{x}$ satisfying $0 \notin \partial f(\bar{x})$ and any constant $\alpha \in (0, 1)$, there exist positive scalars $\epsilon_{\bar{x}}$ and $\eta_{\bar{x}}$ such that for all $\epsilon \in (0, \epsilon_{\bar{x}})$ and $\eta \in (0, \eta_{\bar{x}})$, $0 \notin \limsup_{x \to \bar{x}} G(x, \epsilon)$ and the following descent inequality holds:

$$f(\bar{x} - \eta g) \leq f(\bar{x}) - \alpha \, \eta \|g\|^2, \qquad \forall \, g \in \limsup_{x \to \bar{x}} G(x, \epsilon). \tag{2}$$

Remark 3. Noting that $G(\bar{x}, \epsilon) \subset \limsup_{x \to \bar{x}} G(x, \epsilon)$, the descent inequality implies that for sufficiently small $\epsilon > 0$, any element of $-G(\bar{x}, \epsilon)$ is a descent direction.

Proof. The first statement $0 \notin \limsup_{x \to \bar{x}} G(x, \epsilon)$ for all sufficiently small $\epsilon > 0$ is a direct consequence of property (G2) and our assumption $0 \notin \partial f(\bar{x})$. We only need to show (2). Suppose for contradiction that $0 \notin \partial f(\bar{x})$ but there exist $\{\epsilon_k\} \downarrow 0$, $\{\eta_k\} \downarrow 0$ and $g^k \in \limsup_{x \to \bar{x}} G(x, \epsilon_k)$ such that

$$f(\bar{x} - \eta_k g^k) > f(\bar{x}) - \alpha \, \eta_k \|g^k\|^2, \qquad \forall \, k \in \mathbb{N}. \tag{3}$$

Let $\bar{d} \triangleq \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\| \neq 0$. By property (G2), it holds that $g^k \to \bar{d}$. Next, consider the Clarke directional derivative of f at $\bar{x}$ in the direction $-\bar{d}$:

$$
\begin{aligned}
f^\circ(\bar{x}; -\bar{d}) \quad &= \limsup_{x \to \bar{x}, \, t \downarrow 0} \frac{f(x - t\bar{d}) - f(x)}{t} \geq \limsup_{d \to \bar{d}, \, t \downarrow 0} \frac{f(\bar{x} - td) - f(\bar{x})}{t} \\
&\geq \limsup_{k \to +\infty} \frac{f(\bar{x} - \eta_k g^k) - f(\bar{x})}{\eta_k} \overset{(3)}{\geq} -\alpha \left\| \bar{d} \right\|^2,
\end{aligned} \tag{4}
$$

where the first inequality follows from the local Lipschitz continuity of f. To proceed, we use the optimality condition of $\bar{d} = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|^2$ to obtain $(\bar{d} - u)^\top \bar{d} \leq 0$ for all $u \in \partial f(\bar{x})$. Thus, $f^\circ(\bar{x}; -\bar{d}) = \max_{u \in \partial f(\bar{x})} (-u^\top \bar{d})$ attains its maximum at $u = \bar{d}$. Hence, by (4), $-\|\bar{d}\|^2 = f^\circ(\bar{x}; -\bar{d}) \geq -\alpha \|\bar{d}\|^2$, which implies $\bar{d} = 0$ since $\alpha \in (0, 1)$. This contradicts the assumption $0 \notin \partial f(\bar{x})$. $\qquad \square$

2.2 Construction of descent-oriented subdifferentials

In the first part of this subsection, we illustrate how descent-oriented subdifferentials can be constructed from the minimal norm element of either Goldstein ϵ-subdifferential or the ϵ-subdifferential of convex functions. This perspective enables a unified interpretation of the iteration schemes in (Goldstein) and (Bundle) through the lens of descent-oriented subdifferentials.

Proposition 3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous. Then the single-valued map $G : \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}^n$ defined by $G(x, \epsilon) = \operatorname{argmin}_{v \in \partial_\epsilon^{\mathrm{G}} f(x)} \|v\|$ is a descent-oriented subdifferential for f.

Proof. For any fixed $\epsilon > 0$, the set $\partial_\epsilon^{\mathrm{G}} f(x)$ is non-empty, closed and convex for all x. Thus, G is a single-valued map on $\mathbb{R}^n \times (0, +\infty)$, and $G(\cdot, \epsilon)$ is closed-valued. Since f is locally Lipschitz continuous, the set-valued map $\partial_\epsilon^{\mathrm{G}} f(\cdot)$ is locally bounded for any $\epsilon > 0$. Therefore, its projection map, $G(\cdot, \epsilon)$ is also locally bounded.

We then verify property (G1). Take sequences $\{x^k\} \to \bar{x}$ and $\{\epsilon_k\} \downarrow 0$ such that $G(x^k, \epsilon_k) \to \bar{u}$. We will show that $\bar{u} \in \partial f(\bar{x})$. By noticing $G(x^k, \epsilon_k) \in \partial_{\epsilon_k}^G f(x^k)$ and using Carathéodory's Theorem (see, e.g. [42, Theorem 17.1]), there exist $\lambda_{k,j} \geq 0$ and $y^{k,j} \in \overline{\mathbb{B}}_{\epsilon_k}(x^k)$ for $j = 1, \cdots, n+1$ satisfying

$$G(x^k, \epsilon_k) = \sum_{j=1}^{n+1} \lambda_{k,j} \, g^{k,j} \quad \text{with} \quad g^{k,j} \in \partial f(y^{k,j}) \text{ and } \sum_{j=1}^{n+1} \lambda_{k,j} = 1.$$

For each j, observe that $\{\lambda_{k,j}\}_{k \in \mathbb{N}} \subset [0,1]$, and $\{g^{k,j}\}_{k \in \mathbb{N}}$ is a bounded sequence due to the local Lipschitz continuity of f. By passing to a subsequence if necessary, we may assume that $\lim_{k \to +\infty} \lambda_{k,j} = \bar{\lambda}_j$ and $\lim_{k \to +\infty} g^{k,j} = \bar{g}^j$ for all j. We also have $\bar{g}^j \in \partial f(\bar{x})$ by the outer semicontinuity of ∂f. Thus, the limit point $\bar{u}$ of the sequence $\{G(x^k, \epsilon_k)\}$ admits the representation

$$\bar{u} = \sum_{j=1}^{n+1} \bar{\lambda}_j \, \bar{g}^j \in \text{conv}\{\partial f(\bar{x})\} = \partial f(\bar{x}), \quad \text{where} \sum_{j=1}^{n+1} \bar{\lambda}_j = 1 \text{ and } \bar{\lambda}_j \geq 0.$$

This establishes (G1).

Next, consider any sequence $\{\epsilon_k\} \downarrow 0$ and let $v^k \in \limsup_{x \to \bar{x}} G(x, \epsilon_k)$ for each $k \in \mathbb{N}$. To prove (G2), we need to show that $v^k \to \text{argmin}_{v \in \partial f(\bar{x})} \|v\|$. By definition, there exist sequences $\{x^{k,i}\}_{i \in \mathbb{N}} \to \bar{x}$ and $\{v^{k,i}\}_{i \in \mathbb{N}} \to v^k$ with $v^{k,i} = \text{argmin}_{v \in \partial_{\epsilon_k}^G f(x^{k,i})} \|v\|$ for all $i \in \mathbb{N}$. For each k, select a sufficiently large index i_k such that $\|x^{k,i_k} - \bar{x}\| \leq \epsilon_k$ and $\|v^{k,i_k} - v^k\| \leq 1/k$. Thus,

$$\partial f(\bar{x}) \subset \partial_{\epsilon_k}^G f(x^{k,i_k}) \subset \partial_{2\epsilon_k}^G f(\bar{x}), \qquad \forall \, k \in \mathbb{N}. \tag{5}$$

Since the sequence of sets $\{\partial_{2\epsilon_k}^G f(\bar{x})\}$ is monotone, it follows from [43, Exercise 4.3] that

$$\lim_{k \to +\infty} \partial_{2\epsilon_k}^G f(\bar{x}) = \bigcap_{k \in \mathbb{N}} \text{cl}\left(\partial_{2\epsilon_k}^G f(\bar{x})\right) = \bigcap_{k \in \mathbb{N}} \partial_{2\epsilon_k}^G f(\bar{x}) = \partial f(\bar{x}). \tag{6}$$

Combining (5) and (6), we obtain $\partial_{\epsilon_k}^G f(x^{k,i_k}) \to \partial f(\bar{x})$. By the continuity of the projection onto closed convex sets [43, Proposition 4.9], we conclude that

$$\lim_{k \to +\infty} v^{k,i_k} = \lim_{k \to +\infty} \underset{v \in \partial_{\epsilon_k}^G f(x^{k,i_k})}{\text{argmin}} \|v\| = \underset{v \in \partial f(\bar{x})}{\text{argmin}} \|v\|.$$

Since $\|v^{k,i_k} - v^k\| \leq 1/k$, we have $v^k \to \text{argmin}_{v \in \partial f(\bar{x})} \|v\|$. This proves (G2). $\qquad \square$

We next turn to the ϵ-subdifferential for convex functions. Asplund and Rockafellar [1] proved that the set-valued map $(x, \epsilon) \mapsto \partial_\epsilon f(x)$ is jointly continuous on $\mathbb{R}^n \times (0, +\infty)$ in terms of the Hausdorff distance. Thus, fixing $\epsilon > 0$, the mapping $x \mapsto \partial_\epsilon f(x)$ is continuous. We will use this fact below.

Proposition 4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the single-valued map $G : \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}^n$ defined by $G(x, \epsilon) = \text{argmin}_{v \in \partial_\epsilon f(x)} \|v\|$ is a descent-oriented subdifferential for f.

Proof. For any fixed $\epsilon > 0$, the set $\partial_\epsilon f(x)$ is closed convex for all $x \in \mathbb{R}^n$. Hence, G is a single-valued map, and $G(\cdot, \epsilon)$ is closed-valued. Also, for any $\epsilon > 0$, $G(\cdot, \epsilon)$ is locally bounded as the set-valued map $\partial_\epsilon f(\cdot)$ is locally bounded [1, Corollary 1]. To establish (G1), observe that for any $\bar{x}$,

$$\limsup_{\epsilon \downarrow 0, x \to \bar{x}} G(x, \epsilon) \subset \limsup_{\epsilon \downarrow 0, x \to \bar{x}} \partial_\epsilon f(x) \subset \partial f(\bar{x}),$$

9

where the second inclusion follows directly from the continuity of f and the definition of $\partial_\epsilon f(x)$. We show property (G2) by verifying the sufficient condition (G2′). Fix $\bar{x} \in \mathbb{R}^n$ and consider any sequence $\{\epsilon_k\} \downarrow 0$. Since $\{\partial_{\epsilon_k} f(\bar{x})\}$ is a monotonic sequence of sets, we apply [43, Exercise 4.3] to obtain the set convergence

$$\lim_{k \to +\infty} \partial_{\epsilon_k} f(\bar{x}) = \bigcap_{k \in \mathbb{N}} \mathrm{cl}\left(\partial_{\epsilon_k} f(\bar{x})\right) = \bigcap_{k \in \mathbb{N}} \partial_{\epsilon_k} f(\bar{x}) = \partial f(\bar{x}).$$

Consequently, it follows from [43, Proposition 4.9] that $\mathrm{argmin}_{v \in \partial_{\epsilon_k} f(\bar{x})} \|v\| \to \mathrm{argmin}_{v \in \partial f(\bar{x})} \|v\|$, establishing the second limit in (G2′).

To complete the proof, it remains to show that $G(\cdot, \epsilon)$ is osc for any fixed $\epsilon > 0$. By [1, Proposition 5], we have $\lim_{x \to \bar{x}} \mathbb{H}(\partial_\epsilon f(x), \partial_\epsilon f(\bar{x})) = 0$ for any fixed $\bar{x}$ and $\epsilon > 0$. Since $\partial_\epsilon f(\cdot)$ is locally bounded at $\bar{x}$, it follows from [43, Corollary 5.21] that this convergence in Hausdorff distance is equivalent to the continuity of the set-valued map $\partial_\epsilon f(\cdot)$ at $\bar{x}$. Thus, for any sequence $x^k \to \bar{x}$, we have $\partial_\epsilon f(x^k) \to \partial_\epsilon f(\bar{x})$ and the convergence of their minimal norm elements by using [43, Proposition 4.9]:

$$\lim_{k \to +\infty} G(x^k, \epsilon) = \lim_{k \to +\infty} \mathrm{argmin}_{v \in \partial_\epsilon f(x^k)} \|v\| = \mathrm{argmin}_{v \in \partial_\epsilon f(\bar{x})} \|v\|.$$

Thus, (G2′) hold, and the proof is completed. □

We conclude this section by presenting an alternative construction of descent-oriented subdifferentials based on the Moreau envelope. A function $f : \mathbb{R}^n \to \mathbb{R}$ is ρ-weakly convex if $f + \frac{1}{2}\rho\|\cdot\|^2$ is a convex function. Given $\epsilon > 0$, the Moreau envelope and proximal mapping of f are defined as $e_\epsilon f(x) \triangleq \inf_z \{f(z) + \frac{1}{2\epsilon}\|z - x\|^2\}$ and $P_\epsilon f(x) \triangleq \mathrm{argmin}_z \{f(z) + \frac{1}{2\epsilon}\|z - x\|^2\}$.

Proposition 5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be ρ-weakly convex and bounded from below. For all $\epsilon \in (0, \rho^{-1})$, the Moreau envelope $e_\epsilon f$ is continuously differentiable. Fix $\bar{\epsilon} \in (0, \rho^{-1})$. Then the single-valued map $G : \mathbb{R}^n \times (0, +\infty) \to \mathbb{R}^n$ defined by

$$G(x, \epsilon) = \begin{cases} \nabla e_\epsilon f(x) & \text{when } \epsilon \in (0, \bar{\epsilon}], \\ \nabla e_{\bar{\epsilon}} f(x) & \text{otherwise} \end{cases} \tag{7}$$

is a descent-oriented subdifferential for f.

Proof. Fix any $\bar{x} \in \mathbb{R}^n$. Since $f + \frac{1}{2\rho}\|\cdot\|^2$ is convex, standard results about the Moreau envelope and proximal mapping for convex functions (see, e.g., [43, Theorem 2.26]) imply that for all $\epsilon \in (0, \rho^{-1})$, $P_\epsilon f$ is single-valued and continuous, and $e_\epsilon f$ is continuously differentiable with $\nabla e_\epsilon f(x) = \frac{1}{\epsilon}(x - P_\epsilon f(x))$. Thus, given $\bar{\epsilon} \in (0, \rho^{-1})$, the map G in (7) is well-defined. For each fixed $\epsilon \in (0, \bar{\epsilon}]$, $G(\cdot, \epsilon)$ is locally bounded due to the continuity of $P_\epsilon f$. By the optimality condition for the proximal mapping, $0 \in \partial f(P_\epsilon f(x)) + \frac{1}{\epsilon}(P_\epsilon f(x) - x)$, so $G(x, \epsilon) = \nabla e_\epsilon f(x) = \frac{1}{\epsilon}(x - P_\epsilon f(x)) \in \partial f(P_\epsilon f(x))$ for $\epsilon \in (0, \bar{\epsilon}]$. Since $f(P_\epsilon f(x)) + \frac{1}{2\epsilon}\|P_\epsilon f(x) - x\|^2 \le f(x)$ and f is bounded from below, we obtain $\|P_\epsilon f(x) - x\| \le \sqrt{2\epsilon(f(x) - \inf f)}$, implying $\lim_{x \to \bar{x}, \epsilon \downarrow 0} P_\epsilon f(x) = \bar{x}$. By the outer semicontinuity of ∂f,

$$\limsup_{x \to \bar{x}, \epsilon \downarrow 0} G(x, \epsilon) \subset \limsup_{x \to \bar{x}, \epsilon \downarrow 0} \partial f(P_\epsilon f(x)) \subset \partial f(\bar{x}),$$

establishing property (G1). To proceed, we aim to verify (G2′). The first condition $\limsup_{x \to \bar{x}} G(x, \epsilon) = G(\bar{x}, \epsilon)$ for $\epsilon > 0$ is a direct consequence of the continuity of $P_\epsilon f$. It remains to show that

10

$\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$. Let $\{\epsilon_k\} \downarrow 0$ with $\epsilon_k \in (0, \min\{\bar{\epsilon}, 1/(2\rho)\})$. We first show that the sequence $\{G(\bar{x}, \epsilon_k)\}$ is bounded. Since $f + \frac{1}{2\epsilon_k}\| \cdot -\bar{x}\|^2$ is $(1/\epsilon_k - \rho)$-strongly convex, its minimizer $P_{\epsilon_k} f(\bar{x})$ satisfies

$$f(\bar{x}) \geq \left(f(P_{\epsilon_k} f(\bar{x})) + \frac{1}{2\epsilon_k}\|\bar{x} - P_{\epsilon_k} f(\bar{x})\|^2 \right) + \frac{1}{2}\left(\frac{1}{\epsilon_k} - \rho \right)\|\bar{x} - P_{\epsilon_k} f(\bar{x})\|^2. \tag{8}$$

Additionally, since $f + \frac{1}{2}\rho\|\bar{x} - \cdot\|^2$ is convex, for any $v \in \partial f(\bar{x})$,

$$f(P_{\epsilon_k} f(\bar{x})) + \frac{1}{2}\rho\|\bar{x} - P_{\epsilon_k} f(\bar{x})\|^2 \geq f(\bar{x}) + v^\top \left(P_{\epsilon_k} f(\bar{x}) - \bar{x} \right). \tag{9}$$

Combining (8) and (9), and using the inequality $\frac{1}{2}\left(\epsilon_k\|v\|^2 + \frac{1}{\epsilon_k}\|P_{\epsilon_k} f(\bar{x}) - \bar{x}\|^2 \right) \geq -v^\top \left(P_{\epsilon_k} f(\bar{x}) - \bar{x} \right)$, we obtain that $\|v\|^2 \geq (1 - 2\epsilon_k\rho)\|P_{\epsilon_k} f(\bar{x}) - \bar{x}\|^2/\epsilon_k^2$ for any $v \in \partial f(\bar{x})$. As $\epsilon_k < \min\{\bar{\epsilon}, 1/(2\rho)\}$ and $\epsilon_k \downarrow 0$, it follows that $\|G(\bar{x}, \epsilon_k)\|^2 = \|P_{\epsilon_k} f(\bar{x}) - \bar{x}\|^2/\epsilon_k^2 \leq \inf_{v \in \partial f(\bar{x})} \|v\|^2/(1 - 2\epsilon_k\rho)$. Hence, $\{G(\bar{x}, \epsilon_k)\}$ is bounded and $\limsup_{k \to +\infty} \|G(\bar{x}, \epsilon_k)\| \leq \inf_{v \in \partial f(\bar{x})} \|v\|$.

Next, recall that $\lim_{x \to \bar{x}, \epsilon \downarrow 0} P_\epsilon f(x) = \bar{x}$ and, thus, $P_{\epsilon_k} f(\bar{x}) \to \bar{x}$. By the outer semicontinuity of ∂f and the fact that $G(\bar{x}, \epsilon_k) \in \partial f(P_{\epsilon_k} f(\bar{x}))$, every accumulation point of $\{G(\bar{x}, \epsilon_k)\}$ must belong to $\partial f(\bar{x})$. Moreover, the previous bound $\limsup_{k \to +\infty} \|G(\bar{x}, \epsilon_k)\| \leq \inf_{v \in \partial f(\bar{x})} \|v\|$ ensures that the only possible accumulation point is $\operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$. Therefore, $G(\bar{x}, \epsilon_k) \to \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$. This completes the proof. $\qquad \square$

3 A Descent-Oriented Subgradient Method

Using the descent-oriented subdifferential as a first-order oracle, this section presents a general framework of descent algorithms for minimizing a locally Lipschitz continuous function f. Throughout this section, we assume that a descent-oriented subdifferential $G : \mathbb{R}^n \times (0, +\infty) \rightrightarrows \mathbb{R}^n$ for f is given. The construction of such a map for structured nonsmooth functions will be discussed in the next section.

It follows from Proposition 2 that at any non-stationary point, descent in the objective can be achieved by selecting a sufficiently small parameter ϵ along with a proper stepsize η. Our proposed algorithm progressively reduces ϵ based on the norm of the descent-oriented subgradient at the current iterate, and determines η at a given ϵ via a line search procedure. In general, a diminishing sequence of ϵ is needed to ensure subsequential convergence to stationary points, since property (G1) only forces $G(x, \epsilon)$ approaches elements of $\partial f(x)$ when $\epsilon \downarrow 0$. For structured nonsmooth problems, we further propose an alternative strategy for adjusting ϵ in Section 5, which does not necessarily require ϵ to vanish asymptotically, yet still guarantees asymptotic convergence to stationarity and achieves local linear convergence. We now formally present the descent-oriented subgradient method in Algorithm 1.

Below we establish the well-definedness and convergence of Algorithm 1.

Proposition 6 (Finite termination of the inner loop i in Algorithm 1). Let $k \in \mathbb{N}$ and fix any optimality tolerances $\epsilon_{\text{tol}}, \nu_{\text{tol}} \geq 0$. If $0 \notin \partial f(x^k)$, then the inner loop indexed by i terminates in finitely many iterations, i.e., $i_k < +\infty$.

Proof. By Proposition 2, if $0 \notin \partial f(x^k)$, then there exist constants $\bar{\epsilon}, \bar{\eta} > 0$ such that for all $\epsilon \in (0, \bar{\epsilon})$ and $\eta \in (0, \bar{\eta})$, the following inequality holds:

$$f(x^k - \eta g^k) \leq f(x^k) - \alpha \eta \|g^k\|^2, \qquad \forall g^k \in G(x^k, \epsilon).$$

11

Algorithm 1 A basic descent-oriented subgradient method

1: **Input:** starting point x_0, regularization parameter $\epsilon_{0,0} > 0$, optimality tolerances $\epsilon_{\text{tol}}, \nu_{\text{tol}} \geq 0$, initial stationary target ν_0, reduction factors $\theta_\epsilon, \theta_\nu \in (0,1)$, and the Armijo parameter $\alpha \in (0,1)$

2: **for** $k = 0, 1, \cdots$ **do**
3:     **for** $i = 0, 1, \cdots$ **do**
4:         Set $\epsilon_{k,i} = \epsilon_{k,0} \, 2^{-i}$
5:         Update direction $g^{k,i} \in G(x^k, \epsilon_{k,i})$
6:         **if** $\epsilon_{k,i} \leq \epsilon_{\text{tol}}$ and $\|g^{k,i}\| \leq \nu_{\text{tol}}$ **then**
7:            Set $i_k = i$ and **return** x^k ▷ *approximate stationary*
8:         **for** $j = 0, 1, \cdots, i$ **do**
9:            Set $\eta = \epsilon_{k,0} \, 2^{-j}$ ▷ *line search*
10:           **if** $f(x^k - \eta \, g^{k,i}) \leq f(x^k) - \alpha \, \eta \|g^{k,i}\|^2$ **then**
11:              Set $i_k = i$ and go to line 12
12:     Update $x^{k+1} = x^k - \eta_k \, g^{k,i}$ with $\eta_k = \eta$
13:     **if** $\|g^{k,i_k}\| \leq \nu_k$ **then**
14:         Set $\nu_{k+1} = \theta_\nu \cdot \nu_k$ and $\epsilon_{k+1,0} = \theta_\epsilon \cdot \epsilon_{k,0}$
15:     **else**
16:         Set $\nu_{k+1} = \nu_k$ and $\epsilon_{k+1,0} = \epsilon_{k,0}$

Hence, either x^k is certified as an approximate stationary point with $\epsilon_{k,i} \leq \epsilon_{\text{tol}}$ and $\|g^{k,i}\| \leq \nu_{\text{tol}}$, or the descent condition in line 10 is eventually satisfied for sufficiently large i. The result follows. ☐

Remark 4. The above analysis implies that, for any $k \in \mathbb{N}$, the inner loop would still have the finite termination if we simply used the stepsize $\eta = \epsilon_{k,i}$, rather than performing a backtracking line search over $\{\epsilon_{k,0}, \epsilon_{k,0}/2, \ldots, \epsilon_{k,i}\}$. We adopt the line search strategy in Algorithm 1 to enhance its practical performance.

The following result was originally established by Kiwiel [27, Theorem 3.3] in the convergence analysis of the gradient sampling algorithm. It turns out to be useful in a more general setting, and we restate it here by emphasizing key consequences of a descent inequality, without specifying the update rule for the sequence $\{x^k\}$ and the choice of directions $\{g^k\}$.

Lemma 1. Suppose that f is bounded from below, and there are sequences $\{x^k\}$ and $\{g^k\}$ satisfying the descent inequality

$$f(x^{k+1}) \leq f(x^k) - \alpha \|x^{k+1} - x^k\| \cdot \|g^k\| \quad \text{for some } \alpha \in (0,1).$$

If $\{x^k\}$ has an accumulation point $\bar{x}$, then one of the following holds:
(a) $\{x^k\}$ converges to $\bar{x}$
(b) $\liminf_{k \to +\infty} \max\{\|x^k - \bar{x}\|, \|g^k\|\} = 0$.

We prove the subsequential convergence by using Proposition 2 and Lemma 1.

Theorem 1 (Convergence of Algorithm 1)**.** Assume that f is bounded from below. Let $\{x^k\}$ be a sequence generated by Algorithm 1 with $\epsilon_{\text{tol}} = \nu_{\text{tol}} = 0$. Then, one of the following occurs:
(a) For some $k \in \mathbb{N}$, the inner loop does not terminate. Consequently, x^k is a stationary point.
(b) Any accumulation point of $\{x^k\}$ is a stationary point.

Proof. First, consider the case that the inner loop does not terminate for some $k \in \mathbb{N}$. By Proposition 6, $0 \in \partial f(x^k)$, so (a) holds. Now suppose the inner loop terminates for each k. By Proposition 6, there exists an index i_k for each k such that the descent inequality holds:

$$f(x^{k+1}) = f(x^k - \eta_k\, g^{k,i_k}) \leq f(x^k) - \alpha\, \eta_k \|g^{k,i_k}\|^2.$$

Since f is bounded from below, summing the descent inequality over k gives

$$\sum_{k=0}^{+\infty} \eta_k \|g^{k,i_k}\|^2 = \sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| \cdot \|g^{k,i_k}\| < +\infty. \tag{10}$$

Case 1: Suppose that there exist $\underline{k} \in \mathbb{N}$, $\underline{\nu} > 0$ and $\underline{\epsilon} > 0$ such that $\nu_k = \underline{\nu}$ and $\epsilon_{k,0} = \underline{\epsilon}$ for all $k \geq \underline{k}$. Then, $\|g^{k,i_k}\| > \underline{\nu}$ for all $k \geq \underline{k}$. By (10), it follows that $\eta_k \to 0$ and $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$. Since $\eta_k \geq \epsilon_{k,0}\, 2^{-i_k} \geq \underline{\epsilon}\, 2^{-i_k}$, the condition $\eta_k \to 0$ implies $i_k \to +\infty$. Additionally, from $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$, we deduce that $\{x^k\}$ converges to some point $\bar{x}$. If $0 \in \partial f(\bar{x})$, the proof is done; therefore, assume that $0 \notin \partial f(\bar{x})$.

For each fixed index $i \in \mathbb{N}$, since $i_k \to +\infty$, there exists $k_i > \underline{k}$ such that $i \in \{0, \cdots, i_k\}$ for all $k \geq k_i$. By definition of G, the set-valued map $G(\cdot, \underline{\epsilon}\, 2^{-i})$ is locally bounded at $\bar{x}$, so $\{g^{k,i}\}_{k \geq k_i}$ is a bounded sequence. Let $\{g^{k,i}\}_{k \in N_i}$ be a subsequence convergent to some point $\bar{d}^i$. Thus,

$$\bar{d}^i \in \limsup_{k \to +\infty} G(x^k, \epsilon_{k,i}) = \limsup_{k \to +\infty} G(x^k, \underline{\epsilon}\, 2^{-i}) \qquad \text{for all } i \in \mathbb{N}. \tag{11}$$

For each k with $i_k \geq 1$, due to the termination criterion of the inner loop, the descent inequality fails up to index $i = i_k$, i.e., for all $i \in \{0, \cdots, i_k - 1\}$,

$$f\big(x^k - \epsilon_{k,0}\, 2^{-j} \cdot g^{k,i}\big) > f(x^k) - \epsilon_{k,0}\, 2^{-j}\, \alpha \|g^{k,i}\|^2 \qquad \text{for all } j \in \{0, \cdots, i\}. \tag{12}$$

By choosing $j = i$, passing to the limit as $k(\in N_i) \to +\infty$ in (12), and using the continuity of f, we obtain

$$f\big(\bar{x} - \underline{\epsilon}\, 2^{-i} \cdot \bar{d}^i\big) \geq f(\bar{x}) - \underline{\epsilon}\, 2^{-i} \cdot \alpha \|\bar{d}^i\|^2 \qquad \text{for all } i \in \mathbb{N}. \tag{13}$$

Combining (13) and (11), we conclude that the descent inequality (2) fails to hold with $\epsilon = \eta = \underline{\epsilon}\, 2^{-i}$ for all $i \in \mathbb{N}$. This contradicts Proposition 2, which guarantees the existence of descent directions for sufficiently small ϵ and η. Consequently, in Case 1, $\{x^k\}$ converges to a point $\bar{x}$ satisfying $0 \in \partial f(\bar{x})$.

Case 2: Otherwise, the reduction in line 13 of Algorithm 1 occurs infinitely many times, and we have $(\nu_k, \epsilon_{k,0}) \downarrow 0$. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Since $\{x^k\}$ and $\{g^{k,i_k}\}$ satisfy the descent inequality

$$f(x^{k+1}) \leq f(x^k) - \alpha\, \eta_k \|g^{k,i_k}\|^2 = f(x^k) - \alpha \|x^{k+1} - x^k\| \cdot \|g^{k,i_k}\|,$$

we can apply Lemma 1 to obtain that either $x^k \to \bar{x}$, or $\liminf_{k \to +\infty} \max\{\|x^k - \bar{x}\|, \|g^{k,i_k}\|\} = 0$. Since $\nu_k \downarrow 0$, it always holds that $\liminf_{k \to +\infty} \|g^{k,i_k}\| = 0$. Therefore, in either scenario, we can find an index set $N \in \mathbb{N}_\infty^\sharp$ such that $\max\big\{\|x^k - \bar{x}\|, \|g^{k,i_k}\|\big\} \to_N 0$. As $\epsilon_{k,i_k} \leq \epsilon_{k,0} \downarrow 0$, by property (G1), we have

$$0 = \lim_{k(\in N) \to +\infty} g^{k,i_k} \in \left[\limsup_{k \to +\infty} G(x^k, \epsilon_{k,i_k}) \right] \subset \partial f(\bar{x}).$$

Hence, any accumulation point $\bar{x}$ of $\{x^k\}$ is a stationary point. $\qquad\square$

13

4 Descent Directions via Subgradient Regularization

This section introduces the technique of subgradient regularization for constructing descent-oriented subdifferentials. In contrast to the Goldstein-type and bundle-type constructions discussed in Section 2.2, which rely on enlarged subdifferentials without exploring specific structures, we now focus on a subclass of nonsmooth functions known as marginal functions. For simplicity, we restrict attention to maximal marginal functions. Analogous arguments apply when the maximum is replaced by a minimum, i.e., when the maximal value function f is replaced by $(-f)$ (see Remark 1 following the definition of descent-oriented subdifferentials).

4.1 Subgradient regularization for marginal functions with fixed feasible sets

We start by analyzing the case where the feasible set of the marginal function is fixed, i.e., independent of x. Specifically, we consider the marginal function

$$f(x) \triangleq \max_{y \in Y \subset \mathbb{R}^m} \varphi(x, y), \tag{P}$$

where Y is convex and compact; $\varphi(\cdot, y)$ is differentiable for every fixed $y \in Y$; both $\varphi(\cdot, \cdot)$ and $\nabla_x \varphi(\cdot, \cdot)$ are jointly continuous; and $\varphi(x, \cdot)$ is concave for every fixed x. Under these assumptions, f is locally Lipschitz continuous [43, Theorem 10.31].

Let $Y^*(x)$ denote the solution set of the inner maximization problem for a given x, which must be convex. By Danskin's theorem [10, Theorem 2.1], we have

$$\partial f(x) = \operatorname{conv}\big(S(x)\big) \quad \text{with} \quad S(x) \triangleq \bigcup \big\{ \nabla_x \varphi(x, y) \mid y \in Y^*(x) \big\}. \tag{14}$$

The convex hull is redundant when $\nabla_x \varphi(x, y)$ is affine in y. For any fixed x and $\epsilon > 0$, we define the subgradient-regularized problem associated with (P) as:

$$Y^\epsilon(x) \triangleq \operatorname*{argmax}_{y \in Y} \left\{ \varphi(x, y) - \frac{\epsilon}{2} \|\nabla_x \varphi(x, y)\|^2 \right\}. \tag{P$^\epsilon$}$$

Throughout this subsection, we consider a set-valued map $G : \mathbb{R}^n \times (0, +\infty) \rightrightarrows \mathbb{R}^n$ defined by

$$G(x, \epsilon) \triangleq \bigcup \big\{ \nabla_x \varphi(x, y) \mid y \in Y^\epsilon(x) \big\}. \tag{15}$$

We note that when $\varphi(x, y) = y^\top A x - h(y)$ for a convex function h and a matrix $A \in \mathbb{R}^{m \times n}$, the above construction of $G(x, \epsilon)$ is different from the well-known smoothing technique by Nesterov [40], where the smoothed approximation of f in (P) is defined as $f_\epsilon(x) \triangleq \max_{y \in Y} \{\varphi(x, y) - \epsilon d(y)\}$ for some strongly convex function d. At a given x, the optimal solution of y is always unique and f_ϵ is continuously differentiable. In contrast, $Y^\epsilon(x)$ in (P$^\epsilon$) may not be a singleton, but G is always a descent-oriented subdifferential for f in this case, as shown in the following lemma. The proof can be found in Appendix A.1.

Lemma 2 (Subgradient regularization defines a descent-oriented subdifferential). Consider problem (P), and let $Y^\epsilon(x)$ and $G(x, \epsilon)$ be defined as in (P$^\epsilon$) and (15), respectively. Let $y^\epsilon : \mathbb{R}^n \to \mathbb{R}^m$ be a single-valued map satisfying $y^\epsilon(x) \in Y^\epsilon(x)$ for all $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$. Then for any fixed $\bar{x} \in \mathbb{R}^n$, the following properties hold:

(a) $\lim\limits_{\epsilon\downarrow 0, x\to\bar{x}} \varphi(x, y^\epsilon(x)) = f(\bar{x})$.

(b) $\limsup\limits_{\epsilon\downarrow 0, x\to\bar{x}} \nabla_x\varphi(x, y^\epsilon(x)) \subset \partial f(\bar{x})$. Consequently, the map G satisfies (G1).

(c) For any $\epsilon > 0$, $\limsup_{x\to\bar{x}} \nabla_x\varphi(x, y^\epsilon(x)) \subset \bigcup\{\nabla_x\varphi(\bar{x}, y) \mid y \in Y^\epsilon(\bar{x})\}$. Consequently, G satisfies the first condition in (G2′).

(d) If $S(\bar{x})$ is a convex set, then $\lim_{\epsilon\downarrow 0} \nabla_x\varphi(\bar{x}, y^\epsilon(\bar{x})) = \operatorname{argmin}_{v\in\partial f(\bar{x})} \|v\|$, and G also satisfies the second condition in (G2′). Thus, G is a descent-oriented subdifferential for f.

It can be seen from the above lemma that when $\nabla_x\varphi(x, y)$ is affine in y for any fixed x, the set $S(x)$ is convex for any x, and the mapping G defined in (15) is a descent-oriented subdifferential. In this case, the subgradient-regularized problem (P$^\epsilon$) reduces to a concave maximization over a convex compact set and can be solved to optimal in principle. We also highlight two useful properties in this setting. Part (a) shows that, although $Y^\epsilon(x)$ may not be a singleton, the gradient $\nabla_x\varphi(x, y)$ remains invariant over $y \in Y^\epsilon(x)$.

Lemma 3 (A special case of affine dependence in y). Consider problem (P) with $\varphi(x, y) = g(x) + y^\top\psi(x) - h(y)$, where $g : \mathbb{R}^n \to \mathbb{R}$ and $\psi : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable maps, and $h : \mathbb{R}^m \to \mathbb{R}$ is convex. Let G be defined as in (15). Then the following properties hold:

(a) For any $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$, the set $G(x, \epsilon)$ is a singleton.

(b) Let $\{x^k\} \subset \mathbb{R}^n$ and $\{\epsilon_k\} \subset (0, +\infty)$ be sequences such that $\max\{\|x^k - \bar{x}\|, \|G(x^k, \epsilon_k)\|\} \to_N 0$ for some point $\bar{x}$ and some index set $N \in \mathbb{N}_\infty^\sharp$. Then $\bar{x}$ is a stationary point.

Remark 5. Part (b) implies that, under the specific structure of φ, if an algorithm produces a sequence $\{x^k\}$ with an accumulation point $\bar{x}$ satisfying the condition in (b), then $\bar{x}$ is guaranteed to be a stationary point without the need for $\epsilon_k \to 0$. In contrast, the general convergence analysis of Algorithm 1 in Section 3 requires property (G1), which connects $G(x, \epsilon)$ to $\partial f(x)$ only asymptotically as $\epsilon \downarrow 0$.

Proof. (a) This result can be easily derived using the strong convexity of the squared function $\|\cdot\|^2$ and the linearity of $\nabla_x\varphi(x, y)$ in y. A detailed proof can be found in (e.g., [45, 34]) and is omitted here for brevity. (b) We let $\nabla\psi(x) \in \mathbb{R}^{n\times m}$ be the transpose of the Jacobian matrix of ψ at x. For any $y^k \in Y^{\epsilon_k}(x^k)$, by the optimality condition for problem (P$^\epsilon$), we have

$$0 \in -\psi(x^k) + \partial h(y^k) + \epsilon_k\nabla\psi(x^k)^\top\nabla_x\varphi(x^k, y^k) + \mathcal{N}_Y(y^k). \tag{16}$$

where $\mathcal{N}_Y(\bar{y}) \triangleq \{v \in \mathbb{R}^m \mid v^\top(y - \bar{y}) \leq 0 \text{ for all } y \in Y\}$ is the normal cone of Y at a point $\bar{y} \in Y$. Since Y is compact, by further passing to a subsequence if necessary, we assume that $y^k \to_N \bar{y} \in Y$. It follows from our assumption that $\|G(x^k, \epsilon_k)\| = \|\nabla_x\varphi(x^k, y^k)\| \to_N 0$. By the continuity of $\nabla_x\varphi(\cdot, \cdot)$, we also have $\nabla_x\varphi(\bar{x}, \bar{y}) = 0$. Taking the outer limit in (16) as $k \to +\infty$, and using the continuity of $\psi(\cdot), \nabla\psi(\cdot)$, along with the outer semicontinuity of $\partial h(\cdot)$ and $\mathcal{N}_Y(\cdot)$ (see [43, Propositions 6.6 and 8.7]), we obtain $0 \in [-\psi(\bar{x}) + \partial h(\bar{y}) + \mathcal{N}_Y(\bar{y})] = \partial_y(-\varphi)(\bar{x}, \bar{y}) + \mathcal{N}_Y(\bar{y})$. This implies that $\bar{y} \in Y^*(\bar{x})$. Since $\nabla_x\varphi(\bar{x}, \bar{y}) = 0$, we conclude that

$$0 = \nabla_x\varphi(\bar{x}, \bar{y}) \in \left[\bigcup\{\nabla_x\varphi(\bar{x}, y) \mid y \in Y^*(\bar{x})\}\right] \subset \partial f(\bar{x}),$$

and hence $\bar{x}$ is a stationary point. $\qquad\square$

15

To illustrate the broad applicability of the subgradient-regularization technique, we next present four examples that fall within the formulation of (P) and satisfy the structural assumption in Lemma 3.

Example 1 (Finite max of smooth functions). Let $\{f_i\}_{i=1}^m$ be continuously differentiable. Then

$$f(x) = \max_{1 \le i \le m} f_i(x) = \max_{y \in \Delta^m} \sum_{i=1}^m y_i f_i(x), \tag{17}$$

satisfies the assumption in Lemma 3.

Example 2 (Maximal eigenvalue functions). Consider the function

$$f(x) = \lambda_{\max}(\mathcal{A}(x)),$$

where $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}_m$ is an affine map of the form $\mathcal{A}(x) = A_0 + \mathcal{F}x$ with $\mathcal{F}x = \sum_{i=1}^n x_i A_i$ defined by matrices $A_i \in \mathcal{S}_m$ for $i = 0, 1, \cdots, n$. Here, $\lambda_{\max}(X)$ denotes the largest eigenvalue of the matrix $X \in \mathcal{S}_m$. The function $\lambda_{\max}(\cdot)$ is convex and can be expressed as the support function of the *spectrahedron* set $\mathcal{C}_m \triangleq \{Q \in \mathcal{S}_m \mid Q \succeq 0, \mathrm{Tr}(Q) = 1\}$, which is compact and convex. Using Rayleigh's variational formulation, we can rewrite the objective as a marginal function:

$$\lambda_{\max}(\psi(x)) = \max_{u \in \mathbb{R}^m, \|u\|=1} u^\top \mathcal{A}(x) u = \max_{y \in \mathcal{C}_m} \langle y, \mathcal{A}(x) \rangle.$$

Example 3 (Distributionally robust optimization). Let $\{s_i\}_{i=1}^N$ be observed data and $\ell(\cdot\,; s_i)$ be the corresponding loss. The distributionally robust formulation of the training problem based on ϕ-divergence [31] takes the form of

$$f(x) = \max_{\left\{p \in \Delta^N \,\middle|\, \sum_{i=1}^N \phi(Np_i)/N \le \rho\right\}} \sum_{i=1}^N \left(p_i\, \ell(x; s_i) - \frac{1}{N} \psi(Np_i)\right), \qquad \rho > 0,$$

where $\phi, \psi : [0, +\infty) \to \mathbb{R} \cup \{+\infty\}$ are convex functions satisfying $\phi(1) = \psi(1) = 0$. If $\ell(\cdot\,; s)$ is continuously differentiable, the assumptions required in Lemma 3 hold.

Example 4 (Composition of a convex function with a smooth map). Consider the composite function $f(x) = h(c(x))$, where $c : \mathbb{R}^n \to \mathbb{R}^m$ is C^1-smooth and $h : \mathbb{R}^m \to \mathbb{R}$ is a convex function. By Fenchel duality, we write f as $f(x) = \max_{y \in \mathbb{R}^m} \{y^\top c(x) - h^*(y)\}$. For any fixed x, we denote by $\nabla c(x) \in \mathbb{R}^{n \times m}$ the transpose of the Jacobian matrix of c at x. Then $\partial f(x) = \bigcup \{\nabla c(x)\, y \mid y \in Y^*(x)\}$ where $Y^*(x)$ denotes the solution set of the above maximization problem. We construct a descent-oriented subdifferential via subgradient regularization:

$$G(x, \epsilon) = \bigcup \{\nabla c(x)\, y \mid y \in Y^\epsilon(x)\}, \tag{18}$$

where

$$Y^\epsilon(x) = \operatorname*{argmax}_{y \in \operatorname{dom}(h^*)} \left\{y^\top c(x) - h^*(x) - \frac{\epsilon}{2} \|\nabla c(x)\, y\|^2\right\}.$$

4.2 Subgradient regularization for marginal functions with varying feasible sets

We turn to the case in which the feasible set of the marginal function also depends on x. Consider the following marginal function:

$$f(x) \triangleq \max_{y \in \mathbb{R}^m} \left\{ \varphi_0(x,y), \quad \text{subject to } \varphi_j(x,y) \le 0, \ j = 1, \cdots, r \right\} \tag{P_x}$$

$$= \max_{y \in \mathbb{R}^m} \min_{\lambda \ge 0} \left[\mathcal{L}(x,y;\lambda) \triangleq \varphi_0(x,y) - \sum_{j=1}^r \lambda_j \, \varphi_j(x,y) \right].$$

Here, all functions $\{\varphi_j(x,y)\}_{j=0}^r$ are continuously differentiable. Given (x,y), we define the index set of active constraints as $J(x,y) \triangleq \{j \in \{1, \cdots, r\} \mid \varphi_j(x,y) = 0\}$. For the inner maximization in (P_x), denote the solution set by $Y^*(x)$. We summarize basic assumptions below:

(i) For every x, the function $\varphi_0(x, \cdot)$ is concave, and $\{\varphi_j(x, \cdot)\}_{j=1}^r$ are convex.

(ii) For every x, the feasible set $Y(x) \triangleq \{y \in \mathbb{R}^m \mid \varphi_j(x,y) \le 0, j = 1, \cdots, r\}$ is nonempty, and the mapping $x \mapsto Y(x)$ is locally bounded.

(iii) The Mangasarian-Fromovitz constraint qualification (MFCQ) holds at every x: For any $y \in Y^*(x)$, there exists $w \in \mathbb{R}^m$ such that $\nabla_y \varphi_j(x,y)^\top w < 0$ for all $j \in J(x,y)$.

Under assumption (i), the set of Lagrange multipliers associated with the inner maximization in (P_x) is the same for all $y \in Y^*(x)$. Thus, we can denote this set as $\Lambda(x)$. Assumption (iii) ensures that $\Lambda(x)$ is nonempty and bounded for any fixed x [3, Proposition 5.47]. Furthermore, by [21, Theorem 5.1], assumptions (ii) and (iii) imply the local Lipschitz continuity of f.

We now present two scenarios, each giving rise to a distinct form of subgradient regularization based on the characterization of $\partial f(x)$.

Case 1: We strengthen assumption (iii) by requiring uniqueness of multipliers. Specifically, we replace the MFCQ with:

(iii′) The linear independence constraint qualification (LICQ) holds at every x: For any $y \in Y^*(x)$, the gradients $\{\nabla_y \varphi_j(x,y)\}_{j \in J(x,y)}$ are linearly independent.

Under (iii′), it follows from [21, Corollary 5.4] that

$$\partial f(x) = \mathrm{conv}\left(\bigcup \{\nabla_x \mathcal{L}(x,y;\lambda(x)) \mid y \in Y^*(x)\} \right) \quad \text{for } \lambda(x) = \Lambda(x).$$

If the convex hull operation is superfluous (for instance, when $\{\nabla_x \varphi_j(x, \cdot)\}_{j=0}^r$ are affine), and if an oracle returns the unique multiplier $\lambda(x)$ for each x, we consider the following subgradient-regularized problem:

$$Y^\epsilon(x) \triangleq \operatorname*{argmax}_{y \in \mathbb{R}^m} \left\{ \varphi_0(x,y) - \frac{\epsilon}{2} \|\nabla_x \mathcal{L}(x,y;\lambda(x))\|^2 \right\}$$
$$\text{subject to } \varphi_j(x,y) \le 0, \quad j = 1, \cdots, r.$$

Correspondingly, we define the associated descent-oriented subdifferential as

$$G(x,\epsilon) \triangleq \bigcup \{\nabla_x \mathcal{L}(x,y;\lambda(x)) \mid y \in Y^\epsilon(x)\}. \tag{19}$$

Case 2: Instead of requiring the uniqueness of multipliers as in Case 1, we ensure that $Y^*(x)$ is a singleton and the map $x \mapsto Y^*(x)$ is locally Lipschitz continuous. For simplicity, suppose additionally that $\varphi(x, \cdot)$ is strongly concave for every fixed x. This assumption can be relaxed to strong second-order sufficient optimality conditions (see [3, (5.120)]). Under this condition, by [39, Corollary 4], we have

$$\partial f(x) = \bigcup \left\{ \nabla_x \mathcal{L}(x, y(x); \lambda) \mid \lambda \in \Lambda(x) \right\} \quad \text{for } y(x) = Y^*(x).$$

If an oracle returns the unique optimal solution $y(x) = Y^*(x)$ for each x, we define the following subgradient-regularized problem:

$$\Lambda^\epsilon(x) \triangleq \operatorname*{argmin}_{\lambda \geq 0} \; \left\{ \mathcal{L}(x, y(x); \lambda) + \frac{\epsilon}{2} \left\| \nabla_x \mathcal{L}(x, y(x); \lambda) \right\|^2 \right\}$$
$$\text{subject to } \nabla_y \mathcal{L}(x, y(x); \lambda) = 0.$$

We then construct the associated descent-oriented subdifferential as

$$G(x, \epsilon) \triangleq \bigcup \left\{ \nabla_x \mathcal{L}(x, y(x); \lambda) \mid \lambda \in \Lambda^\epsilon(x) \right\}. \tag{20}$$

We conclude this section by presenting the main results for the two scenarios described above. The proof follows similar arguments as those employed in Lemma 2; for completeness, we provide a concise outline of the proof in Appendix A.2.

Lemma 4 (Subgradient regularization defines a descent-oriented subdifferential). Consider problem (P_x) under assumptions (i)-(iii).
(a) If assumption (iii') holds, and $\{\nabla_x \varphi_j(x, y)\}_{j=0}^r$ are affine in y for any fixed x, then the mapping G defined by (19) is a descent-oriented subdifferential for f.
(b) If $\varphi_0(x, \cdot)$ is strongly concave for any fixed x, then the mapping G defined by (20) is a descent-oriented subdifferential for f.

5   Adaptive Stepsizes and Reduction to the Prox-Linear Method

The focus of last section was on the choice of descent directions. In this section, we shift attention to stepsize control and refine Algorithm 1 by further exploiting problem structure. We recall the nonsmooth composite problem in Example 4:

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} \; f(x) = h(c(x)), \tag{21}$$

where $c : \mathbb{R}^n \to \mathbb{R}^m$ is C^1-smooth and $h : \mathbb{R}^m \to \mathbb{R}$ is a convex function. Throughout this section, we also assume that h is L-Lipschitz continuous with $\inf_y h(y) > -\infty$, and ∇c is β-Lipschitz continuous. For this problem, we will show in Proposition 7 that the direction (18) obtained from subgradient regularization coincides with that of the widely studied prox-linear method. This connection motivates us to study a variant of Algorithm 1 that adaptively adjusts the regularization parameter ϵ and stepsize η.

5.1 A dual perspective of the prox-linear method

This section presents a dual perspective of the prox-linear method through the lens of subgradient regularization. To build intuition, we begin with a concrete example before generalizing in Proposition 7. Consider the finite max of smooth functions $f(x) = \max_{1 \leq i \leq d} f_i(x)$ in Example 1, and recall that a descent-oriented subdifferential can be constructed via subgradient regularization. We now turn to the prox-linear method, which updates x^{k+1} by solving

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f(x; x^k) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\},$$

where $f(x; x^k)$ is a piecewise linear approximation of f at x^k:

$$f(x; x^k) \triangleq \max_{1 \leq i \leq d} \left\{ f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right\} = \max_{y \in \Delta^d} \left\{ \sum_{i=1}^d y_i \left(f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right) \right\}.$$

This leads to the minimax formulation:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \Delta^d} \left\{ \Phi(x, y) \triangleq \sum_{i=1}^d y_i \left(f_i(x^k) + \nabla f_i(x^k)^\top (x - x^k) \right) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right\}.$$

By the minimax theorem [18, Propositions 1.2 and 2.4], we may interchange the order of minimization and maximization. Completing the square reveals that the minimization problem attains its minimum at $x = x^k - \epsilon \sum_{i=1}^d y_i \nabla f_i(x^k)$. Substituting into the maximization problem gives:

$$\operatorname*{maximize}_{y \in \Delta^d} \left\{ \sum_{i=1}^d y_i \, f_i(x^k) - \frac{\epsilon}{2} \left\| \sum_{i=1}^d y_i \nabla f_i(x^k) \right\|^2 \right\},$$

which coincides with the subgradient-regularized problem (P^ϵ) for Example 1. Any solution $\bar{y} \in \Delta^d$ of this problem yields a saddle point $(x^{k+1}, \bar{y})$ of $\Phi(x, y)$, and we recover the prox-linear update as

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \Phi(x, \bar{y}) = x^k - \epsilon\, G(x^k, \epsilon) \quad \text{with} \quad G(x^k, \epsilon) \triangleq \sum_{i=1}^d \bar{y}_i \nabla f_i(x^k).$$

Here, $G(x^k, \epsilon)$ is a descent-oriented subgradient for f, representing a convex combination of gradients $\{\nabla f_i(x^k)\}_{i=1}^d$. This example illustrates how the prox-linear method implicitly constructs descent directions through subgradient regularization, revealing its dual interpretation.

Next, we come back to the general composite problem (21). A descent-oriented subdifferential has been defined in Example 4. To define the prox-linear update, we consider an approximation of f at x by linearizing the inner smooth map c:

$$f(z; x) \triangleq h\left(c(x) + \nabla c(x)^\top (z - x) \right),$$

which is convex in z. It is known that $f(\cdot; x)$ is a *two-sided model* [17] satisfying

$$-\frac{L\beta}{2} \|z - x\|^2 \leq f(z) - f(z; x) \leq \frac{L\beta}{2} \|z - x\|^2. \tag{22}$$

This motivates the prox-linear method, which generates iterates via

$$x^{k+1} = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ f(z; x^k) + \frac{1}{2\epsilon} \|z - x^k\|^2 \right\}. \qquad \text{(Prox-linear)}$$

Complementing the model-based interpretation in (22), the following proposition offers a novel dual perspective by connecting it to subgradient regularization.

Proposition 7 (Equivalence of prox-linear and subgradient regularization)**.** Consider problem (21), and let G be the descent-oriented subdifferential defined in (18) via subgradient regularization. Then (Prox-linear) with stepsize ϵ is equivalent to the update $x^{k+1} = x^k - \epsilon\, G(x^k, \epsilon)$.

Proof. By the convexity of h, we can rewrite (Prox-linear) as a minimax problem using the conjugate function of h:

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \max_{y \in \mathbb{R}^m} \left[y^\top \big(c(x^k) + \nabla c(x^k)^\top (x - x^k) \big) - h^*(y) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right] \right\}. \qquad (23)$$

Observe that for any x, the inner maximum is attained at some $y_x \in \partial h\big(c(x^k) + \nabla c(x^k)^\top(x - x^k)\big)$. Since h is L-Lipschitz continuous, it follows that $\|y_x\| \leq L$ for any x. Consequently, we can write (23) equivalently by restricting y in the compact set $Y = \{y \in \mathbb{R}^m \mid \|y\| \leq L\}$. Similarly, due to the strong convexity of $\|x - x^k\|^2/(2\epsilon)$, we can also restrict x in the compact set $\overline{\mathbb{B}}_\rho(x^k)$, where the radius is given by $\rho \triangleq \max\{\epsilon L \|\nabla c(x^k)\|, \sqrt{2\epsilon\, [h(c(x^k)) - \inf h]}\}$. Thus, the minimax problem (23) is equivalent to

$$\min_{x \in \overline{\mathbb{B}}_\rho(x^k)} \max_{y \in Y} \left[y^\top \big(c(x^k) + \nabla c(x^k)^\top(x - x^k) \big) - h^*(y) + \frac{1}{2\epsilon} \|x - x^k\|^2 \right]. \qquad (24)$$

Since the objective function is convex-concave, $\overline{\mathbb{B}}_\rho(x^k)$ and Y are compact, it follows from the minimax theorem [18, Propositions 1.2 and 2.4] that there exists a saddle point on $\overline{\mathbb{B}}_\rho(x^k) \times Y$. Thus, we can interchange the minimize and maximum in (24). As the inner minimal value is attained at $x = \big(x^k - \epsilon\, \nabla c(x^k)\, y\big) \in \overline{\mathbb{B}}_\rho(x^k)$ for any $y \in Y$, we get a dual solution by direct computation:

$$\bar{y} \in \underset{y \in Y}{\operatorname{argmax}} \left\{ y^\top c(x^k) - h^*(y) - \frac{\epsilon}{2} \big\| \nabla c(x^k)\, y \big\|^2 \right\}. \qquad (25)$$

Because $(x^{k+1}, \bar{y})$ is a saddle point of the minimax problem (24), by the optimality of x^{k+1} for the minimization problem at $\bar{y}$, we immediately obtain $x^{k+1} = x^k - \epsilon\, \nabla c(x^k)\, \bar{y}$. Noting that the feasible region Y in (25) can be omitted without affecting the optimal solution, we have $\nabla c(x^k)\, \bar{y} = G(x^k, \epsilon)$ and, thus, $x^{k+1} = x^k - \epsilon\, G(x^k, \epsilon)$, completing the proof. $\qquad \square$

As a consequence of Proposition 7, we obtain a stronger version of Proposition 2 on uniform non-diminishing ranges of ϵ and η to guarantee the descent inequality (2). The result below follows directly from [17, Lemma 5.1].

Corollary 1 (Descent directions for composite functions)**.** Consider the nonsmooth composite problem (21), and let G be the descent-oriented subdifferential defined in (18). For any $\epsilon \in (0, 1/(2L\beta)]$, the following descent inequality holds for every $\bar{x}$:

$$f(\bar{x} - \epsilon g) \leq f(\bar{x}) - \epsilon \|g\|^2, \qquad \forall\, g \in G(\bar{x}, \epsilon).$$

20

5.2 An adaptive descent-oriented subgradient method

We have seen in previous section that specializing to the composite structure (21), the direction obtained from subgradient regularization is equivalent to that of the prox-linear method. It is known that prox-linear method with a constant stepsize enjoys local linear convergence under an error bound condition [17]. This naturally leads to the question of whether our proposed algorithm can also attain a local linear rate under the same condition.

However, Algorithm 1 with G defined in (18) does not guarantee such a rate. The algorithm typically decreases the parameter ϵ over iterations and employs a diminishing stepsize to ensure subsequential convergence to stationary points. To address this limitation, we propose an adaptive variant of Algorithm 1, presented as Algorithm 2.

Algorithm 2 An adaptive descent-oriented subgradient method

1: **Input:** starting point x_0, regularization parameter $\epsilon_{0,0} > 0$, a non-summable sequence $\{a_t\} \downarrow 0$, initial stationary target ν_0, optimality tolerances $\epsilon_{\mathrm{tol}}, \nu_{\mathrm{tol}} \geq 0$, counter of ratio test $t = 0$, reduction factors $\theta_\epsilon, \theta_\nu \in (0,1)$, and the Armijo parameter $\alpha \in (0,1)$

2: **for** $k = 0, 1, \cdots$ **do**

3: **for** $i = 0, 1, \cdots$ **do**

4:   Set $\epsilon_{k,i} = \epsilon_{k,0}\, 2^{-i}$

5:   Update direction $g^{k,i} \in G(x^k, \epsilon_{k,i})$

6:   **if** $\epsilon_{k,i} \leq \epsilon_{\mathrm{tol}}$ and $\|g^{k,i}\| \leq \nu_{\mathrm{tol}}$ **then**

7:    Set $i_k = i$ and **return** x^k ▷ *approximate stationary*

8:   **for** $j = 0, 1, \cdots, i$ **do**

9:    Set $\eta = \epsilon_{k,0}\, 2^{-j}$ ▷ *line search*

10:    **if** $f(x^k - \eta\, g^{k,i}) \leq f(x^k) - \alpha\, \eta \|g^{k,i}\|^2$ **then**

11:     Set $i_k = i$ and $\eta_k = \mathrm{argmin}\left\{ f(x^k - \eta\, g^{k,i}) \mid \eta \in \{\epsilon_{k,0}\, 2^{-j}, \cdots, \epsilon_{k,0}\, 2^{-i}\} \right\}$

12:     Go to line 13

13:   Update $x^{k+1} = x^k - \eta_k\, g^{k,i}$

14:   **if** $\|g^{k,i_k}\| \leq \nu_k$ **then**

15:    $t \leftarrow (t+1)$

16:    Update $\tilde{g}^k \in G(x^k, \tilde{\epsilon}_k)$ with $\tilde{\epsilon}_k = (a_t)^{1/4}$

17:    **if** $\tilde{\epsilon}_k \leq \epsilon_{\mathrm{tol}}$ and $\|\tilde{g}^k\| \leq \nu_{\mathrm{tol}}$ **then**

18:     **Return** x^k ▷ *approximate stationary*

19:    Set $\nu_{k+1} = \theta_\nu \cdot \nu_k$

20:    **if** $\dfrac{\tilde{\epsilon}_k \|\tilde{g}^k\|}{\left(\epsilon_{k,i_k} \|g^{k,i_k}\|\right)^{1/2}} \leq \dfrac{1}{\epsilon_{k,0}}$ **then**

21:     Set $\epsilon_{k+1,0} = \epsilon_{k,0}$ ▷ *ratio test*

22:    **else**

23:     Set $\epsilon_{k+1,0} = \theta_\epsilon \cdot \epsilon_{k,0}$

24:   **else**

25:    Set $\epsilon_{k+1,0} = \epsilon_{k,0}$ and $\nu_{k+1} = \nu_k$

The key innovation of Algorithm 2 lies in the presence of two sequences $\{\epsilon_{k,0}\}$ and $\{\tilde{\epsilon}_k\}$. The primary sequence $\{\epsilon_{k,0}\}$ is responsible for computing descent directions, as in Algorithm 1, while the auxiliary diminishing sequence $\{\tilde{\epsilon}_k\}$ is used to monitor approximate stationarity through the norm of $\tilde{g}^k \in G(x^k, \tilde{\epsilon}_k)$. We compare $\|\tilde{g}^k\|$ and $\|g^{k,i_k}\|$, where g^{k,i_k} denotes the descent-oriented

subgradient satisfying the descent inequality, to determine whether to reduce $\epsilon_{k,0}$ in the next iteration. Specifically, when the ratio test

$$\frac{\tilde{\epsilon}_k \|\tilde{g}^k\|}{(\epsilon_{k,i_k} \|g^{k,i_k}\|)^{1/2}} \leq \frac{1}{\epsilon_{k,0}}$$

is satisfied, $\epsilon_{k,0}$ remains a constant; otherwise, it is reduced by a factor of θ_ϵ. This mechanism allows Algorithm 2 to maintain a non-diminishing effective stepsize whenever possible, thereby recovering local linear convergence for structured problems while still ensuring subsequential convergence to stationary points in general settings.

5.2.1 Convergence analysis

The inner loop indexed by i in Algorithm 2 is essentially the same as that of Algorithm 1. So the inner loop terminates in finitely many iterations by Proposition 6 and $i_k < +\infty$. We next prove the subsequential convergence.

Theorem 2 (Subsequential convergence of Algorithm 2). Suppose that f is bounded from below. Let $\{x^k\}$ be a sequence generated by Algorithm 2 with $\epsilon_{\text{tol}} = \nu_{\text{tol}} = 0$. One of the following occurs:
(a) For some k, the inner loop does not terminate. Then x^k is a stationary point.
(b) There is an index set $K \in \mathbb{N}_\infty^\sharp$ such that $\tilde{g}^k \to_K 0$. Consequently, for any such subsequence $\tilde{g}^k \to_K 0$, any accumulation point of $\{x^k\}_{k \in K}$ is a stationary point.
(c) Any accumulation point of $\{x^k\}$ is a stationary point.

Proof. If the inner loop does not terminate for some k, then by Proposition 6, we have $0 \in \partial f(x^k)$. Now suppose the inner loop terminates for every k. We will prove that either (b) or (c) holds. According to Proposition 6, there exists an index i_k for each k such that the following descent inequality holds:

$$f(x^{k+1}) = f(x^k - \eta_k g^{k,i_k}) \leq f(x^k) - \alpha \eta_k \|g^{k,i_k}\|^2.$$

Since f is bounded from below, summing the descent inequality over k gives

$$\sum_{k=0}^{+\infty} \eta_k \|g^{k,i_k}\|^2 = \sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| \cdot \|g^{k,i_k}\| < +\infty. \tag{26}$$

Case 1: Suppose there exist $\underline{k} \in \mathbb{N}$, $\underline{\epsilon} > 0$, and $\underline{\nu} > 0$ such that $\epsilon_{k,0} = \underline{\epsilon}$ and $\nu_k = \underline{\nu}$ for all $k \geq \underline{k}$. Then, for all such k, we have $\|g^{k,i_k}\| > \underline{\nu}$, and consequently, $\tilde{g}^k$ is not constructed. Therefore statement (b) does not hold, and it remains to prove (c). Using (26) and the fact that $\|g^{k,i_k}\| > \underline{\nu}$ for all $k \geq \underline{k}$, we can apply the same arguments as in Case 1 of Theorem 1 to conclude that $\{x^k\}$ converges to some point $\bar{x}$ with $0 \in \partial f(\bar{x})$. Therefore, statement (c) holds.

Case 2: Otherwise, there exists $N \in \mathbb{N}_\infty^\sharp$ such that $g^{k,i_k} \to_N 0$. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. By the descent inequality, we apply Lemma 1 to have that either $x^k \to \bar{x}$, or $\liminf_{k \to +\infty} \max\{\|x^k - \bar{x}\|, \|g^{k,i_k}\|\} = 0$. In either scenarios, by passing to a subsequence if necessary, we can assume that $\max\{\|x^k - \bar{x}\|, \|g^{k,i_k}\|\} \to_N 0$.
Case 2.1: If $\epsilon_{k,0} \downarrow 0$, since $\epsilon_{k,i_k} \leq \epsilon_{k,0} \downarrow 0$, property (G1) implies

$$0 = \lim_{k(\in N) \to +\infty} g^{k,i_k} \in \left[\limsup_{k(\in N) \to +\infty} G(x^k, \epsilon_{k,i_k}) \right] \subset \partial f(\bar{x}).$$

Thus, any accumulation point $\bar{x}$ of $\{x^k\}$ satisfies $0 \in \partial f(\bar{x})$, establishing (c).

Case 2.2: Otherwise, the sequence $\{\epsilon_{k,0}\}$ converges monotonically to a positive scalar $\underline{\epsilon}$, and there exists $\underline{k} \in \mathbb{N}$ such that $\epsilon_{k,0} = \underline{\epsilon}$ for all $k \geq \underline{k}$. Define the index set $N' \triangleq \{k \in \mathbb{N} \mid \|g^{k,i_k}\| \leq \nu_k\}$. Since $\|g^{k,i_k}\| \to_N 0$, it follows that $N' \in \mathbb{N}_\infty^\sharp$. Let k_t denote the t-th index in N'. To prove statement (b), we aim to find a subsequence $\tilde{g}^k \to_K 0$ for some index set $K \in \mathbb{N}_\infty^\sharp$ with $K \subset N'$.

Suppose for contradiction that there exists $\delta > 0$ such that $\|\tilde{g}^k\| \geq \delta$ for all $k \in N'$, or equivalently, $\|\tilde{g}^{k_t}\| \geq \delta$ for all $t \in \mathbb{N}$. For any $t \in \mathbb{N}$ with $k_t \geq \underline{k}$, since $\epsilon_{k_t,0} = \underline{\epsilon}$, the ratio test must be satisfied:

$$\frac{\tilde{\epsilon}_{k_t} \|\tilde{g}^{k_t}\|}{\left(\epsilon_{k_t,i_{k_t}} \|g^{k_t,i_{k_t}}\|\right)^{1/2}} \leq \frac{1}{\epsilon_{k_t,0}} = \frac{1}{\underline{\epsilon}}.$$

Observe that $\tilde{\epsilon}_{k_t} = (a_t)^{1/4}$ by the update rule of $\tilde{\epsilon}_k$. For all $t \in \mathbb{N}$ with $k_t \geq \underline{k}$, Rearranging the inequality and using $\epsilon_{k_t,i_{k_t}} = \epsilon_{k_t,0} 2^{-i_{k_t}} \leq \eta_{k_t} \leq \epsilon_{k_t,0} = \underline{\epsilon}$ yield

$$(a_t)^{\frac{1}{4}} \|\tilde{g}^{k_t}\| = \tilde{\epsilon}_{k_t} \|\tilde{g}^{k_t}\| \leq \underline{\epsilon}^{-1} \left(\epsilon_{k_t,i_{k_t}} \|g^{k_t,i_{k_t}}\|\right)^{1/2} \leq \underline{\epsilon}^{-3/4} \left(\eta_{k_t} \|g^{k_t,i_{k_t}}\|^2\right)^{1/4}.$$

Summing both sides over $t \in \mathbb{N}$ with $k_t \geq \underline{k}$ and using the lower bound $\|\tilde{g}^{k_t}\| \geq \delta$, we get

$$\delta^4 \sum_{\{t \in \mathbb{N} \mid k_t \geq \underline{k}\}} a_t \leq \sum_{\{t \in \mathbb{N} \mid k_t \geq \underline{k}\}} a_t \|\tilde{g}^{k_t}\|^4 \leq \underline{\epsilon}^{-3} \cdot \sum_{\{t \in \mathbb{N} \mid k_t \geq \underline{k}\}} \eta_{k_t} \|g^{k_t,i_{k_t}}\|^2$$

$$\leq \underline{\epsilon}^{-3} \cdot \sum_{k \in \mathbb{N}} \eta_k \|g^{k,i_k}\|^2 \overset{(26)}{<} \infty.$$

This implies $\sum_{\{t \in \mathbb{N} \mid k_t \geq \underline{k}\}} a_t < +\infty$, contradicting the fact that $\{a_t\}$ is a non-summable sequence. Hence, there exists a subsequence $\tilde{g}^k \to_K 0$ for some index set $K \in \mathbb{N}_\infty^\sharp$ with $K \subset N'$. Finally, let $\bar{x}$ be any accumulation point of $\{x^k\}_{k \in K}$. Since $\tilde{\epsilon}_{k_t} = (a_t)^{1/4} \downarrow 0$, it follows from property (G1) that $0 \in \partial f(\bar{x})$. This establishes statement (b). $\qquad\square$

5.2.2 Local linear convergence for nonsmooth composite functions

In this section, we analyze convergence rate of Algorithm 2 to solve composite problem (21) under regularity conditions, where $G(x, \epsilon)$ is obtained from the subgradient regularization as in (18). Let us denote the set of stationary points of (21) by $(\partial f)^{-1}(0) \triangleq \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}$. We begin by recalling the notion of metric subregularity.

Definition 2. A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *metrically subregular* at $(\bar{x}, \bar{y}) \in \operatorname{gph} S$ if there exist positive scalars κ and τ such that

$$\operatorname{dist}\left(x, S^{-1}(\bar{y})\right) \leq \kappa \cdot \operatorname{dist}(\bar{y}, S(x)), \qquad \forall\, x \in \mathbb{B}_\tau(\bar{x}).$$

The following lemma provides an error bound based on metric subregularity. It is a uniformed version of [17, Theorem 5.11] in terms of ϵ. A self-contained proof is given in Appendix B.1.

Lemma 5. Consider problem (21) with G defined in (18). Let $\bar{x}$ be a stationary point. If ∂f is metrically subregular at $(\bar{x}, 0)$ with modulus κ, then for any $\bar{\epsilon} > 0$, there exists a constant $\tau > 0$ such that the following error bound holds:

$$\operatorname{dist}\left(x, (\partial f)^{-1}(0)\right) \leq \left((3L\beta\epsilon + 2)\kappa + 2\epsilon\right) \cdot \|G(x, \epsilon)\|, \qquad \forall\, x \in \mathbb{B}_\tau(\bar{x}),\ \epsilon \in (0, \bar{\epsilon}).$$

With all the ingredients, we begin by establishing in Proposition 8 that the stepsize sequence $\{\eta_k\}$ in Algorithm 2 remains uniformly bounded away from zero. Recall that for each iteration k, the stepsize η_k is selected from the set $\{\epsilon_{k,0}, 2^{-1}\epsilon_{k,0}, \cdots, 2^{-i_k}\epsilon_{k,0}\}$ satisfying the descent condition $f(x^k - \eta g^{k,i}) \leq f(x^k) - \alpha\,\eta\|g^{k,i}\|^2$. If the ratio test fails, the upper bound $\epsilon_{k,0}$ is reduced by a factor of θ_ϵ. According to Corollary 1 and given $\alpha \in (0,1)$, any $\epsilon \in (0, 1/(2L\beta)]$ paired with a stepsize $\eta = \epsilon$ ensures the descent condition. Thus, a uniform lower bound on $\{\eta_k\}$ can be guaranteed by showing that $\epsilon_{k,0}$ is only reduced by a factor of θ_ϵ finitely many times; i.e., the ratio test fails only for finitely many iterations under an error bound condition. We defer the proof to Appendix B.2.

Proposition 8 (Eventual constant stepsize)**.** Consider problem (21), and let $\{x^k\}$ be a sequence generated by Algorithm 2 using G defined in (18). Suppose that the level set $X \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ is bounded, and ∂f is metrically subregular at $(\bar{x}, 0)$ for every $\bar{x} \in (\partial f)^{-1}(0) \cap X$. Then there exists a constant $\underline{\epsilon} > 0$ such that $\epsilon_{k,0} = \underline{\epsilon}$ for all sufficiently large k. Consequently, $\eta_k \geq \epsilon_{k,i_k} \geq \min\{\underline{\epsilon}, 1/(4L\beta)\}$ for all sufficiently large k.

To establish local linear convergence, we next state a key inequality from [17, Lemma 5.1] that quantifies the decrease in the objective value after taking a step along the direction $-G(x, \epsilon)$.

Lemma 6. Consider nonsmooth composite problem (21) with G defined in (18). For any $x, y \in \mathbb{R}^n$ and $\epsilon > 0$, we have

$$f(y) \geq f\big(x - \epsilon\,G(x,\epsilon)\big) + \langle G(x,\epsilon), y - x \rangle + \frac{\epsilon}{2}(2 - L\beta\epsilon)\|G(x,\epsilon)\|^2 - \frac{L\beta}{2}\|x - y\|^2. \tag{27}$$

We now impose an additional assumption ensuring a separation of the isocost surfaces of f on the set of stationary points $(\partial f)^{-1}(0)$. This assumption is standard in the literature, and sufficient conditions for it to hold can be found in [35].

Assumption 1. There is a constant $\delta_0 > 0$ such that

$$x \in (\partial f)^{-1}(0), \ y \in (\partial f)^{-1}(0), \ f(x) \neq f(y) \quad \Rightarrow \quad \|x - y\| \geq \delta_0.$$

Finally, we arrive at the main convergence result.

Corollary 2 (Local linear convergence)**.** Consider problem (21) under Assumption 1, and let $\{x^k\}$ be a sequence generated by Algorithm 2 with G defined in (18). Suppose that the level set $X \triangleq \{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ is compact, and ∂f is metrically subregular at $(\bar{x}, 0)$ for every $\bar{x} \in (\partial f)^{-1}(0) \cap X$. Then for all sufficiently large k, the function values $f(x^k)$ and the iterates x^k converge linearly.

Proof. By Proposition 8, there exist $\underline{k} \in \mathbb{N}$ and $\underline{\epsilon} > 0$ such that $\epsilon_{k,0} = \underline{\epsilon}$ and $\eta_k \geq \epsilon_{k,i_k} \geq \underline{\eta} \triangleq \min\{\underline{\epsilon}, 1/(4L\beta)\}$ for all $k \geq \underline{k}$. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. By the continuity of f, $f(\bar{x})$ is an accumulation point of $\{f(x^k)\}$. Since $\{f(x^k)\}$ is non-increasing and bounded from below, it converges and $f(x^k) \to f(\bar{x})$. From the descent inequality

$$f(x^{k+1}) \leq f(x^k) - \alpha\,\eta_k\|g^{k,i_k}\|^2 = f(x^k) - \frac{\alpha}{\eta_k}\|x^{k+1} - x^k\|^2 \tag{28}$$

and the lower bound $\eta_k \geq \underline{\eta}$, we obtain

$$\|g^{k,i_k}\|^2 \leq \frac{f(x^k) - f(x^{k+1})}{\alpha\,\eta_k} \leq \frac{f(x^k) - f(x^{k+1})}{\alpha\,\underline{\eta}} \to 0.$$

24

Thus, $G(x^k, \epsilon_{k,i_k}) = g^{k,i_k} \to 0$. By Lemma 3(b), $\bar{x}$ is a stationary point. Since ∂f is metrically subregular at $(\bar{x}, 0)$, Lemma 5 guarantees the existence of positive scalars $\kappa_{\bar{x}}$ and τ such that

$$\text{dist}\left(x, (\partial f)^{-1}(0)\right) \leq \left((3L\beta\epsilon + 2)\kappa_{\bar{x}} + 2\epsilon\right) \cdot \|G(x, \epsilon)\|, \qquad \forall\, x \in \mathbb{B}_\tau(\bar{x}),\ \epsilon > 0. \tag{29}$$

Now apply the inequality (27) from Lemma 6 with $x = x^k$, $\epsilon = \epsilon_{k,i_k}$, and $y = \hat{x}^k$, where $\hat{x}^k \in (\partial f)^{-1}(0)$ is a point such that $\|x^k - \hat{x}^k\| = \text{dist}\left(x^k, (\partial f)^{-1}(0)\right)$. For all $k \in \mathbb{N}$, we have

$$\begin{aligned}
f\left(x^k - \epsilon_{k,i_k} g^{k,i_k}\right) &- f(\hat{x}^k) \\
&\leq -\langle g^{k,i_k}, \hat{x}^k - x^k \rangle - \frac{\epsilon_{k,i_k}}{2}(2 - L\beta\epsilon_{k,i_k})\|g^{k,i_k}\|^2 + \frac{L\beta}{2}\|x^k - \hat{x}^k\|^2 \\
&\leq \|g^{k,i_k}\|^2 \left(\frac{\|x^k - \hat{x}^k\|}{\|g^{k,i_k}\|} - \frac{\epsilon_{k,i_k}}{2}(2 - L\beta\epsilon_{k,i_k}) + \frac{L\beta}{2}\frac{\|x^k - \hat{x}^k\|^2}{\|g^{k,i_k}\|^2}\right).
\end{aligned}$$

Define $\gamma \triangleq \max\{1, (3L\beta\underline{\epsilon} + 2)\kappa_{\bar{x}} + 2\underline{\epsilon}\}$. Consider an iterate x^k such that $k \geq \underline{k}$ and $\|x^k - \bar{x}\| < \min\{\tau, \delta_0/2\}$. The error bound (29), combined with the fact $\epsilon_{k,i_k} \leq \underline{\epsilon}$, implies that $\|x^k - \hat{x}^k\| \leq \gamma\|g^{k,i_k}\|$. Furthermore, noting that $f(x^{k+1}) \leq f(x^k - \epsilon_{k,i_k} g^{k,i_k})$ due to line 11 of Algorithm 2, we obtain

$$\begin{aligned}
f(x^{k+1}) - f(\hat{x}^k) &\leq \|g^{k,i_k}\|^2 \left((1 + L\beta/2)\gamma^2 - \frac{\epsilon_{k,i_k}}{2}(2 - L\beta\epsilon_{k,i_k})\right) \\
&\overset{(28)}{\leq} \frac{f(x^k) - f(x^{k+1})}{\alpha\,\eta_k}\left((1 + L\beta/2)\gamma^2 - \frac{\epsilon_{k,i_k}}{2}(2 - L\beta\epsilon_{k,i_k})\right) \\
&\leq \left(f(x^k) - f(x^{k+1})\right)\left((1 + L\beta/2)\frac{\gamma^2}{\alpha\underline{\eta}} - \frac{2 - L\beta\underline{\epsilon}}{2\alpha}\right).
\end{aligned}$$

The last inequality uses the bounds $\eta_k \geq \epsilon_{k,i_k} \geq \underline{\eta}$ and $\epsilon_{k,i_k} \leq \epsilon_{k,0} = \underline{\epsilon}$ for $k \geq \underline{k}$. Rearranging terms, we arrive at the recursion:

$$f(x^{k+1}) - f(\hat{x}^k) \leq q\left(f(x^k) - f(\hat{x}^k)\right),$$

where

$$q \triangleq 1 - \left((1 + L\beta/2)\frac{\gamma^2}{\alpha\,\underline{\eta}} - \frac{2 - L\beta\underline{\epsilon}}{2\alpha} + 1\right)^{-1} \in (0, 1).$$

Observe that $\|\hat{x}^k - \bar{x}\| \leq \|\hat{x}^k - x^k\| + \|x^k - \bar{x}\| \leq 2\|x^k - \bar{x}\| < \delta_0$. Since both $\hat{x}^k$ and $\bar{x}$ are stationary points, Assumption 1 implies that $f(\hat{x}^k) = f(\bar{x})$. Consequently, the recursion simplifies to, for all $k \geq \underline{k}$ such that $x^k \in \mathbb{B}_{\min\{\tau, \delta_0/2\}}(\bar{x})$,

$$f(x^{k+1}) - f(\bar{x}) \leq q\left(f(x^k) - f(\bar{x})\right). \tag{30}$$

We next show that once x^k is sufficiently close to $\bar{x}$, all subsequent iterates remain within the neighborhood. Let $k_1 \geq \underline{k}$ be such that $x^{k_1} \in \mathbb{B}_{\min\{\tau, \delta_0/2\}}(\bar{x})$. Suppose, for the sake of contradiction, that a later iterate exits this neighborhood. Then there exists $k_2 \geq k_1 + 1$ as the smallest index such that $x^{k_2} \notin \mathbb{B}_{\min\{\tau, \delta_0/2\}}(\bar{x})$. We can estimate the distance between x^{k_1} and x^{k_2} as follows:

$$\begin{aligned}
\|x^{k_2} - x^{k_1}\| &\leq \sum_{k=k_1}^{k_2-1}\|x^{k+1} - x^k\| \overset{(28)}{\leq} \sum_{k=k_1}^{k_2-1}\sqrt{\eta_k\left(f(x^{k+1}) - f(x^k)\right)/\alpha} \\
&\overset{(30)}{\leq} \sqrt{\underline{\epsilon}\left(f(x^{k_1+1}) - f(\bar{x})\right)/\alpha}\sum_{k=k_1}^{k_2-1}q^{(k-k_1)/2},
\end{aligned}$$

25

which can be made arbitrarily small by choosing k_1 sufficiently large, implying that $x^{k_2} \in \mathbb{B}_{\min\{\tau, \delta_0/2\}}(\bar{x})$, a contradiction. Thus, there exists an index r such that $x^k \in \mathbb{B}_{\min\{\tau, \delta_0/2\}}(\bar{x})$ for all $k \geq r$. Consequently, the geometric decay in (30) holds for all $k \geq r$, and the function values $f(x^k)$ asymptotically converge linearly. Finally, the inequalities (28) and (30), combined with $\|x^k - \bar{x}\| \leq \sum_{i=k}^{+\infty} \|x^i - x^{i+1}\|$, allow us to conclude that the iterates x^k converge linearly to $\bar{x}$ for $k \geq r$. $\qquad \square$

6 Numerical Experiments

In this section, we evaluate the performance of our algorithms on several classes of nonsmooth functions. All experiments were conducted in MATLAB 2023b on a desktop equipped with an Intel Core i7-13700 CPU and 32 GB of RAM. The code is available at `https://github.com/lhyoung99/subgradient-regularization`. For each test example, we consider two types of oracles:

Oracle 1: Given x, return the objective $f(x)$ and a subgradient $v \in \partial f(x)$.

Oracle 2: Given x and $\epsilon > 0$, return the objective $f(x)$ and a descent-oriented subgradient $g \in G(x, \epsilon)$.

The descent-oriented subdifferential G in Oracle 2 is constructed by solving a subgradient-regularized problem, which reduces to a convex quadratic program (see Section 4 for details). Among several solvers tested, MATLAB's built-in function *quadprog* was found to be the most efficient for these subproblems.

When Algorithms 1 and 2 are equipped with the subgradient regularization to generate $G(x, \epsilon)$, we call them SRDescent and SRDescent-adapt, respectively. We summarize the parameter setting used in these two algorithms.

Regularization parameters. We set the initial regularization parameter to $\epsilon_{0,0} = 5$ and apply a reduction factor $\theta_\epsilon = 0.9$. In Algorithm 2, we use the non-summable sequence $\{a_t = 1/t\}$ for the regularization parameter $\tilde{\epsilon}_k$.

Stationary tolerance parameters. The initial target for the stationarity measure is set to $\nu_0 = 10^{-2}$, with a reduction factor $\theta_\nu = 0.5$.

Line search parameter. We use $\alpha = 10^{-4}$ as the Armijo parameter.

For comparison, we consider the following algorithms, all relying on Oracle 1:

Polyak: The subgradient method with Polyak stepsizes that iterates as

$$x^{k+1} = x^k - \frac{f(x^k) - f^*}{\|g^k\|^2} g^k \quad \text{for some } g^k \in \partial f(x^k),$$

where f^* is the optimal objective value.

PBMDC: A proximal bundle method for difference-of-convex functions [15], available at `https://www.oliveira.mat.br/solvers`.

NTDescent: Originally implemented in PyTorch `https://github.com/COR-OPT/ntd.py`, we re-implemented this method in MATLAB to ensure fair runtime comparison. Across several test cases under the same termination conditions, the MATLAB version matches the iteration count of the original implementation while being faster in CPU time.

HANSO: A hybrid nonsmooth optimization algorithm combining BFGS [32] with gradient sampling (GS) [8], available at `https://cs.nyu.edu/~overton/software/hanso/`. We separately

tested GS as an implementable Goldstein-type method with convergence guarantees, and BFGS as a heuristic known for good empirical performance. In many examples, GS suffers from the line search failures, which, as noted in [8], occur when either the maximum number of backtracking steps (set to 50) is reached or the computed direction is not a descent direction.

Unless otherwise specified, all parameters were set to their default values in these algorithms. Performance is evaluated primarily in terms of CPU time and the number of calls to Oracle 1 and Oracle 2, respectively.

6.1 Finite max of convex quadratic functions

We first test our algorithms on the finite max of convex quadratic functions:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \left\{ f(x) = \max_{1 \le i \le m} \left(g_i^\top x + \frac{1}{2} x^\top H_i\, x \right) \right\},$$

where each $H_i \in \mathbb{R}^{n \times n}$ is a randomly generated positive semi-definite matrix, and $\{g_i\}_{i=1}^m$ are random vectors satisfying $\sum_{i=1}^{\lfloor m/2 \rfloor} \lambda_i\, g_i = 0$ and $\sum_{j=1}^{m} \mu_j\, g_j = 0$ for some vector $\lambda \in \mathbb{R}^{\lfloor m/2 \rfloor}$ with $\lambda > 0$ and $\sum_{i=1}^{\lfloor m/2 \rfloor} \lambda_i = 1$, and $\mu \in \mathbb{R}^m$ with $\mu \neq 0$ and $\sum_{j=1}^{m} \mu_j = 0$. Under this setup, $x^* = 0$ is the unique minimizer, and all assumptions in Corollary 2 hold. Thus, by Corollary 2, SRDescent-adapt is expected to converge linearly. Meanwhile, the affine independence and non-degeneracy requirements in [25, 14, 28] do not hold here, so NTDescent and GS lack theoretical linear convergence guarantees for these instances.
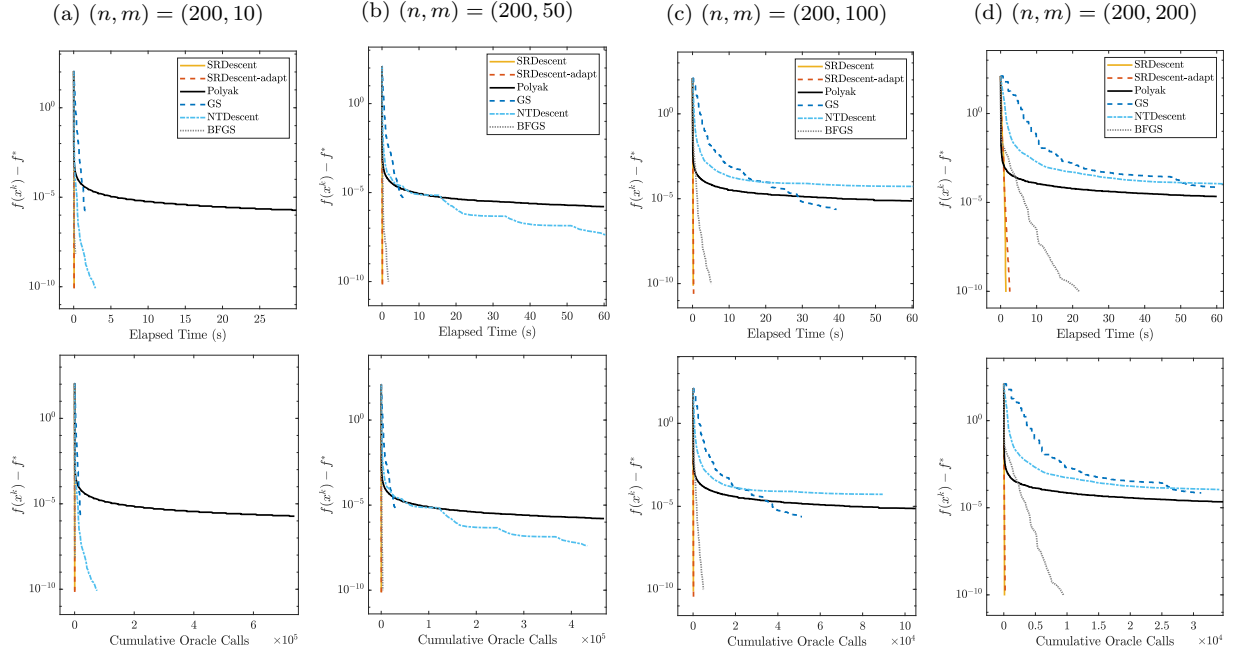

Figure 3: Performance on the max of convex quadratics functions with $n = 200$ and $m \in \{10, 50, 100, 200\}$, all initialized from the same random points. For the Polyak method, we plot the best objective value after k oracle calls, i.e., $\min_{1 \le i \le k} f(x^k)$ rather than $f(x^k)$. In cases $m \in \{10, 50, 100\}$, GS terminates early due to the line search failures.

We generate test problems by fixing $n = 200$ and varying the number of convex quadratic functions $m \in \{10, 50, 100, 200\}$. Figure 3 reports the results. We excluded PBMDC because of its unstable behavior, converging quickly to moderate accuracy but significantly slow afterward; for $(n, m) = (200, 50)$, it already becomes slow near 10^{-7} and behaves worse as m increases. Polyak method converges sublinearly, serving as a baseline. Both SRDescent and SRDescent-adapt exhibit linear convergence and outperform other methods, including BFGS, for larger m. Goldstein-type methods struggle when m grows, possibly due to the bottleneck of sampling more pieces of functions to approximate the Goldstein direction. In addition, GS encounters numerical issues in several cases due to the line search failures, making it less reliable for producing high quality solutions.

6.2   Finite max of nonconvex quadratic functions

We consider Nesterov's nonsmooth Chebyshev-Rosenbrock function:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \left\{ f(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1} \left| x_{i+1} - 2(x_i)^2 + 1 \right| \right\}, \tag{31}$$

which is known for its challenging landscape [23]. This nonconvex function has a unique global minimizer $x^* = (1, 1, \cdots, 1)$, which is also the only stationary point [23, Theorem 1]. Although f is partly smooth with respect to the manifold $\mathcal{M} \triangleq \{ x \mid x_{i+1} - 2(x_i)^2 + 1 = 0, i = 1, \cdots, n-1 \}$, it is not an active manifold around x^* in the sense of [14] and therefore, it is not clear whether NTDescent locally converges nearly linearly for this problem.

As numerically shown in [23], this problem becomes increasingly difficult as n increases. For large n, methods like BFGS result in low accuracy because line search breaks down near $\mathcal{M}$ but not close enough to $\bar{x}$. In our experiments, each algorithm is run until it either achieves an objective gap $f(x^k) - f^*$ below an accuracy of 10^{-2} or 10^{-5}, or exceeds a time limit of $1,000$ seconds. For each dimension n, we test 10 random initial points and consider any run that does not meet the target accuracy as a failure. Table 1 summarizes the detailed results. When $n = 8$ and 10, all methods except SRDescent and SRDescent-adapt frequently fail to reach the high accuracy within the time limit, while our methods are more robust against the challenges posed by this nonsmooth landscape. Additionally, we can see from Table 1 that SRDescent-adapt consistently converges faster than SRDescent, benefiting from the adaptive adjustment of the regularization parameter $\epsilon_{k,0}$.

Beyond algorithms based on Oracle 1 in a black-box setting, we also compare with LiPsMin [19], which explicitly uses the kink structure to build piecewise linear approximations. The results are summarized in Table 2, where the results of LiPsMin are taken from [19]. SRDescent and SRDescent-adapt achieve lower objective values with a comparable number of gradient evaluations as LiPsMin. For $n = 20$, our methods using additional objective and gradient evaluations successfully escape the "pseudo stationary value" of 0.81814 observed in [19] for both LiPsMin and HANSO.

n	method	low accuracy 1e-2				high accuracy 1e-5			
		#fails	time (s)	$f(x^k)$	oracle calls	#fails	time (s)	$f(x^k)$	oracle calls
3	PBMDC	0	5.6e-2	7.9e-3	6.4e+1	0	6.0e-1	9.9e-6	4.6e+2
	GS	0	1.1e-2	8.8e-3	5.5e+2	0	1.5e-1	9.9e-6	1.6e+4
	NTDescent	0	5.7e-3	8.4e-3	2.5e+3	0	2.4e-2	8.2e-6	3.1e+4
	BFGS	0	7.7e-3	9.6e-3	4.0e+2	0	5.3e-2	9.9e-6	2.9e+3
	SRDescent	0	6.1e-3	7.3e-3	4.9e+1	0	5.4e-2	1.0e-5	7.2e+2
	SRDescent-adapt	0	8.2e-3	7.8e-3	1.0e+2	0	5.1e-2	9.9e-6	7.1e+2
5	PBMDC	0	1.6e+0	8.0e-3	1.4e+3	10	— (1.0e+3)	— (1.9e-4)	— (5.1e+4)
	GS	0	1.1e-1	9.8e-3	1.5e+4	6	6.6e+0 (4.9e+0)	1.0e-5 (1.6e-5)	9.8e+5 (7.1e+5)
	NTDescent	0	1.0e-1	9.4e-3	1.7e+5	0	1.1e+2	9.5e-6	2.0e+8
	BFGS	1	3.5e-1 (5.3e-1)	8.9e-3 (1.4e-1)	1.6e+4 (2.4e+4)	10	— (2.3e+0)	— (1.7e-2)	— (8.2e+4)
	SRDescent	0	3.5e-2	7.9e-3	5.5e+2	0	1.5e+0	9.1e-6	2.4e+4
	SRDescent-adapt	0	5.1e-2	7.9e-3	8.5e+2	0	2.4e+0	9.1e-6	4.7e+4
8	PBMDC	8	2.4e-2 (1.0e+3)	7.5e-3 (1.6e-2)	3.6e+1 (2.9e+5)	10	— (1.0e+3)	— (1.4e-2)	— (2.4e+5)
	GS	0	9.9e+0	9.7e-3	1.7e+6	10	— (5.5e+1)	— (9.2e-4)	— (8.1e+6)
	NTDescent	0	2.6e+2	9.2e-3	4.6e+8	9	1.5e-2 (1.0e+3)	9.9e-6 (2.6e-3)	1.6e+4 (1.7e+9)
	BFGS	6	1.2e-2 (8.9e+0)	7.9e-3 (2.8e-1)	7.5e+1 (2.6e+5)	10	— (6.8e+0)	— (1.7e-1)	— (1.8e+5)
	SRDescent	0	1.7e+0	7.4e-3	2.6e+4	0	1.1e+2	9.2e-6	1.7e+6
	SRDescent-adapt	0	2.2e+0	7.4e-3	4.0e+4	0	4.0e+1	9.2e-6	6.2e+5
10	PBMDC	10	— (1.0e+3)	— (1.3e-1)	— (6.5e+4)	10	— (1.0e+3)	— (1.3e-1)	— (6.4e+4)
	GS	7	6.6e+1 (1.1e+2)	9.1e-3 (1.9e-2)	1.2e+7 (1.9e+7)	10	— (1.0e+2)	— (1.5e-2)	— (1.8e+7)
	NTDescent	7	1.2e-2 (1.0e+3)	8.6e-3 (4.2e-1)	3.3e+3 (1.7e+9)	10	— (1.0e+3)	— (3.0e-1)	— (1.7e+9)
	BFGS	6	5.5e-3 (1.0e+1)	9.1e-3 (3.0e-1)	1.0e+2 (2.5e+5)	10	— (7.6e+0)	— (1.8e-1)	— (1.8e+5)
	SRDescent	0	3.0e+1	7.8e-3	4.7e+5	8	4.0e+2 (1.0e+3)	6.6e-6 (5.5e-5)	6.1e+6 (1.5e+7)
	SRDescent-adapt	0	2.4e+1	7.8e-3	4.1e+5	0	6.4e+2	9.3e-6	9.5e+6

Table 1: Comparison on Nesterov's nonsmooth Chebyshev-Rosenbrock function (31) over 10 random initial points. The column "#fails" counts the number of runs (out of 10) that do not reach the target accuracy within 1,000 seconds, or terminate early due to the line search failures. For successful runs, the average runtime, final objective values, and number of oracle calls are reported without parentheses; parenthetical values represent averages over failed runs.

method	$n = 5$			$n = 10$			$n = 20$		
	$f(x^k)$	#f	#∇f	$f(x^k)$	#f	#∇f	$f(x^k)$	#f	#∇f
LiPsMin	6.4e-2	1001	15095	0.81744	1001	30371	0.81814	1001	42150
SRDescent	5.4e-5	3949	15092	0.80087	7768	30368	0.81814	2218	42142
	2.4e-7	1.3e+5	5.0e+5	2.7e-5	1.1e+7	1.0e+8	0.81804	5.3e+7	1.0e+9
SRDescent-adapt	2.1e-5	3975	15092	0.7914	7792	30368	0.81814	2204	42142
	4.2e-10	1.3e+5	5.0e+5	2.6e-6	1.1e+7	1.0e+8	0.81773	5.3e+7	1.0e+9

Table 2: Comparison on Nesterov's nonsmooth Chebyshev-Rosenbrock function (31) with a fixed initial point x^0, where $x_i^0 = 0.5$ for odd i and $x_i^0 = -0.5$ otherwise. "#f" and "#∇f" denote the number of objective and gradient evaluations. For SRDescent(-adapt), the first row reports the final objective using a comparable number of gradient evaluations as LiPsMin, while the second row shows results obtained with additional objective and gradient evaluations.

6.3 Finite min of convex quadratic functions

In this part, we test our algorithms on the finite min of convex quadratic functions:

$$f(x) = \min_{1 \leq i \leq m} \frac{1}{2}\|A_i\, x - b_i\|^2,$$

where each $A_i \in \mathbb{R}^{d \times n}$ with $d \geq n$ has linearly independent columns, and $b_i = A_i\, x^*$ for a solution $x^* \sim \mathcal{N}\big(0, n^{-1/2}I_n\big)$. By construction, x^* is a global minimizer with $f^* = 0$, which is also the unique stationary point. This function is not subdifferentially regular [43, Definition 7.25] at points where multiple quadratic components are active, creating the so-called "downward cusps".

We fix $n = d = 300$ and vary the number of pieces m. Each algorithm is tested from the same random initial point. Figure 4 summarizes run time and oracle calls averaged over 100 random instances, and Figure 5 plots results on a random instance for $m \in \{10, 50, 100\}$. SRDescent(-adapt) and BFGS show the best overall performance. Their oracle calls required for achieving a target accuracy of 10^{-8} remain stable across different m. However, BFGS tends to be more efficient than SRDescent(-adapt) as m increases to 350. While Polyak and GS perform well for small m, their results start to have high deviations for large m. In Figure 6, we also include an additional comparison for $m = 50$ and varying dimensions n.

We remark that although the finite min of quadratic functions is not subdifferentially regular near the minimizer x^*, gradients of all component quadratic functions vanish at this point. This suggests that, unlike in the case of the finite max of quadratics, combining gradients of all active components may be unnecessary. Nonetheless, the observed slow convergence of NTDescent and GS indicates additional challenges inherent to this class of nonsmooth functions. Further investigation into methods for handling non-regular functions is left for future work.
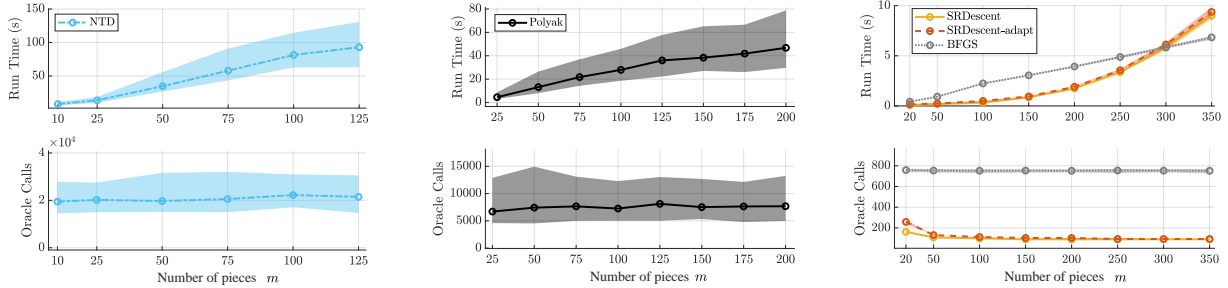

Figure 4: Run time and oracle calls on the finite min of convex quadratic functions over 100 randomly generated instances with $n = d = 300$ and varying numbers of pieces m. The lines represent the median and shaded areas indicate inter-quartiles. NTDescent terminates when the objective gap $f(x^k) - f^*$ is below 10^{-6} and other methods terminate when the objective gap is below 10^{-8}.

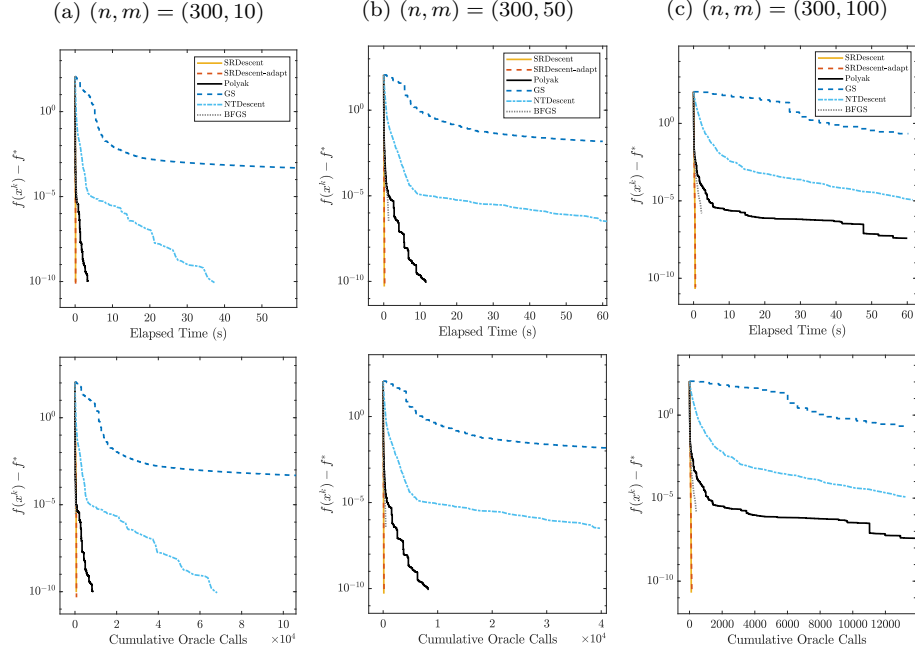(a) $(n, m) = (300, 10)$ (b) $(n, m) = (300, 50)$ (c) $(n, m) = (300, 100)$

Figure 5: Performance on the finite min of convex quadratic functions for $n = d = 300$ and varying numbers of pieces $m \in \{10, 50, 100\}$, all initialized from the same randomly generated points. For Polyak, we report the best objective value after k oracle calls, i.e., $\min_{1 \le i \le k} f(x^k)$, rather than $f(x^k)$. For all cases, BFGS terminates early due to the line search failures.

(a) $(n, m) = (100, 50)$ (b) $(n, m) = (250, 50)$ (c) $(n, m) = (500, 50)$ (d) $(n, m) = (1000, 50)$
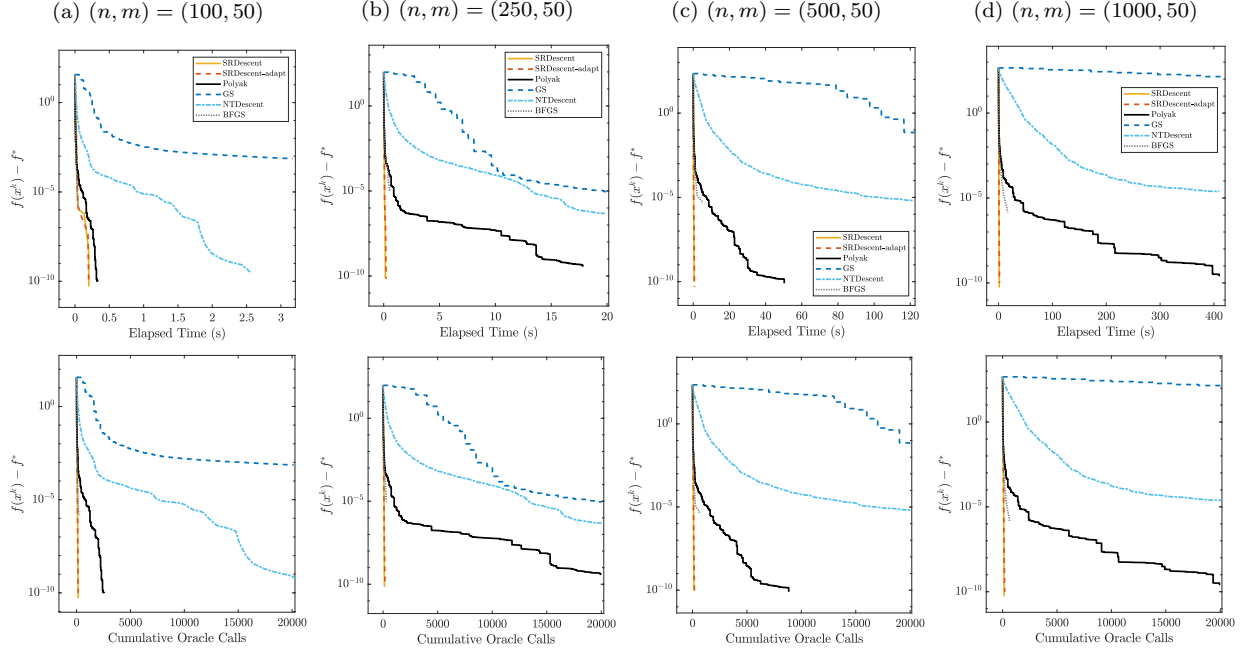
Figure 6: Performance on the finite min of convex quadratic functions for a fixed numbers of pieces $m = 50$ and varying dimensions $n \in \{100, 250, 500, 1000\}$, all initialized from the same randomly generated points. For Polyak, we report the best objective value after k oracle calls, i.e., $\min_{1 \le i \le k} f(x^k)$, rather than $f(x^k)$. For all cases, BFGS terminates early due to the line search failures.

6.4 Marginal functions with varying feasible sets

In the final example, we study the minimization of a marginal function given by:

$$f(x) = \min_{y \in \mathbb{R}^m} \left\{ (c + Dx)^\top y + \frac{1}{2}\, y^\top Q y + \|x\|^4 \right\}$$
$$\text{subject to } b - Ax - \mathbf{1} \le Wy \le b - Ax\,.$$

Here, the vectors and matrices are generated randomly: the entries of $b, c, A,$ and D are independently drawn from $\mathcal{N}(0, m^{-1/2})$, the matrix $W \in \mathbb{R}^{m \times m}$ has linearly independent rows, and Q is positive semi-definite. The vector $\mathbf{1}$ denotes the m-dimensional all-ones vector. Under this setup, $(-f)$ satisfies assumptions (i)-(ii), and (iii$'$) LICQ in Section 4.2. Without the term $\|x\|^4$, the marginal function is known to be piecewise linear-quadratic on its domain of finiteness [12, Proposition 5.2.2]. The additional term $\|x\|^4$ ensures that f is bounded from below.

We fix $n = 300$ and vary the dimension m of the inner minimization problem. We run SRDescent with optimality tolerances $\epsilon_{\text{tol}} = \nu_{\text{tol}} = 10^{-3}$. Figure 7 presents the results for run time and oracle calls. We focus solely on SRDescent because different algorithms converge to different stationary values, complicating direct comparisons. Additionally, we observe that SRDescent-adapt exhibits performance nearly identical to that of SRDescent. As shown in Figure 7, increasing m leads to a notable increase in run time and oracle calls. This indicates that the problem becomes more challenging as the dimension of the inner minimization grows.
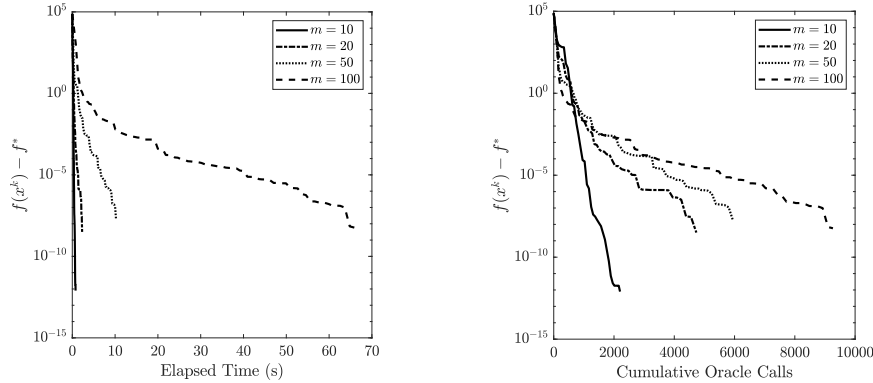

Figure 7: Run time and oracle calls of SRDescent on marginal functions with varying feasible sets for a fixed dimension $n = 300$ and varying $m \in \{10, 20, 50, 100\}$. For each instance, we use the final objective value as an estimation of f^* to compute the objective gap.

7 Conclusions

In this work, we proposed a unifying principle for designing convergent descent methods in nonsmooth optimization through the concept of descent-oriented subgradients. This framework reveals the common structure of Goldstein-type methods and bundle methods, and motivates a novel technique for constructing descent directions, which we term subgradient regularization. Our work opens several promising directions for future research. A key question is to develop computationally efficient descent-oriented subdifferentials for broader classes of nonsmooth functions. Additionally,

extending linear convergence guarantees under regularity conditions to more general settings, as well as establishing iteration complexity bounds, are important topics for further exploration.

References

[1] Edgar Asplund and RT Rockafellar. Gradients of convex functions. *Transactions of the American Mathematical Society*, 139:443–467, 1969.

[2] Felipe Atenas, Claudia Sagastizábal, Paulo JS Silva, and Mikhail Solodov. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1):89–115, 2023.

[3] J Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Science & Business Media, 2013.

[4] James V Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33(3):260–279, 1985.

[5] James V Burke, Frank E Curtis, Adrian S Lewis, Michael L Overton, and Lucas EA Simões. Gradient sampling methods for nonsmooth optimization. *Numerical Nonsmooth Optimization: State of the Art Algorithms*, pages 201–225, 2020.

[6] James V Burke and Michael C Ferris. A Gauss-Newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[7] James V Burke, Adrian S Lewis, and Michael L Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002.

[8] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[9] Xiaojun Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming*, 134:71–99, 2012.

[10] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[11] Frank H Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[12] Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, 2021.

[13] 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. *Advances in Neural Information Processing Systems*, 35:6692–6703, 2022.

[14] Damek Davis and Liwei Jiang. A local nearly linearly convergent first-order method for nonsmooth functions with quadratic growth. *Foundations of Computational Mathematics*, 2024.

[15] Welington de Oliveira. Proximal bundle methods for nonsmooth DC programming. *Journal of Global Optimization*, 75:523–563, 2019.

[16] Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.

[17] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[18] Ivar Ekeland and Roger Temam. *Convex Analysis and Variational Problems*. SIAM, 1999.

[19] Sabrina Fiege, Andrea Walther, and Andreas Griewank. An algorithm for nonsmooth optimization by successive piecewise linearization. *Mathematical Programming*, 177:343–370, 2019.

[20] Roger Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Mathematical Programming Studies*, 17:67–76, 1982.

[21] Jacques Gauvin and François Dubeau. Differential properties of the marginal function in mathematical programming. *Optimality and Stability in Mathematical Programming*, pages 101–119, 1982.

[22] Allen A Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.

[23] Mert Gürbüzbalaban and Michael L Overton. On Nesterov's nonsmooth Chebyshev–Rosenbrock functions. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1282–1289, 2012.

[24] Warren Hare and Claudia Sagastizábal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.

[25] Elias Salomão Helou, Sandra A Santos, and Lucas EA Simões. On the local convergence analysis of the gradient sampling method for finite max-functions. *Journal of Optimization Theory and Applications*, 175:137–157, 2017.

[26] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I and II*, volume 305 and 306. Springer science & business media, 1996.

[27] Krzysztof C Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

[28] Siyu Kong and Adrian S Lewis. Lipschitz minimization and the Goldstein modulus. *arXiv preprint arXiv:2405.12655*, 2024.

[29] Claude Lemaréchal. An extension of Davidon methods to nondifferentiable problems. In *Nondifferentiable Optimization*, pages 95–109. Springer, 1975.

[30] Claude Lemaréchal and François Oustry. Nonsmooth algorithms to solve semidefinite programs. In *Advances in linear matrix inequality methods in control*, pages 57–77. SIAM, 2000.

[31] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[32] Adrian S Lewis and Michael L Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141:135–163, 2013.

[33] Adrian S Lewis and Stephen J Wright. A proximal method for composite minimization. *Mathematical Programming*, 158:501–546, 2016.

[34] Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[35] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[36] Robert Mifflin. A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In *Nondifferential and Variational Techniques in Optimization*, pages 77–90. Springer, Berlin, Heidelberg, 1982.

[37] Robert Mifflin and Claudia Sagastizábal. On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. *SIAM Journal on Optimization*, 11(2):547–571, 2000.

[38] Robert Mifflin and Claudia Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Mathematical programming*, 104:583–608, 2005.

[39] 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.

[40] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

[41] Stephen M Robinson. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. *Mathematical Programming*, 86:41–50, 1999.

[42] R Tyrrell Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

[43] R Tyrrell Rockafellar and Roger JB Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, New York, 2009.

[44] 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 *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

[45] P Tseng. Descent methods for convex essentially smooth minimization. *Journal of optimization theory and applications*, 71(3):425–463, 1991.

[46] Philip Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable Optimization*, pages 145–173. Springer, 1975.

[47] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

A Omitted Proofs in Section 4

A.1 Proof of Lemma 2

Before proving Lemma 2, we first establish a technical result that quantifies the objective value and gradient norm associated with a solution to the subgradient-regularized problem (P^ϵ).

Lemma 7. Consider problem (P), and let $Y^\epsilon(x)$ be defined as in (P^ϵ). Let $y^\epsilon : \mathbb{R}^n \to \mathbb{R}^m$ be a single-valued map with $y^\epsilon(x) \in Y^\epsilon(x)$ for all $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$. Then for all such (x, ϵ),

(a) (Error estimate for the objective value)

$$f(x) - \frac{\epsilon}{2} \inf_{y \in Y^*(x)} \|\nabla_x \varphi(x, y)\|^2 \leq \varphi(x, y^\epsilon(x)) \leq f(x).$$

(b) (Bound on the gradient norm)

$$\|\nabla_x \varphi(x, y^\epsilon(x))\| \leq \inf_{y \in Y^*(x)} \|\nabla_x \varphi(x, y)\|.$$

Proof. (a) The upper bound $\varphi(x, y^\epsilon(x)) \leq f(x)$ follows immediately from the definition of $f(x)$. For the lower bound, we use the optimality of $y^\epsilon(x) \in Y^\epsilon(x)$. For any $y \in Y^*(x)$, we have

$$\varphi(x, y) - \frac{\epsilon}{2}\|\nabla_x \varphi(x, y)\|^2 \leq \varphi(x, y^\epsilon(x)) - \frac{\epsilon}{2}\|\nabla_x \varphi(x, y^\epsilon(x))\|^2 \leq \varphi(x, y^\epsilon(x)). \qquad (32)$$

Taking the supremum over $y \in Y^*(x)$ on the left-hand side yields

$$f(x) - \frac{\epsilon}{2} \inf_{y \in Y^*(x)} \|\nabla_x \varphi(x, y)\|^2 \leq \varphi(x, y^\epsilon(x)).$$

Here, the infimum is attained and finite since $Y^*(x)$ is compact and $\nabla_x \varphi(x, \cdot)$ is continuous.

(b) For any $y \in Y^*(x)$, combining the inequality $\varphi(x, y^\epsilon(x)) \leq f(x)$ with (32), we have

$$f(x) - \frac{\epsilon}{2}\|\nabla_x \varphi(x, y)\|^2 = \varphi(x, y) - \frac{\epsilon}{2}\|\nabla_x \varphi(x, y)\|^2 \leq f(x) - \frac{\epsilon}{2}\|\nabla_x \varphi(x, y^\epsilon(x))\|^2.$$

Thus, $\|\nabla_x \varphi(x, y^\epsilon(x))\| \leq \|\nabla_x \varphi(x, y)\|$. Taking the infimum over $y \in Y^*(x)$ completes the proof. $\qquad \square$

We now proceed to prove Lemma 2, using the results of Lemma 7.

Proof of Lemma 2. (a) Based on Lemma 7(a), if we can further show that

$$\limsup_{x \to \bar{x}} \left(\inf_{y \in Y^*(x)} \|\nabla_x \varphi(x, y)\|^2 \right) < +\infty, \qquad (33)$$

the proof is then completed by taking the limits as $\epsilon \downarrow 0$ and $x \to \bar{x}$ and using the continuity of f Observe that $\inf_{y \in Y^*(x)} \|\nabla_x \varphi(x, y)\| \leq \sup_{v \in \partial f(x)} \|v\|$ because $\partial f(x) = \text{conv}\left(\bigcup\{\nabla_x \varphi(x, y) \mid y \in Y^*(x)\}\right)$. From [11, Proposition 2.1.2], the term $\sup_{v \in \partial f(x)} \|v\|$ is upper bounded by the local

Lipschitz constant of f at x. Thus, (33) is a direct consequence of the locally Lipschitz continuity of f at $\bar{x}$, and (a) follows.

(b) Let $\{\epsilon_k\} \downarrow 0$ and $\{x^k\} \to \bar{x}$. We aim to show that the sequence $\{\nabla_x \varphi(x^k, y^{\epsilon_k}(x^k))\}$ is bounded and that any of its accumulation points belongs to $\partial f(\bar{x})$. Boundedness follows immediately from Lemma 7(b) and our assumption that $\nabla_x \varphi(\cdot, \cdot)$ is continuous. Now consider a convergent subsequence $\{\nabla_x \varphi(x^k, y^{\epsilon_k}(x^k))\}_{k \in N}$ for some index set $N \in \mathbb{N}_\infty^\sharp$. Since Y is compact, the corresponding subsequence $\{y^{\epsilon_k}(x^k)\}_{k \in N} \subset Y$ has at least one accumulation point $\bar{y} \in Y$. Let $y^{\epsilon_k}(x^k) \to_{N'} \bar{y}$ along a further subsequence indexed by $N' \subset N$. By the continuity of $\varphi(\cdot, \cdot)$ and part (a), we have $\varphi(x^k, y^{\epsilon_k}(x^k)) \to_{N'} \varphi(\bar{x}, \bar{y}) = f(\bar{x})$, so $\bar{y} \in Y^*(\bar{x})$. Using the continuity of $\nabla_x \varphi(\cdot, \cdot)$, we obtain

$$\lim_{k(\in N) \to +\infty} \nabla_x \varphi(x^k, y^{\epsilon_k}(x^k)) = \lim_{k(\in N') \to +\infty} \nabla_x \varphi(x^k, y^{\epsilon_k}(x^k)) = \nabla_x \varphi(\bar{x}, \bar{y}) \in \partial f(\bar{x}).$$

We can conclude that any accumulation point of $\{\nabla_x \varphi(x^k, y^{\epsilon_k}(x^k))\}$ is an element of $\partial f(x)$. This completes the proof of part (b).

(c) Fix $\epsilon > 0$ and take any sequence $\{x^k\} \to \bar{x}$. By Lemma 7(b) and the continuity of $\nabla_x \varphi(\cdot, \cdot)$, the sequence $\{\nabla_x \varphi(x^k, y^\epsilon(x^k))\}$ is bounded. To establish the inclusion

$$\limsup_{k \to +\infty} \nabla_x \varphi(x^k, y^\epsilon(x^k)) \subset \bigcup \{\nabla_x \varphi(\bar{x}, y) \mid y \in Y^\epsilon(\bar{x})\}, \tag{34}$$

it suffices to show that every accumulation point of the sequence $\{y^\epsilon(x^k)\}$ belongs to $Y^\epsilon(\bar{x})$. Recall that

$$y^\epsilon(x) \in Y^\epsilon(x) = \operatorname*{argmax}_{y \in Y} \left\{ \varphi(x, y) - \frac{\epsilon}{2} \|\nabla_x \varphi(x, y)\|^2 \right\}.$$

By assumption, both $\varphi(\cdot, \cdot)$ and $\nabla_x \varphi(\cdot, \cdot)$ are continuous. Hence, the objective function of the parametric maximization above is continuous in x for every fixed $y \in Y$. It then follows from [43, Theorem 1.17] that any accumulation point of $\{y^\epsilon(x^k)\}$ must lie in $Y^\epsilon(\bar{x})$.

Observe that (34) holds for any single-valued selection $y^\epsilon : \mathbb{R}^n \to \mathbb{R}^m$ satisfying $y^\epsilon(x) \in Y^\epsilon(x)$ for all $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$. It follows that $\limsup_{x \to \bar{x}} G(x, \epsilon) \subset G(\bar{x}, \epsilon)$; equivalently, $\limsup_{x \to \bar{x}} G(x, \epsilon) = G(\bar{x}, \epsilon)$. Thus, G satisfies the first condition in (G2′).

(d) Suppose that $S(\bar{x})$ is a convex set. It is immediate from (14) that $\partial f(\bar{x}) = \bigcup \{\nabla_x \varphi(\bar{x}, y) \mid y \in Y^*(x)\}$. Take any sequence $\{\epsilon_k\} \downarrow 0$. By Lemma 7, we know that

$$\|\nabla_x \varphi(\bar{x}, y^{\epsilon_k}(\bar{x}))\| \leq \inf_{y \in Y^*(\bar{x})} \|\nabla_x \varphi(\bar{x}, y)\| = \inf_{v \in \partial f(\bar{x})} \|v\|,$$

so the sequence $\{\nabla_x \varphi(\bar{x}, y^{\epsilon_k}(\bar{x}))\}$ is bounded and has at least one accumulation point. Additionally, by part (b), every accumulation point of this sequence belongs to $\partial f(\bar{x})$. Moreover, the previous bound $\|\nabla_x \varphi(\bar{x}, y^{\epsilon_k}(\bar{x}))\| \leq \inf_{v \in \partial f(\bar{x})} \|v\|$ ensures that the only possible accumulation point is $\operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$. Therefore, $\{\nabla_x \varphi(\bar{x}, y^{\epsilon_k}(\bar{x}))\}$ is a bounded sequence with a unique accumulation point. Finally, if the sequence did not converge, it would have at least two distinct accumulation points, contradicting uniqueness of the minimal norm subgradient. Thus, $\nabla_x \varphi(\bar{x}, y^{\epsilon_k}(\bar{x})) \to \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$, and the result follows because

$$\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \lim_{\epsilon \downarrow 0} \bigcup \{\nabla_x \varphi(\bar{x}, y^\epsilon(\bar{x})) \mid y^\epsilon(\bar{x}) \in Y^\epsilon(\bar{x})\} = \operatorname*{argmin}_{v \in \partial f(\bar{x})} \|v\|.$$

$\square$

A.2 Proof of Lemma 4

Proof. (a) First, it is easy to see that for each fixed $\epsilon > 0$, $G(\cdot, \epsilon)$ is closed-valued and locally bounded since each $\nabla_x \varphi_j$ is continuous and $Y(\cdot)$ is locally bounded. We proceed by presenting technical results parallel to Lemma 7. Let $y^\epsilon : \mathbb{R}^n \to \mathbb{R}^m$ be a single-valued map such that $y^\epsilon(x) \in Y^\epsilon(x)$ for all $(x, \epsilon) \in \mathbb{R}^n \times (0, +\infty)$. By the optimality of $y^\epsilon(x)$, we obtain the following estimate for the objective value:

$$f(x) - \frac{\epsilon}{2} \inf_{y \in Y^\epsilon(x)} \|\nabla_x \mathcal{L}(x, y; \lambda(x))\|^2 \leq \varphi_0(x, y^\epsilon(x)) \leq f(x), \tag{35}$$

along with a bound on the gradient norm:

$$\|\nabla_x \mathcal{L}(x, y^\epsilon(x); \lambda(x))\| \leq \inf_{y \in Y^*(x)} \|\nabla_x \mathcal{L}(x, y; \lambda(x))\| = \inf_{v \in \partial f(x)} \|v\|. \tag{36}$$

To verify property (G1), consider any sequences $\{x^k\} \to \bar{x}$ and $\{\epsilon_k\} \downarrow 0$. Note that the sequence $\{\nabla_x \mathcal{L}(x^k, y^{\epsilon_k}(x^k); \lambda(x^k))\}$ is bounded due to (36) and the local Lipschitz continuity of f. Since $y^{\epsilon_k}(x^k) \in Y(x^k)$, assumption (ii) ensures that $\{y^{\epsilon_k}(x^k)\}$ is bounded. Moreover, any accumulation point of $\{y^{\epsilon_k}(x^k)\}$ must lie in $Y^*(\bar{x})$ because $\varphi_0(x^k, y^{\epsilon_k}(x^k)) \to f(\bar{x})$ by (35). In addition, the multiplier sequence $\{\lambda(x^k)\}$ is bounded due to the MFCQ, and its accumulation point must be $\lambda(\bar{x})$ by the continuity of each $\nabla_y \varphi_j(\cdot, \cdot)$. Hence,

$$\limsup_{k \to +\infty} \nabla_x \mathcal{L}(x^k, y^{\epsilon_k}(x^k); \lambda(x^k)) \subset \bigcup \{\nabla_x \mathcal{L}(\bar{x}, y; \lambda(\bar{x})) \mid y \in Y^*(\bar{x})\} = \partial f(\bar{x}),$$

which implies property (G1); that is, $\limsup_{x \to \bar{x}, \epsilon \downarrow 0} G(x, \epsilon) \subset \partial f(\bar{x})$.

To show $\limsup_{x \to \bar{x}} G(x, \epsilon) = G(\bar{x}, \epsilon)$ for any fixed $\epsilon > 0$, consider any sequence $\{x^k\} \to \bar{x}$. The argument proceeds similarly to the one used above. The key difference is that, for fixed $\epsilon > 0$, any accumulation point of the sequence $\{y^\epsilon(x^k)\}$ must belong to $Y^\epsilon(\bar{x})$, which follows from the continuity of $\varphi(\cdot, \cdot)$ and each $\nabla_x \varphi_j(\cdot, \cdot)$.

Finally, to establish $\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$, we observe that property (G1) ensures $\limsup_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) \subset \partial f(\bar{x})$, while (36) guarantees that all vectors in $G(\bar{x}, \epsilon)$ have norm at most $\inf_{v \in \partial f(\bar{x})} \|v\|$. Combining these two results and using the uniqueness of the minimal norm subgradient, we can deduce that $\lim_{\epsilon \downarrow 0} G(\bar{x}, \epsilon) = \operatorname{argmin}_{v \in \partial f(\bar{x})} \|v\|$.

So far, we have verified properties (G1) and (G2$'$), and thus conclude that the mapping G defined in (19) is a descent-oriented subdifferential.

(b) The proof proceeds analogously under the assumption that $\varphi_0(x, \cdot)$ is strongly concave for each fixed x. In this case, both the error estimate for the objective value and the bound on the gradient norm follow from the same reasoning as in part (a). Properties (G1) and (G2$'$) can be verified by applying analogous boundedness and convergence arguments. Therefore, the mapping G defined in (20) is also a descent-oriented subdifferential in this case. □

B Omitted Proofs in Section 5.2.2

To proceed, consider the proximal mapping parameterized by x:

$$z^\epsilon(x) \triangleq \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ h\big(c(x) + \nabla c(x)^\top (z - x)\big) + \frac{1}{2\epsilon} \|z - x\|^2 \right\}, \quad \epsilon > 0. \tag{37}$$

The prox-linear update can then be expressed as $x^{k+1} = z^\epsilon(x^k)$. By Proposition 7, this yields

$$x^{k+1} = z^\epsilon(x^k) = x^k - \epsilon\, G(x^k, \epsilon) \quad \Rightarrow \quad G(x^k, \epsilon) = \frac{x^k - z^\epsilon(x^k)}{\epsilon}.$$

It is not difficult to see that any point $\bar{x} \in (\partial f)^{-1}(0)$ is the unique optimal solution of the strongly convex problem (37) parameterized by $\bar{x}$, i.e., $z^\epsilon(\bar{x}) = \bar{x}$. To understand the behavior of $G(x, \epsilon) = (x - z^\epsilon(x))/\epsilon$ near $\bar{x}$, we study a uniform stability of the solution map $z^\epsilon(x)$ around a neighborhood of $(\partial f)^{-1}(0)$, as established in the lemma below.

Lemma 8 (Uniform Holder stability near stationary points)**.** Consider problem (21), and denote its set of stationary points by $(\partial f)^{-1}(0)$. Let $z^\epsilon(x)$ be defined as in (37). For any compact set $X \subset \mathbb{R}^n$ and any $\bar{\epsilon} > 0$, there exist positive constants δ and D such that

$$\sup_{\epsilon \in (0, \bar{\epsilon})} \|z^\epsilon(x) - x\| \leq D \cdot \mathrm{dist}\big(x, (\partial f)^{-1}(0) \cap X\big)^{1/2}$$

for all x with $\mathrm{dist}\big(x, (\partial f)^{-1}(0) \cap X\big) \leq \delta$.

Remark 6 (Comparison with the pointwise stability)**.** Applying a general perturbation result for optimal solutions [3, Proposition 4.37] to problem (37), for each fixed stationary point $\bar{x}$, we have $\|z^\epsilon(x) - z^\epsilon(\bar{x})\| = \|z^\epsilon(x) - \bar{x}\| = \mathcal{O}\big(\|x - \bar{x}\|^{1/2}\big)$ and thus $\|z^\epsilon(x) - x\| = \mathcal{O}\big(\|x - \bar{x}\|^{1/2}\big)$. In contrast, our result holds uniformly for all x around $(\partial f)^{-1}(0) \cap X$, rather than a single point. Moreover, our bound depends on $\mathrm{dist}(x, (\partial f)^{-1}(0) \cap X)$, which is not only tighter than $\|x - \bar{x}\|$ for stationary point $\bar{x} \in X$ but can also be further controlled by metric subregularity in the proof of Proposition 8 (cf. Appendix B.2).

Proof. Define $\Phi(x) \triangleq \big\{(z, u) \in \mathbb{R}^n \times \mathbb{R}^m \mid u = c(x) + \nabla c(x)^\top(z - x)\big\}$ as the feasible region of problem (37) parameterized by x, and rewrite problem (37) as

$$\big(z^\epsilon(x), u^\epsilon(x)\big) \triangleq \operatorname*{argmin}_{(z,u) \in \Phi(x)} \left\{ h(u) + \frac{1}{2\epsilon}\|z - x\|^2 \right\}, \qquad \epsilon > 0. \tag{38}$$

For any $\bar{x} \in (\partial f)^{-1}(0)$, it is immediate that $\big(z^\epsilon(\bar{x}), u^\epsilon(\bar{x})\big) = \big(\bar{x}, c(\bar{x})\big)$.

Step 1 (Uniform stability of the feasible region $\Phi(x)$)**.** Note that the feasible set $\Phi(x)$ is defined by an equality constraint for which the linear independence constraint qualification holds. By the stability theorem in [3, Theorem 2.87], for any stationary point $\bar{x} \in (\partial f)^{-1}(0) \cap X$, there exists a constant $C_{\bar{x}}$ such that for all (z, u, x') near $(\bar{x}, c(\bar{x}), \bar{x})$,

$$\mathrm{dist}\big((z, u), \Phi(x')\big) \leq C_{\bar{x}}\big\|c(x') + \nabla c(x')^\top(z - x') - u\big\|. \tag{39}$$

Since ∂f is osc and X is compact, the set $(\partial f)^{-1}(0) \cap X$ is compact. By the continuity of $c(\cdot)$, the set $S \triangleq \{(\bar{x}, c(\bar{x}), \bar{x}) \mid \bar{x} \in (\partial f)^{-1}(0) \cap X\}$ is also compact. By the compactness of S and the bound in (39), we can find a uniform constant C and a neighborhood $U(S)$ of S such that

$$\mathrm{dist}\big((z, u), \Phi(x')\big) \leq C\big\|c(x') + \nabla c(x')^\top(z - x') - u\big\|, \qquad \forall\, (z, u, x') \in U(S). \tag{40}$$

For any x, let $\hat{x} \in (\partial f)^{-1}(0) \cap X$ be such that $\|x - \hat{x}\| = \mathrm{dist}\big(x, (\partial f)^{-1}(0) \cap X\big)$. For all x in a sufficiently small neighborhood of $(\partial f)^{-1}(0) \cap X$, we have $(\hat{x}, c(\hat{x}), x) \in U(S)$, so applying (40) with $(z, u, x') = (\hat{x}, c(\hat{x}), x)$ yields

$$\mathrm{dist}\big((\hat{x}, c(\hat{x})), \Phi(x)\big) \leq C\big\|c(x) + \nabla c(x)^\top(\hat{x} - x) - c(\hat{x})\big\| \leq C\beta\|x - \hat{x}\|^2/2 \leq C_1\|x - \hat{x}\| \tag{41}$$

for some constant C_1, where the second inequality follows from the β-Lipschitz continuity of ∇c. Next, for all x sufficiently close to $(\partial f)^{-1}(0) \cap X$, take a neighborhood $\mathcal{B}(\bar{x}, c(\bar{x}))$ around $(\bar{x}, c(\bar{x}))$ such that $\mathcal{B}(\bar{x}, c(\bar{x})) \times \{\hat{x}\} \subset U(S)$ and $(z^\epsilon(x), u^\epsilon(x)) \in \mathcal{B}(\bar{x}, c(\bar{x}))$ for $\epsilon \in (0, \bar{\epsilon})$. For such x, we apply (40) again with $(z, u) \in [\Phi(x) \cap \mathcal{B}(\bar{x}, c(\bar{x}))]$ and $x' = \hat{x}$ to obtain

$$
\begin{aligned}
\mathrm{dist}\big((z, u), \Phi(\hat{x})\big) &\leq\ C \big\| c(\hat{x}) + \nabla c(\hat{x})^\top (z - \hat{x}) - c(x) - \nabla c(x)^\top (z - x) \big\| \\
&\leq\ C \big\| c(\hat{x}) + \nabla c(\hat{x})^\top (x - \hat{x}) - c(x) \big\| + C \big\| (\nabla c(\hat{x}) - \nabla c(x))^\top (z - x) \big\| \\
&\leq\ C\beta \|x - \hat{x}\|^2 / 2 + C\beta \|x - \hat{x}\| \|z - x\|,
\end{aligned}
$$

where the last inequality uses the β-Lipschitz continuity of ∇c. Consequently, for all x in a neighborhood of $(\partial f)^{-1}(0) \cap X$,

$$
\mathbb{D}\Big(\Phi(x) \cap \mathcal{B}(\bar{x}, c(\bar{x})),\ \Phi(\hat{x})\Big) \leq \sup_{(z,u) \in [\Phi(x) \cap \mathcal{B}(\bar{x}, c(\bar{x}))]} \mathrm{dist}\big((z, u), \Phi(\hat{x})\big) \leq C_2 \|x - \hat{x}\| \tag{42}
$$

for some constant C_2. So far, we have derived two bounds (41) and (42), measuring the local distance between $\Phi(x)$ and $\Phi(\hat{x})$, from the stability theorem.

Step 2 (A quadratic inequality to bound $\|z^\epsilon(x) - x\|$). From now on, we fix x in a neighborhood of $(\partial f)^{-1}(0) \cap X$ such that (41) and (42) hold. Recall that $\hat{x} \in (\partial f)^{-1}(0) \cap X$ is chosen so that $\|x - \hat{x}\| = \mathrm{dist}(x, (\partial f)^{-1}(0) \cap X)$. Since $\Phi(\hat{x})$ is a closed set, let $(z, u) \in \Phi(\hat{x})$ be such that $\big\|(z^\epsilon(x), u^\epsilon(x)) - (z, u)\big\| = \mathrm{dist}\big((z^\epsilon(x), u^\epsilon(x)), \Phi(\hat{x})\big)$. Then,

$$
\|u^\epsilon(x) - u\| \leq \mathrm{dist}\Big((z^\epsilon(x), u^\epsilon(x)), \Phi(\hat{x})\Big) \overset{(42)}{\leq} C_2 \|x - \hat{x}\|. \tag{43}
$$

Since $\hat{x}$ is a stationary point of problem (21), $0 \in \nabla c(\hat{x})\, \partial h(c(\hat{x}))$. By the convexity of h, the subgradient inequality yields $h(u) = h\big(c(\hat{x}) + \nabla c(\hat{x})^\top (z - \hat{x})\big) \geq h(c(\hat{x}))$ for any x. Hence,

$$
\begin{aligned}
h(u^\epsilon(x)) + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2 - h(c(\hat{x})) &\geq h(u^\epsilon(x)) - h(u) + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2 \\
&\overset{(43)}{\geq} -LC_2\|x - \hat{x}\| + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2.
\end{aligned} \tag{44}
$$

To bound the same quantity from above, we fix any pair $(z', u') \in \Phi(x)$. Since $(z^\epsilon(x), u^\epsilon(x))$ is the optimal solution to (38), we have $h(u^\epsilon(x)) + \frac{1}{2\epsilon}\|z^\epsilon(x) - x\|^2 \leq h(u') + \frac{1}{2\epsilon}\|z' - x\|^2$. Subtracting $h(c(\hat{x}))$ from both sides and rearranging terms, we obtain

$$
\begin{aligned}
&h(u^\epsilon(x)) + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2 - h(c(\hat{x})) \\
&\leq h(u') - h(c(\hat{x})) + \frac{1}{2\epsilon}\big(\|z' - x\|^2 - \|\hat{x} - x\|^2\big) + \frac{1}{2\epsilon}\big(\|\hat{x} - x\|^2 - \|z^\epsilon(x) - x\|^2 + \|z^\epsilon(x) - \hat{x}\|^2\big) \\
&\leq L\|u' - c(\hat{x})\| + \frac{1}{2\epsilon}\|z' + \hat{x} - 2x\| \|z' - \hat{x}\| + \frac{1}{\epsilon}\|x - \hat{x}\| \|z^\epsilon(x) - \hat{x}\| \\
&\leq \frac{1}{2\epsilon} \max\big\{2\epsilon L, \|z' + \hat{x} - 2x\|\big\} \cdot \big\|(z', u') - (\hat{x}, c(\hat{x}))\big\| + \frac{1}{\epsilon}\|x - \hat{x}\| \|z^\epsilon(x) - \hat{x}\|.
\end{aligned}
$$

Using the estimate $\|z' + \hat{x} - 2x\| \leq \|z' - \hat{x}\| + 2\|\hat{x} - x\| \leq \|(z', u') - (\hat{x}, c(\hat{x}))\| + 2\|\hat{x} - x\|$ and minimizing the right-hand side over $(z', u') \in \Phi(x)$, we get

$$
\begin{aligned}
& h(u^\epsilon(x)) + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2 - h(c(\hat{x})) \\
& \leq \frac{1}{2\epsilon} \max\left\{ 2\epsilon L, \, \mathrm{dist}\big((\hat{x}, c(\hat{x})), \Phi(x)\big) + 2\|\hat{x} - x\| \right\} \cdot \mathrm{dist}\big((\hat{x}, c(\hat{x})), \Phi(x)\big) \\
& \quad + \frac{1}{\epsilon}\|x - \hat{x}\|\|z^\epsilon(x) - \hat{x}\| \\
& \overset{(41)}{\leq} \frac{1}{2\epsilon} \max\left\{ 2\epsilon L, (C_1 + 2)\|x - \hat{x}\| \right\} \cdot C_1\|x - \hat{x}\| + \frac{1}{\epsilon}\|x - \hat{x}\|\|z^\epsilon(x) - \hat{x}\|.
\end{aligned}
\tag{45}
$$

Combining the lower bound (44) and the upper bound (45), we obtain the quadratic inequality:

$$
\beta_1 + \frac{1}{2\epsilon}\|z^\epsilon(x) - \hat{x}\|^2 \leq \beta_2 + \frac{1}{\epsilon}\|x - \hat{x}\|\|z^\epsilon(x) - \hat{x}\|,
$$

where $\beta_1 \triangleq -LC_2\|x - \hat{x}\|$ and $\beta_2 \triangleq (2\epsilon)^{-1} \max\{2\epsilon L, (C_1 + 2)\|x - \hat{x}\|\} \cdot C_1\|x - \hat{x}\|$. Solving this inequality in terms of $\|z^\epsilon(x) - \hat{x}\|$ gives

$$
\begin{aligned}
\|z^\epsilon(x) - \hat{x}\| & \leq \|x - \hat{x}\| + \sqrt{\|x - \hat{x}\|^2 - 2\epsilon(\beta_1 - \beta_2)} \leq 2\|x - \hat{x}\| + \sqrt{2\epsilon(\beta_2 - \beta_1)} \\
& = 2\|x - \hat{x}\| + \sqrt{C_1 \max\{2\epsilon L, (C_1 + 2)\|x - \hat{x}\|\} + 2\epsilon LC_2} \cdot \sqrt{\|x - \hat{x}\|}.
\end{aligned}
$$

Finally, since $\|z^\epsilon(x) - x\| \leq \|z^\epsilon(x) - \hat{x}\| + \|x - \hat{x}\|$, taking supreme over $\epsilon \in (0, \bar{\epsilon})$, we can find a constant $D > 0$ such that

$$
\sup_{\epsilon \in (0, \bar{\epsilon})} \|z^\epsilon(x) - x\| \leq D\|x - \hat{x}\|^{\frac{1}{2}} = D \cdot \mathrm{dist}\big(x, (\partial f)^{-1}(0) \cap X\big)^{\frac{1}{2}}
$$

for all x in a neighborhood of $(\partial f)^{-1}(0) \cap X$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

B.1 Proof of Lemma 5

Lemma 9 ([17, Theorem 5.3])**.** Consider problem (21) with G defined in (18), and let $z^\epsilon(x)$ be the unique solution to the prox-linear subproblem (37). For any $\epsilon > 0$, there exists a point $\hat{x}$ such that $\|z^\epsilon(x) - \hat{x}\| \leq \|z^\epsilon(x) - x\|$ and $\mathrm{dist}(0, \partial f(\hat{x})) \leq (3L\beta + 2/\epsilon)\|z^\epsilon(x) - x\|$.

We now prove Lemma 5 using Lemmas 8 and 9.

Proof of Lemma 5. Let $\tau > 0$ be such that the inequality of metric subregularity holds for all $x \in \mathbb{B}_\tau(\bar{x})$. Applying Lemma 8 with $X = \{\bar{x}\}$, there exist positive scalars δ and D such that

$$
\sup_{\epsilon \in (0, \bar{\epsilon})} \|z^\epsilon(x) - x\| \leq D\|x - \bar{x}\|^{\frac{1}{2}}, \qquad \forall\, x \in \mathbb{B}_\delta(\bar{x}).
\tag{46}
$$

Fix any $\varepsilon \in (0, \delta)$ such that $\varepsilon + 2D\sqrt{\varepsilon} \leq \tau$, and consider a point $x \in \mathbb{B}_\varepsilon(\bar{x}) \subset \mathbb{B}_\delta(\bar{x})$. By Lemma 9, there exists a point $\hat{x}$ satisfying $\|z^\epsilon(x) - \hat{x}\| \leq \|z^\epsilon(x) - x\|$ and $\mathrm{dist}(0, \partial f(\hat{x})) \leq (3L\beta + 2/\epsilon)\|z^\epsilon(x) - x\|$. Combining this with the bound (46), we obtain

$$
\begin{aligned}
\|\bar{x} - \hat{x}\| & \leq \|\bar{x} - x\| + \|x - z^\epsilon(x)\| + \|z^\epsilon(x) - \hat{x}\| \leq \|\bar{x} - x\| + 2\|z^\epsilon(x) - x\| \\
& \overset{(46)}{\leq} \|\bar{x} - x\| + 2D\|x - \bar{x}\|^{\frac{1}{2}} \leq \varepsilon + 2D\sqrt{\varepsilon} \leq \tau.
\end{aligned}
$$

Hence $\hat{x} \in \mathbb{B}_\tau(\bar{x})$, and $\mathrm{dist}(\hat{x}, (\partial f)^{-1}(0)) \le \kappa \cdot \mathrm{dist}(0, \partial f(\hat{x}))$ by metric subregularity. Then,

$$\begin{aligned}
\mathrm{dist}(x, (\partial f)^{-1}(0)) &\le \|x - z^\epsilon(x)\| + \|z^\epsilon(x) - \hat{x}\| + \mathrm{dist}(\hat{x}, (\partial f)^{-1}(0)) \\
&\le 2\|z^\epsilon(x) - x\| + \kappa \cdot \mathrm{dist}(0, \partial f(\hat{x})) \le (2 + (3L\beta + 2/\epsilon)\kappa) \cdot \|z^\epsilon(x) - x\|.
\end{aligned}$$

Finally, since $\|z^\epsilon(x) - x\| = \epsilon\|G(x, \epsilon)\|$, the result follows. $\qquad\square$

B.2 Proof of Proposition 8 (Eventual constant stepsize)

Proof. Suppose, for the sake of contradiction, that $\epsilon_{k,0} \downarrow 0$. Then there exists an index set $N \in \mathbb{N}_\infty^\sharp$ such that $\|g^{k,i_k}\| \to_N 0$, and the ratio test fails for all $k \in N$, i.e.,

$$\frac{\tilde{\epsilon}_k\|\tilde{g}^k\|}{\|g^{k,i_k}\|^{1/2}} > \frac{(\epsilon_{k,i_k})^{1/2}}{\epsilon_{k,0}}. \tag{47}$$

By Corollary 1 and given $\alpha \in (0,1)$, any $\epsilon_{k,i} \in (0, 1/(2L\beta)]$ with stepsize $\eta = \epsilon_{k,i}$ ensures the descent condition. Thus, $\epsilon_{k,i_k} \ge \min\{\epsilon_{k,0}, (4L\beta)^{-1}\}$, which implies that the right-hand side of (47) diverges to infinity as $k(\in N) \to +\infty$. Since $f(x^k) \le f(x^0)$ for all k, the sequence $\{x^k\}_{k\in N} \subset X$ is bounded. Let $\{x^k\}_{k\in N'}$ be a convergent subsequence such that $x^k \to_{N'} \bar{x}$ for some $N' \in \mathbb{N}_\infty^\sharp$ with $N' \subset N$. Because $\|g^{k,i_k}\| \to_N 0$, it follows from (G1) that $0 \in \partial f(\bar{x})$. By the continuity of f, we have $f(x^k) \to_{N'} f(\bar{x})$ and $f(\bar{x}) \le f(x^0)$, so $\bar{x} \in (\partial f)^{-1}(0) \cap X$.

Since ∂f is metrically subregular at $\bar{x}$ with some modulus $\kappa_{\bar{x}} > 0$, Lemma 5 with $\bar{\epsilon} = \epsilon_{0,0}$ guarantees the existence of a positive number τ such that the error bound property holds:

$$\mathrm{dist}(x, (\partial f)^{-1}(0)) \le ((3L\beta\epsilon + 2)\kappa_{\bar{x}} + 2\epsilon) \cdot \|G(x, \epsilon)\|, \qquad \forall\, x \in \mathbb{B}_\tau(\bar{x}),\, \epsilon \in (0, \epsilon_{0,0}).$$

In particular, for all sufficiently large $k \in N'$, we have

$$\mathrm{dist}(x^k, (\partial f)^{-1}(0)) \le ((3L\beta\epsilon_{k,i_k} + 2)\kappa_{\bar{x}} + 2\epsilon_{k,i_k}) \cdot \|g^{k,i_k}\|. \tag{48}$$

For each $k \in N'$, let $\hat{x}^k \in (\partial f)^{-1}(0)$ be such that $\|x^k - \hat{x}^k\| = \mathrm{dist}(x^k, (\partial f)^{-1}(0))$. Observing that the right-hand side of (48) converges to 0 as $k(\in N') \to +\infty$, we have $\|x^k - \hat{x}^k\| \to_{N'} 0$, and hence $\hat{x}^k \to_{N'} \bar{x}$. This yields the existence of a compact set $\widehat{X}$ containing $\{\hat{x}^k\}_{k\in N'}$. By Lemma 8 with $\bar{\epsilon} = (a_0)^{1/4}$, there exists $D > 0$ such that for all sufficiently large $k \in N'$,

$$\left\|z^{\tilde{\epsilon}_k}(x^k) - x^k\right\| \le D \cdot \mathrm{dist}\left(x^k, (\partial f)^{-1}(0) \cap \widehat{X}\right)^{\frac{1}{2}} = D\|x^k - \hat{x}^k\|^{\frac{1}{2}} = D \cdot \mathrm{dist}(x^k, (\partial f)^{-1}(0))^{\frac{1}{2}}.$$

Combining this with the error bound (48), we obtain

$$\tilde{\epsilon}_k\|\tilde{g}^k\| = \left\|z^{\tilde{\epsilon}_k}(x^k) - x^k\right\| \le D((3L\beta\epsilon_{k,i_k} + 2)\kappa_{\bar{x}} + 2\epsilon_{k,i_k})^{\frac{1}{2}} \cdot \|g^{k,i_k}\|^{\frac{1}{2}}.$$

Dividing both sides by $\|g^{k,i_k}\|^{1/2}$ gives

$$\frac{\tilde{\epsilon}_k\|\tilde{g}^k\|}{\|g^{k,i_k}\|^{1/2}} \le D((3L\beta\epsilon_{k,i_k} + 2)\kappa_{\bar{x}} + 2\epsilon_{k,i_k})^{\frac{1}{2}} \longrightarrow_{N'} 0,$$

which contradicts the divergence $\tilde{\epsilon}_k\|\tilde{g}^k\|/\|g^{k,i_k}\|^{1/2} \to +\infty$ in (47). Therefore, the sequence $\{\epsilon_{k,0}\}$ cannot converge to zero. There exists $\underline{\epsilon} > 0$ such that $\epsilon_{k,0} = \underline{\epsilon}$ for all sufficiently large k. Using Corollary 1 again, we conclude that

$$\eta_k \ge \epsilon_{k,i_k} \ge \min\{\epsilon_{k,0}, (4L\beta)^{-1}\} \ge \min\{\underline{\epsilon}, (4L\beta)^{-1}\}, \qquad \forall\, k \in \mathbb{N}.$$

$\qquad\square$