

# SUBDIFFERENTIALLY POLYNOMIALLY BOUNDED FUNCTIONS AND GAUSSIAN SMOOTHING-BASED ZERO-ORDER OPTIMIZATION\*

MING LEI<sup>†</sup>, TING KEI PONG<sup>‡</sup>, SHUQIN SUN<sup>§</sup>, AND MAN-CHUNG YUE<sup>¶</sup>

**Abstract.** We study the class of subdifferentially polynomially bounded (SPB) functions, which is a rich class of locally Lipschitz functions that encompasses all Lipschitz functions, all gradient- or Hessian-Lipschitz functions, and even some nonsmooth locally Lipschitz functions. We show that SPB functions are compatible with Gaussian smoothing (GS), in the sense that the GS of any SPB function is well-defined and satisfies a descent lemma akin to gradient-Lipschitz functions, with the Lipschitz constant replaced by a polynomial function. Leveraging this descent lemma, we propose GS-based zeroth-order optimization algorithms with an adaptive stepsize strategy for minimizing SPB functions, and we analyze their convergence rates with respect to both relative and absolute stationarity measures. Finally, we also establish the iteration complexity for achieving a  $(\delta, \epsilon)$ -approximate stationary point, based on a novel quantification of Goldstein stationarity via the GS gradient that could be of independent interest.

**Key words.** Gaussian smoothing, zeroth-order optimization, subdifferentially polynomially bounded functions, Goldstein stationarity

**MSC codes.** 49J52, 90C30, 90C56

**DOI.** 10.1137/24M1659911

**1. Introduction.** Zeroth-order optimization (a.k.a. derivative-free optimization) refers to optimization problems where the objective function can be accessed only through a zeroth-order oracle, a routine for evaluating the function at a prescribed point. Zeroth-order optimization often arises in situations where one aims at optimally exploring or configuring physical environments using experimental data or computer simulations, and has attracted intense research over the last few decades. We refer the reader to the expositions [11, 23] and references therein for classic works and recent developments on zeroth-order optimization.

A prominent zeroth-order optimization algorithm is Nesterov and Spokoiny's random search method [23, 29, 30, 33] developed based on the concept of Gaussian smoothing (GS), whose definition is recalled here for convenience.

\*Received by the editors May 7, 2024; accepted for publication (in revised form) March 16, 2025; published electronically June 19, 2025.

<https://doi.org/10.1137/24M1659911>

**Funding:** The second author was supported in part by a Research Scheme of the Research Grants Council of Hong Kong SAR, China (project T22-504/21R). The third author was supported in part by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (project 2023QYJ08). The fourth author was supported in part by the Hong Kong Research Grants Council under GRF project 15304422.

<sup>†</sup>College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, People's Republic of China (leim@cuit.edu.cn).

<sup>‡</sup>Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China (tk.pong@polyu.edu.hk).

<sup>§</sup>Key Laboratory of Optimization Theory and Applications at China West Normal University of Sichuan Province, School of Mathematics Education, China West Normal University, Nanchong, People's Republic of China (sunshuqinsusan@163.com).

<sup>¶</sup>Musketeers Foundation Institute of Data Science and Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong, People's Republic of China (mcyue@hku.hk).

DEFINITION 1.1 ([30, section 2]). Let  $\sigma > 0$  and let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a Lebesgue measurable function. The Gaussian smoothing of  $f$  is defined as

$$f_\sigma(x) = \mathbb{E}_{u \sim \mathcal{N}(0, I)}[f(x + \sigma u)],$$

where  $\mathcal{N}(0, I)$  denotes the  $d$ -dimensional standard Gaussian distribution.

As a convolution of  $f$  with the Gaussian kernel, the GS  $f_\sigma$  enjoys many desirable properties. For example, it was shown in [30, section 2] to inherit convexity and Lipschitz continuity from  $f$ . Moreover, it was shown that  $\nabla f_\sigma$  is Lipschitz continuous whenever  $f$  is globally Lipschitz. This latter fact was leveraged in [30, section 7] to establish the *first worst-case complexity result* for a (stochastic) zeroth-order method for minimizing a nonsmooth nonconvex *globally* Lipschitz function. The work [30] has stimulated a surge of studies on GS-based zeroth-order optimization algorithms; see, e.g., [1, 2, 13, 21, 24, 26, 31, 35, 36]. It should, however, be pointed out that the advantage of GS-based algorithms over classical finite-difference methods is still under discussion [22].

To the best of our knowledge, most existing works on GS-based zeroth-order optimization algorithms, if not all, require the objective function itself, its gradient, or its Hessian to be Lipschitz continuous. Such assumptions not only ensure that the GS is well-defined and its gradient can be unbiasedly approximated by random samples of  $f(x + \sigma u)u/\sigma$  or  $(f(x + \sigma u) - f(x))u/\sigma$  (with  $u \sim \mathcal{N}(0, I)$ ), but also play a crucial role in the convergence analysis of the corresponding GS-based zeroth-order optimization algorithms. Nonetheless, these Lipschitz assumptions may not hold in many practical applications, including hyperparameter tuning [15], distributionally robust optimization [20, 38], neural network training [40], adversarial attacks [8], and  $\mathcal{H}_\infty$  control [19]. It is thus important to study less stringent Lipschitz assumptions to widen the applicability of zeroth-order optimization.

A similar issue concerning Lipschitz assumptions also arises in the context of first-order methods, where the global Lipschitzness of the gradient is instrumental to the algorithmic design and analysis. As an attempt to relax the Lipschitz requirement in the study of first-order methods, various notions of generalized smoothness [9, 20, 27, 40] have been recently proposed and led to the development and analysis of new first-order methods for these classes of generalized smooth functions. While it may be tempting to adapt these notions to the study of zeroth-order optimization, it is unclear how this can be done even for the special case of GS-based zeroth-order optimization algorithms.

Recently, a class of locally Lipschitz functions with Lipschitz modulus growing at most *polynomially* was introduced in [3, Assumption 1] to study stochastic optimization. In this paper, we further study this class of locally Lipschitz functions and develop new GS-based zeroth-order optimization algorithms for minimizing this class of functions. Our main contributions are threefold.

1. We study the class of subdifferentially polynomially bounded (SPB) functions, which is the subclass of locally Lipschitz functions with a Lipschitz modulus that grows at most *polynomially*. The class of SPB functions is rich, encompassing not only all functions that are Lipschitz, or gradient-Lipschitz, or Hessian-Lipschitz, but also certain nonsmooth locally Lipschitz functions, such as functions arising from neural networks; see Examples 3.1(v)–(vi). We show that if  $f$  is SPB, then its GS  $f_\sigma$  is well-defined and continuously differentiable; moreover,  $f_\sigma$  and its partial derivatives are SPB too. We also establish a relationship between  $\nabla f_\sigma$  and the Goldstein  $\delta$ -subdifferential of

an SPB function  $f$ , which allows us to quantify the approximate stationarity of a point  $x$  with respect to  $f$  by measuring  $\nabla f_\sigma(x)$ . The Goldstein  $\delta$ -subdifferential is a commonly used subdifferential for studying stationarity of nonsmooth functions [17, 41], and our result can be viewed as an extension of [30, Theorem 2] and [24, Theorem 3.1] from globally Lipschitz to SPB functions.

2. We devise GS-based zeroth-order algorithms for SPB minimization under two different settings: the constrained convex setting (where the objective function  $f$  is convex) and the unconstrained nonconvex setting (where the feasible region  $\Omega = \mathbb{R}^d$ ). Our algorithms update the iterate  $x^k$  by moving along an approximate negative gradient direction with an adaptive stepsize depending inversely on a polynomial of  $\|x^k\|$ , and the approximate gradient is obtained as the random vector  $(f(x^k + \sigma u) - f(x^k))u/\sigma$  with  $u \sim \mathcal{N}(0, I)$ . We analyze the iteration complexity of the proposed algorithms. The crux of our analysis is a novel descent lemma for  $f_\sigma$  analogous to the standard descent lemma for Lipschitz differentiable functions, where the Lipschitz constant is replaced by a polynomial function.
3. In the unconstrained nonconvex setting, the above-mentioned complexity result is with respect to the GS  $f_\sigma$  but not the original objective function  $f$ . Therefore, we also analyze the iteration complexity of our proposed algorithms for computing a  $(\delta, \epsilon)$ -stationary point, a notion of approximate Goldstein stationary point, with respect to the original objective function.

The remainder of this paper unfolds as follows. We present the notation and preliminary materials in section 2. Section 3 introduces subdifferentially polynomially bounded functions and studies their properties in relation to GS and Goldstein  $\delta$ -stationarity. In sections 4 and 5, we prove the descent lemma and develop our GS-based zeroth-order algorithms for minimizing SPB functions.

**2. Notation and preliminaries.** Throughout this paper, we let  $\mathbb{R}^d$  denote the Euclidean space of dimension  $d$  equipped with the standard inner product  $\langle \cdot, \cdot \rangle$ . For any  $x \in \mathbb{R}^d$ , we let  $\|x\|$  denote its Euclidean norm, and  $\mathbb{B}(x, r)$  denote the closed ball in  $\mathbb{R}^d$  with center  $x$  and radius  $r \geq 0$ . We use  $\mathbb{B}_r$  to denote  $\mathbb{B}(0, r)$ , and further use  $\mathbb{B}$  to denote  $\mathbb{B}_1$ . We let  $I = [e_1, e_2, \dots, e_d]$  denote the  $d \times d$  identity matrix, where  $e_i \in \mathbb{R}^d$  is the  $i$ th canonical basis vector for  $i = 1, \dots, d$ , i.e.,  $(e_i)_j = 1$  if  $j = i$  and  $(e_i)_j = 0$  otherwise.

For a subset  $D \subseteq \mathbb{R}^d$ , we let  $D^c$ ,  $\partial D$ , and  $\text{conv}(D)$  denote its complement, boundary, and convex hull, respectively; we also denote the characteristic function of  $D$  by

$$\mathbf{1}_D(x) = \begin{cases} 1 & \text{if } x \in D, \\ 0 & \text{if } x \notin D. \end{cases}$$

For a closed set  $S \subseteq \mathbb{R}^d$ , the distance from an  $x \in \mathbb{R}^d$  to  $S$  is defined as  $\text{dist}(x, S) = \inf_{y \in S} \|x - y\|$ . For a closed convex set  $S$ , the (unique) projection of an  $x \in \mathbb{R}^d$  onto  $S$  is denoted by  $P_S(x)$ ; also, the normal cone of  $S$  at any  $x \in S$  is defined as

$$N_S(x) = \{y \in \mathbb{R}^d : \langle y, u - x \rangle \leq 0 \quad \forall u \in S\}.$$

For a locally Lipschitz function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , the Clarke directional derivative of  $f$  (see [10, page 25]) at any  $x \in \mathbb{R}^d$  in the direction  $v \in \mathbb{R}^d$  is defined as

$$f^\circ(x; v) = \limsup_{\substack{x' \rightarrow x, t \downarrow 0}} \frac{f(x' + tv) - f(x')}{t},$$

and the Clarke subdifferential of  $f$  (see [10, page 27]) at  $x$  is the set

$$\partial_C f(x) = \{s \in \mathbb{R}^d : \langle s, v \rangle \leq f^\circ(x; v) \quad \forall v \in \mathbb{R}^d\}.$$

The Clarke directional derivative and Clarke subdifferential are related as follows:

$$f^\circ(x; v) = \max_{s \in \partial_C f(x)} \langle s, v \rangle;$$

also, letting  $\Upsilon_f$  be the set of points at which  $f$  is not differentiable, we have

$$(2.1) \quad \partial_C f(x) = \text{conv}(\{s \in \mathbb{R}^d : \exists \{x^k\} \subset \mathbb{R}^d \setminus \Upsilon_f \text{ with } x^k \rightarrow x \text{ and } \nabla f(x^k) \rightarrow s\});$$

see [10, Proposition 2.1.2(b)] and [10, Theorem 2.5.1].

Next, for any  $\delta > 0$ , the Goldstein  $\delta$ -subdifferential [17] of  $f$  at  $x \in \mathbb{R}^d$  is the set

$$(2.2) \quad \partial_G^\delta f(x) = \text{conv}\left(\bigcup_{y \in \mathbb{B}(x, \delta)} \partial_C f(y)\right).$$

Note that at any  $x \in \mathbb{R}^d$ , both the Clarke subdifferential and Goldstein  $\delta$ -subdifferential are compact convex sets.

**3. Subdifferentially polynomially bounded functions.** In this section, we study the class of subdifferentially polynomially bounded (SPB) functions.<sup>1</sup>

**DEFINITION 3.1** (subdifferentially polynomially bounded functions). *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be locally Lipschitz continuous. We say that  $f$  is subdifferentially polynomially bounded (SPB) if there exist  $R_1 \geq 0$ ,  $R_2 > 0$ , and an integer  $m \geq 0$  with  $R_1 = 0$  if and only if  $m = 0$  such that<sup>2</sup>*

$$(3.1) \quad \sup_{\zeta \in \partial_C f(x)} \|\zeta\| \leq R_1 \|x\|^m + R_2 \quad \forall x \in \mathbb{R}^d.$$

The class of SPB functions on  $\mathbb{R}^d$  is denoted by  $\mathcal{SPB}(\mathbb{R}^d)$ .

Note that using calculus rules for the Clarke subdifferential (see Corollary 2 of [10, Proposition 2.3.3]), one can show that  $\mathcal{SPB}(\mathbb{R}^d)$  is a vector space. Also,  $\mathcal{SPB}(\mathbb{R}^d)$  generalizes the class of globally Lipschitz functions, which correspond to the case of  $R_1 = 0$  in (3.1); see Example 3.1(i) below. In fact, the SPB class is much richer than that and covers a wide variety of functions that arise naturally in many contemporary applications. Here, we present some concrete examples of SPB functions.

*Example 3.1.*

- (i) If  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is globally Lipschitz continuous with Lipschitz continuity modulus  $L > 0$ , then we have from [10, Proposition 2.1.2(a)] that  $\sup_{u \in \partial_C f(x)} \|u\| \leq L$  for all  $x \in \mathbb{R}^d$ . Consequently,  $f$  is SPB.
- (ii) Every polynomial function is SPB.
- (iii) Any continuously differentiable function with a Lipschitz gradient is SPB. To see this, let  $g$  be such a function; then there exists  $L > 0$  such that

$$\|\nabla g(x)\| \leq \|\nabla g(x) - \nabla g(0)\| + \|\nabla g(0)\| \leq L\|x\| + \|\nabla g(0)\| \quad \forall x \in \mathbb{R}^d,$$

showing that  $g$  is SPB (with  $m = 1$ ). The converse is, however, not true, as an SPB function with  $m = 1$  is not necessarily differentiable.

<sup>1</sup>Condition (3.1) with  $R_1 = R_2$  is equivalent to [3, Assumption 1.1] when the function  $f(x, s)$  there is constant in  $s$ , i.e., in the deterministic setting.

<sup>2</sup>We adopt the convention  $0^0 = 1$  when  $x = 0$  and  $m = 0$  (in which case  $R_1 = 0$ ).

- (iv) Let  $f = g \circ h$ , where  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $h : \mathbb{R}^d \rightarrow \mathbb{R}^n$ . Assume that  $g$  and all component functions of  $h$  are SPB. Then one can deduce from [10, Theorem 2.3.9] that  $f$  is SPB.
- (v) In machine learning, one is often interested in approximating the unknown relationship between an independent variable  $x \in \mathbb{R}^d$  and a dependent variable  $y \in \mathbb{R}$ . An  $L$ -layer neural network is a parametric approximation of the form

$$(3.2) \quad y = \psi(x; w) = \varrho_L(W_L(\varrho_{L-1}(W_{L-1}(\cdots \varrho_1(W_1(x)) \cdots))),$$

where for  $\ell = 1, \dots, L$ ,  $\varrho_\ell : \mathbb{R} \rightarrow \mathbb{R}$  is the activation function for the  $\ell$ th layer (for any vector  $z$ , the notation  $\varrho_\ell(z)$  is understood as the vector obtained by applying the activation function  $\varrho_\ell$  entrywise to  $z$ ),  $W_\ell : \mathbb{R}^{p_\ell} \rightarrow \mathbb{R}^{p_{\ell+1}}$  is an affine mapping for some positive integers  $p_\ell$  and  $p_{\ell+1}$  (with  $p_1 = d$  and  $p_{L+1} = 1$ ), and  $w$  is called the parameter and represents the vector of all coefficients defining the maps  $W_1, \dots, W_L$ ; see [4, section 6.2] for details. Common activation functions include  $\varrho(t) = t$  (often used for the output layer),  $\varrho(t) = \tanh(t)$ ,  $\varrho(t) = \ln(1 + e^t)$ ,  $\varrho(t) = \max\{0, t\}$ ,  $\varrho(t) = \max\{0, t\} + \alpha \min\{0, t\}$  with  $\alpha > 0$ ,  $\varrho(t) = \frac{1}{1+e^{-t}}$ , and piecewise polynomial functions. With any of these activation functions, Example 3.1(iv) implies that the neural network function  $\psi(\cdot; w)$  is SPB for any fixed parameter  $w$ .

- (vi) Suppose that we are given a sample  $\{(x_i, y_i)\}_{i=1}^n$  of  $n$  data points for approximating the unknown relationship between  $x$  and  $y$ . Naturally, we want to find the best parameter  $w$  so that the function  $\psi(\cdot; w)$  in (3.2) fits the sample data as well as possible, a process called “training.” A popular formulation for the best parameter  $w$  is given by the least squares criterion

$$\min_w \sum_{i=1}^n (y_i - \psi(x_i; w))^2.$$

Again by Example 3.1(iv), the objective function in this problem is SPB.

**3.1. Gaussian smoothing of SPB functions.** We next study the properties of SPB functions in relation to Gaussian smoothing (GS) [23, 29, 30, 33]. More precisely, we will show that for any SPB function, both the GS and its gradient are well-defined and that the class  $\mathcal{SPB}(\mathbb{R}^d)$  is closed under the GS transformation. Towards that end, we record a simple property of SPB functions that will be repeatedly used in the paper. Specifically, we express the Lipschitz modulus of an SPB function in terms of a sum of functions in  $x$  and the displacement  $y - x$ . This explicit dependence on the displacement is crucial for our subsequent analysis, especially in section 4.

**LEMMA 3.2.** *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1). Then*

$$|f(x) - f(y)| \leq (2^{m-1}R_1\|x\|^m + 2^{m-1}R_1\|y - x\|^m + R_2)\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

*Proof.* From [10, Theorem 2.3.7], we have

$$f(x) - f(y) \in \{\langle \zeta, x - y \rangle : \zeta \in \partial_C f(x + \alpha(y - x)), \alpha \in (0, 1)\}.$$

In view of this and (3.1), we deduce further that

$$\begin{aligned} |f(x) - f(y)| &\leq \sup_{\alpha \in (0, 1)} \{R_1\|x + \alpha(y - x)\|^m + R_2\}\|x - y\| \\ &\leq (2^{m-1}R_1\|x\|^m + 2^{m-1}R_1\|y - x\|^m + R_2)\|x - y\|, \end{aligned}$$

where the second inequality follows from the fact that  $\alpha \in (0, 1)$  and the convexity of the function  $\|\cdot\|^m$  when  $m \geq 1$ , and the inequality holds trivially when  $m = 0$  (in which case  $R_1 = 0$ ) with the convention  $0^0 = 1$ .  $\square$

When the function  $f$  is not globally Lipschitz, its GS  $f_\sigma$  is not necessarily defined. This is exemplified by the function  $f(x) = e^{\|x\|^4}$ . The theorem below asserts that for any SPB function  $f$ , the GS  $f_\sigma$  and its gradient  $\nabla f_\sigma$  are both well-defined. An explicit formula for  $\nabla f_\sigma$  is also proved. The crux for the proof of the formula lies in the interchangeability of differentiation and integration, which we achieve by using the theory of Schwartz spaces and tempered distributions. Roughly speaking, a Schwartz space consists of  $C^\infty$  functions whose derivatives of any order decay faster than any polynomial (e.g.,  $e^{-\|\cdot\|^2/2}$ ; see [18, Example 2.2.2]), and tempered distributions form its topological dual.

**THEOREM 3.3** (well-definedness of GS and its gradient). *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1). Then its GS  $f_\sigma$ , given in Definition 1.1, is well-defined. Moreover, the gradient of  $f_\sigma$  is given by*

$$(3.3) \quad \nabla f_\sigma(x) = \sigma^{-1} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [f(x + \sigma u) u]$$

and is well-defined and continuous.

*Proof.* For any  $x \in \mathbb{R}^d$ , we have

$$(3.4) \quad \begin{aligned} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [|f(x + \sigma u)|] &\leq \mathbb{E}_{u \sim \mathcal{N}(0, I)} [|f(x + \sigma u) - f(x)|] + |f(x)| \\ &\leq \mathbb{E}_{u \sim \mathcal{N}(0, I)} [(2^{m-1} R_1 \|x\|^m + 2^{m-1} R_1 \sigma^m \|u\|^m + R_2) \cdot \sigma \|u\|] + |f(x)| < \infty, \end{aligned}$$

where the second inequality follows from Lemma 3.2. Therefore,  $f_\sigma$  is well-defined.

We next prove (3.3) and the well-definedness of the expectation there. First, by the definition of GS, we have

$$(3.5) \quad f_\sigma(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x + \sigma u) e^{-\frac{\|u\|^2}{2}} du = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \int_{\mathbb{R}^d} f(y) e^{-\frac{\|x-y\|^2}{2\sigma^2}} dy.$$

Note that  $e^{-\|\cdot\|^2/(2\sigma^2)}$  is a Schwartz function on  $\mathbb{R}^d$  according to [18, Example 2.2.2]. On the other hand, we have the following inequality for each  $r > 0$  based on Lemma 3.2:

$$\begin{aligned} \int_{\mathbb{B}_r} |f(x)| dx &\leq r^d \alpha(d) \cdot (|f(0)| + \max_{x \in \mathbb{B}_r} |f(x) - f(0)|) \\ &\leq r^d \alpha(d) \cdot (|f(0)| + [2^{m-1} R_1 r^m + R_2] r) =: \mathfrak{U}(r), \end{aligned}$$

where  $\alpha(d)$  is the volume of the  $d$ -dimensional unit ball  $\mathbb{B}$ . Since  $\mathfrak{U}(r) = O(r^{m+d+1})$  as  $r \rightarrow \infty$ , in view of [37],<sup>3</sup> we see that  $\mathfrak{F}(g) := \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \int_{\mathbb{R}^d} f(y) g(y) dy$  is a continuous linear functional on Schwartz spaces (i.e., a tempered distribution).

The next part follows closely the proof of [18, Theorem 2.3.20], which is included for self-containedness. Specifically, from (3.5), we see that for any  $h \in \mathbb{R} \setminus \{0\}$  and any  $i \in \{1, \dots, d\}$ ,

$$\begin{aligned} \frac{f_\sigma(x + h e_i) - f_\sigma(x)}{h} &= \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \int_{\mathbb{R}^d} f(y) \frac{e^{-\frac{\|x + h e_i - y\|^2}{2\sigma^2}} - e^{-\frac{\|x - y\|^2}{2\sigma^2}}}{h} dy \\ &= \mathfrak{F}((e^{-\frac{\|x + h e_i - y\|^2}{2\sigma^2}} - e^{-\frac{\|x - y\|^2}{2\sigma^2}})/h). \end{aligned}$$

<sup>3</sup>Specifically, see the last example on page 106.

Since  $(e^{-\frac{\|x+he_i-y\|^2}{2\sigma^2}} - e^{-\frac{\|x-y\|^2}{2\sigma^2}})/h \rightarrow -\frac{x_i-y_i}{\sigma^2}e^{-\frac{\|x-y\|^2}{2\sigma^2}}$  as  $h \rightarrow 0$  in Schwartz spaces according to [18, Exercise 2.3.5(a)] and  $\mathfrak{F}$  is a tempered distribution, we conclude upon passing to the limit as  $h \rightarrow 0$  in the above display that

$$\begin{aligned}\nabla f_\sigma(x) &= \left[ \mathfrak{F} \left( -\frac{x_i-y_i}{\sigma^2} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \right) \right]_{i=1}^d = -\frac{1}{(2\pi)^{\frac{d}{2}} \sigma^{d+2}} \int_{\mathbb{R}^d} f(y)(x-y) e^{-\frac{\|x-y\|^2}{2\sigma^2}} dy \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} \sigma} \int_{\mathbb{R}^d} f(x+\sigma u) u e^{-\frac{\|u\|^2}{2}} du.\end{aligned}$$

This proves (3.3) and the well-definedness of the integral.

Finally, the continuity of  $\nabla f_\sigma$  follows immediately from the above integral representation and the dominated convergence theorem, where the required integrability assumption can be established in a similar way to (3.4).  $\square$

The next result shows in particular that  $\mathcal{SPB}(\mathbb{R}^d)$  is closed under the GS transformation and that if  $f$  is SPB, so are its partial derivatives. The case  $m = 0$  was already established in [30].

**THEOREM 3.4.** *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1, R_2$ , and  $m$  as in (3.1) and let  $f_\sigma$  be defined as in Definition 1.1. Then the following statements hold:*

(i) *It holds that*

$$(3.6) \quad |f_\sigma(x) - f_\sigma(y)| \leq (\mathfrak{A} + \mathfrak{B}\|x\|^m + \mathfrak{C}\|y-x\|^m)\|x-y\| \quad \forall x, y \in \mathbb{R}^d,$$

where  $\mathfrak{A} = 2^{2m-2}R_1\sigma^m(m+d)^{\frac{m}{2}} + R_2$ ,  $\mathfrak{B} = 2^{2m-2}R_1$  and  $\mathfrak{C} = 2^{m-1}R_1$ . In particular,  $f_\sigma$  is SPB.

(ii) *It holds that*

$$(3.7) \quad \|\nabla f_\sigma(x) - \nabla f_\sigma(y)\| \leq (\mathcal{A} + \mathcal{B}\|x\|^m + \mathcal{C}\|y-x\|^m)\|x-y\| \quad \forall x, y \in \mathbb{R}^d,$$

where  $\mathcal{A} = 2^{2m-2}R_1\sigma^{m-1}(m+1+d)^{\frac{m+1}{2}} + \sigma^{-1}R_2\sqrt{d}$ ,  $\mathcal{B} = 2^{2m-2}\sigma^{-1}R_1\sqrt{d}$  and  $\mathcal{C} = 2^{m-1}\sigma^{-1}R_1\sqrt{d}$ . In particular,  $\frac{\partial f_\sigma}{\partial x_i}$  is SPB for any  $i$ .

*Proof.* The case  $m = 0$  (in which case we have  $R_1 = 0$ ) was studied in [30], with item (i) proved in the display before [30, eqn. (12)], and item (ii) proved in [30, Lemma 2].

We next consider the case  $m \geq 1$ . We first observe from Lemma 3.2 that for every  $x, y$  and  $u \in \mathbb{R}^d$ ,

$$(3.8) \quad \begin{aligned} &|f(x+\sigma u) - f(y+\sigma u)| \\ &\leq (2^{m-1}R_1\|x+\sigma u\|^m + 2^{m-1}R_1\|y-x\|^m + R_2)\|y-x\| \\ &\leq (2^{2m-2}R_1\sigma^m\|u\|^m + 2^{2m-2}R_1\|x\|^m + 2^{m-1}R_1\|y-x\|^m + R_2)\|x-y\|, \end{aligned}$$

where the second inequality follows from the convexity of  $\|\cdot\|^m$  when  $m \geq 1$ .

To prove (i), from (3.8), one has for any  $x \neq y$  that

$$\begin{aligned} \frac{|f_\sigma(x) - f_\sigma(y)|}{\|x-y\|} &\leq \frac{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[|f(x+\sigma u) - f(y+\sigma u)|]}{\|x-y\|} \\ &\leq \mathbb{E}_{u \sim \mathcal{N}(0,I)}[(2^{2m-2}R_1\sigma^m\|u\|^m + 2^{2m-2}R_1\|x\|^m + 2^{m-1}R_1\|y-x\|^m + R_2)] \\ &= (2^{2m-2}R_1\|x\|^m + 2^{m-1}R_1\|y-x\|^m + R_2) + 2^{2m-2}R_1\sigma^m\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m] \\ &\leq (2^{2m-2}R_1\|x\|^m + 2^{m-1}R_1\|y-x\|^m + R_2) + 2^{2m-2}R_1\sigma^m(m+d)^{\frac{m}{2}}, \end{aligned}$$

where the last inequality follows from [30, Lemma 1]. This proves (3.6).

Now, fix any  $x \in \mathbb{R}^d$  and  $v \in \mathbb{R}^d$  with  $\|v\| = 1$ . We have for any  $\xi \in \partial_C f_\sigma(x)$  that

$$\langle \xi, v \rangle \leq \limsup_{x' \rightarrow x, t \downarrow 0} \frac{f_\sigma(x' + tv) - f_\sigma(x')}{t} \leq \limsup_{x' \rightarrow x, t \downarrow 0} (\mathfrak{A} + \mathfrak{B}\|x'\|^m + \mathfrak{C}t^m) \leq \mathfrak{A} + \mathfrak{B}\|x\|^m.$$

Consequently, it holds that  $\|\xi\| \leq \mathfrak{A} + \mathfrak{B}\|x\|^m$ , showing that  $f_\sigma \in \mathcal{SPB}(\mathbb{R}^d)$ .

To prove (ii), we notice from (3.3) that

$$\|\nabla f_\sigma(x) - \nabla f_\sigma(y)\| \leq \sigma^{-1} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [|f(x + \sigma u) - f(y + \sigma u)| \cdot \|u\|].$$

Combining the above display with (3.8), we have for any  $x \neq y$  that

$$\begin{aligned} & \frac{\|\nabla f_\sigma(x) - \nabla f_\sigma(y)\|}{\|x - y\|} \\ & \leq 2^{2m-2} \mathbf{R}_1 \sigma^{m-1} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [\|u\|^{m+1}] \\ & \quad + \sigma^{-1} (2^{2m-2} \mathbf{R}_1 \|x\|^m + 2^{m-1} \mathbf{R}_1 \|y - x\|^m + \mathbf{R}_2) \mathbb{E}_{u \sim \mathcal{N}(0, I)} [\|u\|] \\ & \leq 2^{2m-2} \mathbf{R}_1 \sigma^{m-1} (m+1+d)^{\frac{m+1}{2}} + \sigma^{-1} (2^{2m-2} \mathbf{R}_1 \|x\|^m + 2^{m-1} \mathbf{R}_1 \|y - x\|^m + \mathbf{R}_2) \sqrt{d}, \end{aligned}$$

where the last inequality follows from [30, Lemma 1]. This proves (3.7).

The claim that  $\frac{\partial f_\sigma}{\partial x_i} \in \mathcal{SPB}(\mathbb{R}^d)$  can now be proved in a similar way to the proof of  $f_\sigma \in \mathcal{SPB}(\mathbb{R}^d)$  in item (i).  $\square$

**3.2. Approximate Goldstein stationarity.** In this subsection, we explore the relationship between the GS gradient and the Goldstein  $\delta$ -subdifferential for SPB functions. We start with the following auxiliary lemma concerning the tail of the Gaussian integral. We let  $W_{-1}$  denote the negative real branch of the Lambert  $W$  function (see, e.g., [5, 16, 32]); this function is defined as the inverse of the function  $t \mapsto te^t$  with domain  $[-1/e, 0)$  and range  $(-\infty, -1]$ .

LEMMA 3.5. *For any  $\nu > 0$  and  $M \geq [-d \cdot W_{-1}(-\nu^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$ , it holds that*

$$\int_{\|u\| \geq M} e^{-\frac{\|u\|^2}{2}} du \leq \nu;$$

here, we set by convention that  $W_{-1}(t) = 0$  if  $t < -1/e$ .

*Proof.* Fix any  $\nu > 0$  and  $M \geq \wp := [-d \cdot W_{-1}(-\nu^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$ . For any  $R \geq 0$ ,

$$\int_{\|u\| \geq R} e^{-\frac{\|u\|^2}{2}} du = (2\pi)^{\frac{d}{2}} \cdot \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\|u\| \geq R} e^{-\frac{\|u\|^2}{2}} du = (2\pi)^{\frac{d}{2}} [1 - F(R^2; d)],$$

where  $F(\cdot; k)$  is the cumulative distribution function of the chi-squared distribution with  $k$  degrees of freedom. Thus, the desired conclusion is equivalent to

$$(3.9) \quad 1 - \nu(2\pi)^{-\frac{d}{2}} \leq F(M^2; d).$$

We now prove (3.9). Note that if  $\nu \geq (2\pi)^{\frac{d}{2}}$ , then we have  $1 - \nu(2\pi)^{-\frac{d}{2}} \leq 0 \leq F(M^2; d)$ . Hence, (3.9) is valid in this case. We next consider the case  $\nu < (2\pi)^{\frac{d}{2}}$ . In this case, we have  $\nu^{\frac{2}{d}} < 2\pi$  and hence  $M \geq \wp > \sqrt{d}$ .<sup>4</sup> Then, from [12, Lemma 2.2] (see also [34, Proposition 5.3.1]), we have

$$(3.10) \quad 1 - F(M^2; d) = 1 - F\left(\frac{M^2}{d}; d\right) \leq \left(\frac{M^2}{d} e^{1 - \frac{M^2}{d}}\right)^{\frac{d}{2}}.$$

<sup>4</sup>The second inequality holds because  $W_{-1}(t) < -1$  when  $t \in (-1/e, 0)$ ; see [25, page 2].



Now, note that we have the following equivalence for any  $R \geq 0$ :

$$(3.11) \quad \left( \frac{R^2}{d} e^{1 - \frac{R^2}{d}} \right)^{\frac{d}{2}} \leq \frac{\nu}{(2\pi)^{\frac{d}{2}}} \iff \left( -\frac{R^2}{d} \right) e^{-\frac{R^2}{d}} \geq -e^{-1} \frac{\nu^{\frac{2}{d}}}{2\pi}.$$

Then we have from the definition of Lambert  $W$  function that the rightmost inequality (and hence both inequalities) in (3.11) holds with  $R = M$  since  $M \geq \varphi$ . Thus, the desired conclusion follows from (3.11) and (3.10).  $\square$

In the next theorem, we show that for all sufficiently small  $\sigma > 0$ , some Goldstein  $\delta$ -subgradients can be approximated by the GS gradient  $\nabla f_\sigma$ . Specifically, we derive a sufficient condition on  $\sigma$ , in the form of an *explicit* upper bound depending on  $\delta$  and  $\varepsilon$ , for the GS gradient to reside in a  $(1 + \|x\|^m)\varepsilon$  neighborhood of the Goldstein  $\delta$ -subdifferential. This inclusion is proved as follows. We first express the GS gradient as an expectation of the original gradient  $\nabla f(x + \sigma u)$  (which exists almost everywhere by Rademacher's theorem) with respect to the random vector  $u \sim \mathcal{N}(0, I)$ ; see (3.12) below. It follows from (2.2) that  $\nabla f(x + \sigma u)$  constitutes a Goldstein subgradient for any realization of  $u$  close enough to  $x$ . Therefore, by dividing the expectation in (3.12) into two integrals, one over a small ball centered at  $x$  and the other over the complement, we can see that the GS gradient is an approximate Goldstein subgradient. The integral over the complement contributes to the approximation error, which can be controlled by using Lemma 3.5.

Similar results have been derived under a globally Lipschitz continuity assumption on  $f$ ; see, e.g., [30, Theorem 2] and [24, Theorem 3.1]. In particular, the proof of [24, Theorem 3.1] was based on an analogue of (3.12) for globally Lipschitz continuous  $f$ . We would also like to point out that the representation (3.12) can be seen as a variant of general results on convolution and differentiation such as [39, section 4.2.5] and [6, Lemma 9.1]. Here we include an elementary proof of (3.12) to highlight the role of polynomial boundedness of the subdifferential.

**THEOREM 3.6** (GS gradient as approximate Goldstein  $\delta$ -subgradient). *Let  $f \in \text{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1). Let  $\nabla f_\sigma$  and  $\partial_G^\delta f$  be given as in (3.3) and (2.2), respectively. Then the following hold:*

(i) *For every  $x \in \mathbb{R}^d$ , it holds that*

$$(3.12) \quad \nabla f_\sigma(x) = \mathbb{E}_{u \sim \mathcal{N}(0, I)} [\nabla f(x + \sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u)],$$

*where  $\mathfrak{D}_\sigma = \{u \in \mathbb{R}^d : f \text{ is differentiable at } x + \sigma u\}$ .*

(ii) *For every  $\delta > 0$  and  $\varepsilon > 0$ , it holds that*

$$\nabla f_\sigma(x) \in \partial_G^\delta f(x) + (1 + \|x\|^m)\varepsilon \cdot \mathbb{B} \quad \forall \sigma \in (0, \bar{\sigma}] \text{ and } \forall x \in \mathbb{R}^d,$$

*where*

$$(3.13) \quad \bar{\sigma} = \min \left\{ \left[ \frac{\varepsilon}{2^{m+1}R_1(m+d)^{\frac{m}{2}}} \right]^{\frac{1}{m}}, 1, \frac{\delta}{H} \right\},$$

$$(3.14) \quad H = [-d \cdot W_{-1}(-\eta_1^{2/d}/(2\pi e))]^{\frac{1}{2}},$$

$$(3.15) \quad \eta_1 = \min\{\varepsilon \mathcal{P}^{-1}, (2\pi)^{\frac{d}{2}} - 0.5\},$$

$$(3.16) \quad \mathcal{P} = 4R_2 + 2^{m+1}R_1(m+d)^{\frac{m}{2}},$$

*and  $W_{-1}$  is the negative real branch of the Lambert  $W$  function.*<sup>5</sup>

<sup>5</sup>Since  $\eta_1 \leq (2\pi)^{\frac{d}{2}} - \frac{1}{2}$ , we have  $\eta_1^{2/d}/(2\pi e) < 1/e$  and hence  $W_{-1}(-\eta_1^{2/d}/(2\pi e)) < -1$ . Thus,  $H \in (\sqrt{d}, \infty)$ .

*Proof.* Fix any  $x \in \mathbb{R}^d$ . From Theorem 3.3, the GS  $f_\sigma$  and its gradient  $\nabla f_\sigma$  are well-defined at  $x$ . For notational simplicity, for any  $\sigma > 0$ , let  $\mathfrak{D}_\sigma = \{u \in \mathbb{R}^d : f \text{ is differentiable at } x + \sigma u\}$ . Then it follows from Rademacher's theorem that the complement  $\mathfrak{D}_\sigma^c$  has Lebesgue measure zero.

To prove (i), note from Lemma 3.2 that for each  $u$  and  $h \in \mathbb{R}^d$

$$(3.17) \quad \begin{aligned} |f(x+h+\sigma u) - f(x+\sigma u)| &\leq [2^{m-1}R_1\|x+\sigma u\|^m + 2^{m-1}R_1\|h\|^m + R_2]\|h\| \\ &\leq (2^{2m-2}R_1\|x\|^m + 2^{2m-2}R_1\sigma^m\|u\|^m + 2^{m-1}R_1\|h\|^m + R_2)\|h\|, \end{aligned}$$

where the second inequality follows from the convexity of  $\|\cdot\|^m$  when  $m \geq 1$  and the inequality holds when  $m = 0$  (in which case  $R_1 = 0$ ) with the convention  $0^0 = 1$ .

We prove (3.12) by contradiction. First,  $\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u)]$  exists as  $f$  is SPB. Suppose to the contrary  $\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u)] \neq \nabla f_\sigma(x)$ . Define

$$h_x = \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u)] - \nabla f_\sigma(x).$$

Then  $h_x \neq 0$  and we have from the differentiability of  $f_\sigma$  (see Theorem 3.3) that

$$\begin{aligned} 0 &= \lim_{t \rightarrow 0} \frac{f_\sigma(x+th_x) - f_\sigma(x) - \langle \nabla f_\sigma(x), th_x \rangle}{\|th_x\|} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \lim_{t \rightarrow 0} \int_{\mathbb{R}^d} \frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle \nabla f_\sigma(x), th_x \rangle}{\|th_x\|} \cdot e^{-\frac{\|u\|^2}{2}} du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \lim_{t \rightarrow 0} \int_{\mathfrak{D}_\sigma} \frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle \nabla f_\sigma(x), th_x \rangle}{\|th_x\|} \cdot e^{-\frac{\|u\|^2}{2}} du \\ &\stackrel{(a)}{=} \mathbb{E}_{u \sim \mathcal{N}(0,I)} \left[ \frac{\langle \nabla f(x+\sigma u) - \nabla f_\sigma(x), h_x \rangle}{\|h_x\|} \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u) \right] \stackrel{(b)}{=} \|h_x\|, \end{aligned}$$

where (a) follows from (3.17), the dominated convergence theorem, and the fact that

$$\begin{aligned} \lim_{t \rightarrow 0} \left| \frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle \nabla f_\sigma(x), th_x \rangle}{\|th_x\|} - \frac{\langle \nabla f(x+\sigma u) - \nabla f_\sigma(x), h_x \rangle}{\|h_x\|} \right| \\ = \lim_{t \rightarrow 0} \left| \frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle \nabla f(x+\sigma u), th_x \rangle}{\|th_x\|} \right| = 0, \end{aligned}$$

which holds thanks to the differentiability of  $f$  at  $x+\sigma u$  when  $u \in \mathfrak{D}_\sigma$ , and (b) follows from the definition of  $h_x$ . This contradicts the fact that  $h_x \neq 0$ . Thus, (3.12) holds.

We now prove (ii) by using the integral representation in (3.12) to relate  $\nabla f_\sigma(x)$  to  $\partial_G^\delta f(x)$ . To this end, we let  $M > 0$  and notice that for any  $\sigma > 0$  we have

$$\begin{aligned} \Delta_h &:= \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma}(u)] - \frac{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M}(u)]}{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M}(u)]} \\ &= \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M^c}(u)] \\ (3.18) \quad &+ \left( 1 - \frac{1}{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M}(u)]} \right) \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M}(u)], \end{aligned}$$

where we recall that  $\mathbb{B}_M = \{u \in \mathbb{R}^d : \|u\| \leq M\}$  and  $\mathbb{B}_M^c$  is its complement.

For the first term on the second line of (3.18), we have

$$\begin{aligned} &\|\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x+\sigma u) \cdot \mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M^c}(u)]\| \\ (3.19) \quad &\stackrel{(a)}{\leq} \mathbb{E}_{u \sim \mathcal{N}(0,I)}[(R_1\|x+\sigma u\|^m + R_2) \cdot \mathbf{1}_{\mathfrak{D}_\sigma \cap \mathbb{B}_M^c}(u)] \stackrel{(b)}{\leq} \Xi_1 + \Xi_2 \end{aligned}$$

with

$$\begin{aligned}\Xi_1 &= (2^{m-1}R_1\|x\|^m + R_2)\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)], \\ \Xi_2 &= (2^{m-1}R_1\sigma^m)\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m \cdot \mathbb{1}_{\mathbb{B}_M^c}(u)],\end{aligned}$$

where we invoked (3.1) in (a), and used the convexity of  $\|\cdot\|^m$  when  $m \geq 1$  and the fact that  $R_1 = 0$  when  $m = 0$ . Now, observe that

$$\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m \cdot \mathbb{1}_{\mathbb{B}_M^c}(u)] \leq \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m] \leq (m+d)^{\frac{m}{2}},$$

where the second inequality follows from [30, Lemma 1]. Thus, for  $m \geq 1$ , if we choose a finite positive  $\sigma$  such that  $\sigma \leq [\frac{\varepsilon}{2^{m+1}R_1(m+d)^{\frac{m}{2}}}]^{\frac{1}{m}}$ , then

$$(3.20) \quad \Xi_2 = (2^{m-1}R_1\sigma^m)\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m \cdot \mathbb{1}_{\mathbb{B}_M^c}(u)] \leq 0.25\varepsilon \leq 0.25\varepsilon(1 + \|x\|^m).$$

As for  $m = 0$ , since  $R_1 = 0$ , we conclude that for any  $\sigma > 0$ ,  $\Xi_2 = 0 \leq 0.25\varepsilon(1 + \|x\|^m)$ . Thus, for  $m \geq 0$ , if  $\sigma \in (0, \infty)$  and  $\sigma \leq [\frac{\varepsilon}{2^{m+1}R_1(m+d)^{\frac{m}{2}}}]^{\frac{1}{m}}$ , then (3.20) holds.

Next, choose  $\eta = \varepsilon(2\pi)^{\frac{d}{2}}(4\max\{2^{m-1}R_1, R_2\})^{-1}$ . Then, by setting  $\nu = \eta$  in Lemma 3.5, we see that whenever  $M \geq [-d \cdot W_{-1}(-\eta^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$ ,

$$(3.21) \quad \begin{aligned}\Xi_1 &= (2^{m-1}R_1\|x\|^m + R_2)\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)] \\ &\leq \max\{2^{m-1}R_1, R_2\}(1 + \|x\|^m)\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)] \leq 0.25\varepsilon(1 + \|x\|^m).\end{aligned}$$

Additionally, for the above  $\eta$  and the  $\eta_1$  in (3.15), we have

$$\eta_1 \leq \varepsilon[4R_2 + 2^{m+1}R_1(m+d)^{\frac{m}{2}}]^{-1} \leq \eta.$$

Combining this conclusion with (3.19), (3.20), and (3.21), we know that for any finite positive  $\sigma \leq [\frac{\varepsilon}{2^{m+1}R_1(m+d)^{\frac{m}{2}}}]^{\frac{1}{m}}$  and  $M \geq [-d \cdot W_{-1}(-\eta_1^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$  with  $\eta_1$  as in (3.15), we have

$$(3.22) \quad \|\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x + \sigma u) \cdot \mathbb{1}_{\mathcal{D}_\sigma \cap \mathbb{B}_M^c}(u)]\| \leq 0.5\varepsilon(1 + \|x\|^m).$$

We next estimate the second term on the second line of (3.18). We first notice that  $\int_{\mathbb{B}_M} e^{-\|u\|^2/2} du = (2\pi)^{\frac{d}{2}} - \int_{\mathbb{B}_M^c} e^{-\|u\|^2/2} du$ . So, if  $\int_{\mathbb{B}_M^c} e^{-\|u\|^2/2} du \leq (2\pi)^{\frac{d}{2}} - 0.5$ , then  $\int_{\mathbb{B}_M} e^{-\|u\|^2/2} du \geq 0.5$ . In view of Lemma 3.5 and the definition of  $\eta_1$ , this happens when we choose  $M \geq [-d \cdot W_{-1}(-\eta_1^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$ , since  $\eta_1 \leq (2\pi)^{\frac{d}{2}} - 0.5$ .

On top of this choice of  $M$ , if we further choose  $\sigma \leq 1$ , then the term in the last line of (3.18) can be upper bounded as follows:

$$(3.23) \quad \begin{aligned}& \left\| \left( 1 - \frac{1}{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M}(u)]} \right) \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\nabla f(x + \sigma u) \cdot \mathbb{1}_{\mathcal{D}_\sigma \cap \mathbb{B}_M}(u)] \right\| \\& \leq \frac{\tilde{K}}{\mathbb{E}_{u \sim \mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M}(u)]} \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|\nabla f(x + \sigma u)\| \cdot \mathbb{1}_{\mathcal{D}_\sigma \cap \mathbb{B}_M}(u)] \\& \stackrel{(a)}{\leq} 2(2\pi)^{\frac{d}{2}} \tilde{K} \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|\nabla f(x + \sigma u)\| \cdot \mathbb{1}_{\mathcal{D}_\sigma \cap \mathbb{B}_M}(u)] \\& \stackrel{(b)}{\leq} 2(2\pi)^{\frac{d}{2}} \tilde{K} \mathbb{E}_{u \sim \mathcal{N}(0,I)}[R_1\|x + \sigma u\|^m + R_2] \\& \stackrel{(c)}{\leq} 2(2\pi)^{\frac{d}{2}} \tilde{K} [(2^{m-1}R_1\|x\|^m + R_2) + 2^{m-1}R_1\sigma^m \mathbb{E}_{u \sim \mathcal{N}(0,I)}[\|u\|^m]] \\& \stackrel{(d)}{\leq} 2(2\pi)^{\frac{d}{2}} \tilde{K} [(2^{m-1}R_1\|x\|^m + R_2) + 2^{m-1}R_1(m+d)^{\frac{m}{2}}] \\& \leq 2(2\pi)^{\frac{d}{2}} \tilde{K} [R_2 + 2^{m-1}R_1(m+d)^{\frac{m}{2}}](1 + \|x\|^m),\end{aligned}$$

where  $\tilde{K} := \mathbb{E}_{u \sim \mathcal{N}(0, I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)]$ , and (a) holds because  $(2\pi)^{\frac{d}{2}} \mathbb{E}_{u \sim \mathcal{N}(0, I)}[\mathbb{1}_{\mathbb{B}_M}(u)] = \int_{\mathbb{B}_M} e^{-\|u\|^2/2} du \geq 0.5$ , (b) holds upon using (3.1) and enlarging the domain of integration, (c) follows from the convexity of  $\|\cdot\|^m$  when  $m \geq 1$  and the inequality holds trivially as an equality when  $m = 0$  because  $R_1 = 0$ , and (d) follows from [30, Lemma 1] and the choice that  $\sigma \leq 1$ . In view of (3.23), we can now invoke Lemma 3.5 to deduce that if we choose  $M \geq [-d \cdot W_{-1}(-\eta_1^{\frac{2}{d}}/(2\pi e))]^{\frac{1}{2}}$  with  $\eta_1$  as in (3.15) and  $\sigma \leq 1$ , then

$$(3.24) \quad \left\| \left(1 - \frac{1}{\mathbb{E}_{u \sim \mathcal{N}(0, I)}[\mathbb{1}_{\mathbb{B}_M}(u)]}\right) \mathbb{E}_{u \sim \mathcal{N}(0, I)}[\nabla f(x + \sigma u) \cdot \mathbb{1}_{\mathcal{D}_{\sigma} \cap \mathbb{B}_M}(u)] \right\| \leq \frac{\varepsilon}{2}(1 + \|x\|^m).$$

Thus, we conclude that (3.22) and (3.24) will both hold as long as we choose  $M \geq H$  defined as in (3.14) and  $\sigma \leq \tilde{\sigma} = \min\left\{\left[\frac{\varepsilon}{2^{m+1}R_1(m+d)\frac{m}{2}}\right]^{\frac{1}{m}}, 1\right\}$ . Hence, we have

$$(3.25) \quad \|\Delta_h\| \leq \varepsilon(1 + \|x\|^m) \quad \text{whenever } M \geq H \text{ and } \sigma \leq \tilde{\sigma},$$

where  $\Delta_h$  is defined as in (3.18).

Finally, for  $M = H$  and any  $\sigma \leq \tilde{\sigma} = \min\left\{\left[\frac{\varepsilon}{2^{m+1}R_1(m+d)\frac{m}{2}}\right]^{\frac{1}{m}}, 1, \frac{\delta}{H}\right\}$ , we have  $\sigma M \leq \delta$ . Hence, by (2.1) and the definition of Goldstein  $\delta$ -subdifferential,

$$(3.26) \quad \frac{1}{\mathbb{E}_{u \sim \mathcal{N}(0, I)}[\mathbb{1}_{\mathcal{D}_{\sigma} \cap \mathbb{B}_M}(u)]} \mathbb{E}_{u \sim \mathcal{N}(0, I)}[\nabla f(x + \sigma u) \cdot \mathbb{1}_{\mathcal{D}_{\sigma} \cap \mathbb{B}_M}(u)] \in \partial_G^\delta f(x).$$

The desired conclusion follows from (3.12), (3.26), (3.25), and the definition of  $\Delta_h$  in (3.18).  $\square$

*Remark 3.7* (simplified expressions for the choice of  $\sigma$ ). We present more explicit upper bounds for  $\sigma$  in Theorem 3.6(ii). Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1). Let

$$(3.27) \quad 0 < \varepsilon < \min\{5R_2, 1\} \quad \text{and} \quad 0 < \delta < 1.$$

To provide an estimate on the corresponding  $\bar{\sigma}$  in (3.13), we define

$$(3.28) \quad \mathfrak{M}_1 = [(2\pi)^{\frac{d}{2}} - 0.5]\mathcal{P}, \quad \mathfrak{M}_2 = (\pi e/5)^{\frac{d}{2}}\mathcal{P},$$

where  $\mathcal{P}$  is defined as in (3.16). Since  $\mathcal{P} \geq 4R_2$ , we can deduce that

$$\mathfrak{M}_1 \geq 5R_2 \quad \text{and} \quad \mathfrak{M}_2 \geq (\pi e/5)^{d/2} 4R_2 \geq 5R_2.$$

This means that for the  $\varepsilon$  and  $\delta$  chosen as in (3.27), we indeed have

$$(3.29) \quad 0 < \varepsilon < \min\{\mathfrak{M}_1, \mathfrak{M}_2, 1\} \quad \text{and} \quad 0 < \delta < 1.$$

Using the definitions of  $\mathfrak{M}_1$  and  $\mathfrak{M}_2$  above and the definition of  $\eta_1$  in (3.15), we can then deduce that<sup>6</sup>

$$(3.30) \quad \eta_1 = \varepsilon \mathcal{P}^{-1} \quad \text{and} \quad 0 < \eta_1^{\frac{2}{d}}/(2\pi e) < 1/10 < 1/e.$$

<sup>6</sup>Specifically, we deduce from  $\varepsilon < \mathfrak{M}_1$  and the definition of  $\eta_1$  that  $\eta_1 = \varepsilon \mathcal{P}^{-1}$ , and then from  $\varepsilon < \mathfrak{M}_2$  and the definition of  $\eta_1$  that  $\eta_1^{\frac{2}{d}}/(2\pi e) < 1/10$ .

Next, recall that for  $0 < h < \frac{1}{10}$ , it holds that

$$W_{-1}(-h) \geq \frac{e}{e-1} \ln(h) \geq \frac{e}{e-1} \frac{0.1 \ln(0.1)}{h} > -\frac{1}{2h},$$

where the first inequality follows from [25, eqn. (8)], and the second inequality holds because  $t \mapsto t \ln t$  is decreasing on  $[0, 0.1]$ . The above display together with (3.30) implies that for the  $H$  given in (3.14),

$$(3.31) \quad H \leq \sqrt{d\pi e} \eta_1^{-\frac{1}{d}}.$$

Finally, since  $H > \sqrt{d}$  (see footnote 5) and we chose  $0 < \delta < 1$  as stated in (3.27), it follows from (3.13) that  $\bar{\sigma} = \min\{\lceil \frac{\varepsilon}{2^{m+1}R_1(m+d)^{\frac{m}{2}}} \rceil^{\frac{1}{m}}, \frac{\delta}{H}\}$ . Therefore, for the  $\varepsilon$  and  $\delta$  chosen as in (3.27), upon combining (3.29), (3.30), and (3.31) and recalling that  $R_1 = 0$  if and only if  $m = 0$ , we have that the inclusion in Theorem 3.6(ii) holds whenever

$$(3.32) \quad \sigma \leq \begin{cases} \min\{[2^{m+1}R_1(m+d)^{\frac{m}{2}}]^{-\frac{1}{m}}, \delta \mathcal{P}^{-1/d}/\sqrt{d\pi e}\} \cdot \varepsilon^{\max\{\frac{1}{m}, \frac{1}{d}\}} & \text{if } m \geq 1, \\ \delta \mathcal{P}^{-1/d} \varepsilon^{\frac{1}{d}}/\sqrt{d\pi e} & \text{if } m = 0. \end{cases}$$

Let  $\zeta$  be an accumulation point of the set  $\{\nabla f_\sigma(x)\}_{\sigma>0}$  as  $\sigma \rightarrow 0^+$ . Since  $\partial_G^\delta f(x)$  and  $(1+\|x\|^m)\varepsilon \cdot \mathbb{B}$  are both compact,  $\partial_G^\delta f(x) + (1+\|x\|^m)\varepsilon \cdot \mathbb{B}$  is closed. Theorem 3.6 then implies that  $\zeta \in \partial_G^\delta f(x) + (1+\|x\|^m)\varepsilon \cdot \mathbb{B}$ . Using the definition of Clarke subdifferential and limiting arguments, we obtain the following corollary.

**COROLLARY 3.8** (GS gradient consistency). *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  and let  $\nabla f_\sigma$  and  $\partial_C f$  be given as in (3.3) and (2.1), respectively. Then for each  $x \in \mathbb{R}^d$ , every accumulation point of  $\{\nabla f_\sigma(x)\}_{\sigma>0}$  as  $\sigma \rightarrow 0^+$  belongs to  $\partial_C f(x)$ .*

**4. GS-based algorithms for minimizing SPB functions.** In this section, we consider the optimization problem

$$(4.1) \quad \begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f(x) \\ \text{s.t.} \quad & x \in \Omega, \end{aligned}$$

where  $\Omega \subseteq \mathbb{R}^d$  is a closed convex set with an easy-to-compute projection, and  $f$  is an SPB function that can only be accessed through its zeroth-order oracle (i.e., a routine for computing the function value at any prescribed point). Problem (4.1) therefore falls into the category of zeroth-order (or derivative-free) optimization problems; see, e.g., [11, 23] and references therein for classical works and recent developments on zeroth-order optimization.

From Theorem 3.3, for any  $x \in \mathbb{R}^d$  and  $\sigma > 0$ , both  $f(x + \sigma u)u/\sigma$  and  $(f(x + \sigma u) - f(x))u/\sigma$  are unbiased estimators of the GS gradient  $\nabla f_\sigma(x)$ , where  $u$  is the standard Gaussian random vector. If the smoothing parameter  $\sigma$  is small, they can serve as random ascent directions for  $f$  at  $x$ . Observing that they can be computed with one or two evaluations of  $f$ , it is then tempting to develop zeroth-order algorithms based on these gradient estimators. Indeed, there is a wealth of literature on such zeroth-order algorithms; see, e.g., [1, 2, 21, 26, 30]. We refer to them as GS-based algorithms. Most existing works rely on the global Lipschitz continuity of  $\nabla f_\sigma$ , which not only offers an obvious choice of stepsize but also greatly facilitates the convergence analysis. One technical novelty in this paper lies in the convergence analysis of GS-based zeroth-order algorithms for SPB functions that do not possess a globally Lipschitz GS gradient.

A core idea for the design of our GS-based algorithms is as follows. For any SPB  $f$ , Theorem 3.4(ii) asserts that  $\nabla f_\sigma$  is locally Lipschitz with a polynomially bounded Lipschitz modulus of order  $O(\|x\|^m + 1)$  for some  $m \geq 0$ . This naturally suggests the adaptive stepsize that scales as  $(\|x\|^m + 1)^{-1}$ . Based on this observation, we will develop several algorithms for different subclasses of problem (4.1) in subsequent subsections.

We will develop algorithms for problem (4.1) under two different settings: the constrained convex setting (where the objective function  $f$  is convex) and the unconstrained nonconvex setting (where the feasible region  $\Omega = \mathbb{R}^d$ ). Before that, we first establish some auxiliary lemmas, which will be useful for the algorithmic analysis.

**4.1. Auxiliary lemmas for complexity analysis.** A key instrument for convergence analysis of optimization algorithms is the descent lemma, which is often proved for functions with globally Lipschitz gradients; see, e.g., [28, Theorem 2.1.5]. We present a descent lemma of  $f_\sigma$  for SPB functions  $f$ , whose gradients  $\nabla f_\sigma$  are not necessarily globally Lipschitz.

**LEMMA 4.1** (descent lemma). *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1), let  $f_\sigma$  be defined as in Definition 1.1, and let  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  be given as in (3.7). Then*

$$f_\sigma(x) - f_\sigma(y) \leq \langle \nabla f_\sigma(y), x - y \rangle + \left[ \frac{\mathcal{A} + \mathcal{B}\|y\|^m}{2} + \frac{\mathcal{C}\|x - y\|^m}{m + 2} \right] \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

*Proof.* For any  $x, y \in \mathbb{R}^d$ , we have

$$\begin{aligned} f_\sigma(x) - f_\sigma(y) - \langle \nabla f_\sigma(y), x - y \rangle &= \int_0^1 \langle \nabla f_\sigma(y + t(x - y)) - \nabla f_\sigma(y), x - y \rangle dt \\ &\leq \int_0^1 \|\nabla f_\sigma(y + t(x - y)) - \nabla f_\sigma(y)\| dt \cdot \|x - y\| \\ &\stackrel{(a)}{\leq} \int_0^1 [\mathcal{A} + \mathcal{B}\|y\|^m + \mathcal{C}\|t(x - y)\|^m] t dt \cdot \|x - y\|^2 = \left[ \frac{\mathcal{A} + \mathcal{B}\|y\|^m}{2} + \frac{\mathcal{C}\|x - y\|^m}{m + 2} \right] \|x - y\|^2, \end{aligned}$$

where (a) follows from Theorem 3.4(ii).  $\square$

We remark that another descent lemma for functions without a Lipschitz gradient has been formulated and studied in a very recent work (see [27, Definition 2.1]) based on the notion of directional smoothness and utilized to analyze gradient descent. Using their descent lemma, they further proposed an adaptive stepsize for gradient descent, which is implicitly defined by a nonlinear equation involving the current and *next* iterates and requires a root-finding procedure to compute. In contrast, the adaptive stepsize derived from Lemma 4.1 (see Theorems 4.5 and 4.8 below) only depends on the current iterate and is given by an *explicit* formula.

The next lemma quantifies the approximation error of  $f_\sigma$ .

**LEMMA 4.2.** *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1) and let  $f_\sigma$  be defined as in Definition 1.1. Then it holds that*

$$|f_\sigma(x) - f(x)| \leq \mathcal{M}(x) \cdot \sigma \quad \forall x \in \mathbb{R}^d,$$

where  $\mathcal{M} : \mathbb{R}^d \rightarrow \mathbb{R}_+$  is the function

$$\mathcal{M}(x) := (2^{m-1}R_1\|x\|^m + R_2)\sqrt{d} + 2^{m-1}R_1\sigma^m(m+1+d)^{\frac{m+1}{2}}.$$

*Proof.* Notice that for all  $x \in \mathbb{R}^d$ , we have

$$\begin{aligned} |f_\sigma(x) - f(x)| &\leq \mathbb{E}_{u \sim \mathcal{N}(0, I)} [|f(x + \sigma u) - f(x)|] \\ &\leq \mathbb{E}_{u \sim \mathcal{N}(0, I)} [(2^{m-1} R_1 \|x\|^m + 2^{m-1} R_1 \sigma^m \|u\|^m + R_2) \cdot \sigma \|u\|] \\ &\leq [(2^{m-1} R_1 \|x\|^m + R_2) \sqrt{d} + 2^{m-1} R_1 \sigma^m (m+1+d)^{\frac{m+1}{2}}] \cdot \sigma = \mathcal{M}(x) \cdot \sigma, \end{aligned}$$

where the second inequality follows from Lemma 3.2 and the last inequality follows from [30, Lemma 1].  $\square$

The next two lemmas concern the random vector  $(\frac{f(x+\sigma u)-f(x)}{\sigma})u$  with  $u \sim \mathcal{N}(0, I)$ .

LEMMA 4.3. *Let  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1, R_2$ , and  $m$  as in (3.1), let  $x \in \mathbb{R}^d$  and  $\sigma > 0$ , and let  $p$  be a nonnegative integer. Define  $F(u) = \frac{1}{\sigma} [f(x + \sigma u) - f(x)]u$ . Then*

$$(4.2) \quad \mathbb{E}_{u \sim \mathcal{N}(0, I)} [\|F(u)\|^p] \leq \mathcal{H}_{(p)}(\|x\|^{mp} + 1),$$

where  $\mathcal{H}_{(\cdot)}$  is the function such that  $\mathcal{H}_{(0)} := 1$ , and when  $p \geq 1$ ,

$$\mathcal{H}_{(p)} := 3^{p-1} \max\{2^{(m-1)p} R_1^p (2p+d)^p, R_2^p (2p+d)^p + 2^{(m-1)p} R_1^p \sigma^{mp} [(m+2)p+d]^{\frac{(m+2)p}{2}}\}.$$

*Proof.* Note that (4.2) holds trivially for  $p = 0$ . However, for  $p \geq 1$ , from Lemma 3.2, for any  $u \in \mathbb{R}^d$ , we have

$$|f(x + \sigma u) - f(x)| \leq (2^{m-1} R_1 \|x\|^m + 2^{m-1} R_1 \sigma^m \|u\|^m + R_2) \sigma \|u\|.$$

It follows from the convexity of  $\|\cdot\|^p$  that

$$\begin{aligned} \left| \frac{f(x + \sigma u) - f(x)}{\sigma} \right|^p \|u\|^p &\leq [3^{p-1} 2^{(m-1)p} R_1^p \|x\|^{mp} + 3^{p-1} R_2^p] \|u\|^{2p} + 3^{p-1} 2^{(m-1)p} R_1^p \sigma^{mp} \|u\|^{(m+2)p}. \end{aligned}$$

Taking the expectation on both sides of the above display with respect to  $u \sim \mathcal{N}(0, I)$  and invoking [30, Lemma 1] for upper bounding moments of the form  $\mathbb{E}_{u \sim \mathcal{N}(0, I)} [\|u\|^k]$  for  $k \geq 1$ , we see that  $\mathbb{E}_{u \sim \mathcal{N}(0, I)} [\|F(u)\|^p]$  can be bounded from above by

$$[3^{p-1} 2^{(m-1)p} R_1^p \|x\|^{mp} + 3^{p-1} R_2^p] (2p+d)^p + 3^{p-1} 2^{(m-1)p} R_1^p \sigma^{mp} [(m+2)p+d]^{(m+2)p/2}.$$

The relation (4.2) now follows immediately.  $\square$

LEMMA 4.4. *Let  $x$  be a random vector and  $g$  be a nonnegative, lower semicontinuous function. Assume that*

$$(4.3) \quad \mathbb{E}_x [\|x\|^2] \leq \beta_c \quad \text{and} \quad \mathbb{E}_x \left[ \frac{g(x)}{1 + \|x\|^n} \right] \leq \alpha_c$$

for some integer  $n \geq 1$  and positive numbers  $\alpha_c$  and  $\beta_c$ . Then

$$\mathbb{E}_x \left[ g(x)^{\frac{1}{2\lceil n/2 \rceil}} \right] \leq \left( 1 + \sqrt{\beta_c} \right) (2\alpha_c)^{\frac{1}{2\lceil n/2 \rceil}}.$$

---

**Algorithm 4.1** GS-based zeroth-order algorithm for convex problem (4.1).

---

- 1: **Input:** Initial point  $x^0 \in \Omega$ ,  $\{\tau_k\} \subset (0, 1]$  and  $\sigma > 0$ . Let  $m$  be defined as in (3.1) corresponding to our  $f \in \mathcal{SPB}(\mathbb{R}^d)$ .
- 2: **for**  $k = 0, 1, 2, \dots$  **do**
- 3:     Generate  $u^k \sim \mathcal{N}(0, I)$  and form  $v^k = \frac{1}{\sigma}[f(x^k + \sigma u^k) - f(x^k)]u^k$ .
- 4:     Compute

$$x^{k+1} = P_\Omega \left( x^k - \tau_k \cdot \frac{v^k}{\|x^k\|^m + 1} \right).$$

- 5: **end for**

---

*Proof.* Notice that

$$(4.4) \quad \begin{aligned} \mathbb{E}_x \left[ \left( \frac{g(x)^{\frac{1}{2\lceil n/2 \rceil}}}{1 + \|x\|} \right)^2 \right] &\stackrel{(a)}{\leq} \mathbb{E}_x \left[ \frac{g(x)^{\frac{1}{\lceil n/2 \rceil}}}{1 + \|x\|^2} \right] \stackrel{(b)}{\leq} \left( \mathbb{E}_x \left[ \left( \frac{g(x)^{\frac{1}{\lceil n/2 \rceil}}}{1 + \|x\|^2} \right)^{\lceil n/2 \rceil} \right] \right)^{\frac{1}{\lceil n/2 \rceil}} \\ &= \left( \mathbb{E}_x \left[ \frac{g(x)}{(1 + \|x\|^2)^{\lceil n/2 \rceil}} \right] \right)^{\frac{1}{\lceil n/2 \rceil}} \stackrel{(c)}{\leq} \left( 2\mathbb{E}_x \left[ \frac{g(x)}{1 + \|x\|^n} \right] \right)^{\frac{1}{\lceil n/2 \rceil}} \leq (2\alpha_c)^{\frac{1}{\lceil n/2 \rceil}}, \end{aligned}$$

where (a) holds because  $1 + \|x\|^2 \leq (1 + \|x\|)^2$ , (b) follows from Jensen's inequality, (c) holds since  $1 + \|x\|^n \leq 2(1 + \|x\|^{2\lceil n/2 \rceil}) \leq 2(1 + \|x\|^2)^{\lceil n/2 \rceil}$ , and the last inequality follows from (4.3). On the other hand, note that  $\mathbb{E}_x[\|x\|^2] \leq \beta_c$  implies  $\mathbb{E}_x[\|x\|] \leq \sqrt{\beta_c}$  and hence  $\mathbb{E}_x[(1 + \|x\|)^2] \leq (1 + \sqrt{\beta_c})^2$ . Combining this with (4.4) gives

$$(1 + \sqrt{\beta_c})^2 (2\alpha_c)^{\frac{1}{\lceil n/2 \rceil}} \geq \mathbb{E}_x[(1 + \|x\|)^2] \cdot \mathbb{E}_x \left[ \left( \frac{g(x)^{\frac{1}{2\lceil n/2 \rceil}}}{1 + \|x\|} \right)^2 \right] \geq \left( \mathbb{E}_x \left[ g(x)^{\frac{1}{2\lceil n/2 \rceil}} \right] \right)^2. \quad \square$$

**4.2. Convex SPB minimization.** Here, we assume in addition that the objective function  $f$  is convex, rendering problem (4.1) a convex optimization problem. The specific algorithm for this case is presented in Algorithm 4.1. Algorithm 4.1 can be seen as a natural extension of the algorithm in [30, eqn. (39)]; indeed, when  $m = 0$ , our algorithm essentially reduces to their algorithm (with an extra scaling factor of 0.5 in our stepsize). Here, we scale the stepsize by  $\|x^k\|^m + 1$  to account for the lack of Lipschitzness.

The following result establishes the convergence rate of Algorithm 4.1. Due to the weak assumptions,  $\{x^k\}$  can be unbounded in general. Therefore, a *relative* optimality measure that is rescaled by  $\|x^k\|^m + 1$  is adopted. Complexity result based on the standard measure  $f(x^k) - f(x^*)$  will be derived in Corollary 4.7 under additional assumptions.

**THEOREM 4.5** (complexity bound for Algorithm 4.1). *Consider problem (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1, R_2$ , and  $m$  as in (3.1). Assume additionally that  $f$  is convex and there exists an optimal solution  $x^*$  for (4.1). Then the sequence  $\{x^k\}$  generated by Algorithm 4.1 satisfies that*

$$(4.5) \quad 0 \leq \mathbb{E}[\|x^k - z\|^2] - \mathbb{E}[\|x^{k+1} - z\|^2] + \mathcal{H}_{(2)} \tau_k^2 \quad \forall z \in \arg \min_{u \in \Omega} f_\sigma(u),$$

and for any  $T > 0$ ,



$$(4.6) \quad \min_{0 \leq k \leq T} \mathbb{E} \left[ \frac{f(x^k) - f(x^*)}{\|x^k\|^m + 1} \right] \leq \frac{1}{2 \sum_{k=0}^T \tau_k} \left[ \|x^0 - x^*\|^2 + \mathcal{H}_{(2)} \sum_{k=0}^T \tau_k^2 \right] + \mathcal{M}(x^*) \cdot \sigma,$$

where  $\mathcal{M}(\cdot)$  and  $\mathcal{H}_{(\cdot)}$  are defined as in Lemma 4.2 and Lemma 4.3, respectively.

*Proof.* Since the projection operator is nonexpansive, we see that for any  $z \in \Omega$ ,

$$\begin{aligned} \|x^{k+1} - z\|^2 &= \|P_\Omega(x^k - \alpha_k v^k) - z\|^2 \leq \|x^k - z - \alpha_k v^k\|^2 \\ &= \|x^k - z\|^2 - 2\alpha_k \langle v^k, x^k - z \rangle + \alpha_k^2 \|v^k\|^2, \end{aligned}$$

where  $\alpha_k = \tau_k / (\|x^k\|^m + 1)$ . Taking the expectation on both sides of the above inequality with respect to the random variable  $v^k$ , we can obtain from (3.3) that

$$(4.7) \quad 2\alpha_k \langle \nabla f_\sigma(x^k), x^k - z \rangle \leq \|x^k - z\|^2 - \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|x^{k+1} - z\|^2] + \alpha_k^2 \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^2].$$

Next, by Definition 1.1 and the convexity of  $f$ ,  $f_\sigma$  is also convex. Then we have the following subgradient inequality:

$$(4.8) \quad f_\sigma(x^k) - f_\sigma(z) \leq \langle \nabla f_\sigma(x^k), x^k - z \rangle.$$

In addition, we have from Lemma 4.3 with  $p = 2$  that

$$(4.9) \quad \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^2] \leq \mathcal{H}_{(2)} (\|x^k\|^{2m} + 1) \leq \mathcal{H}_{(2)} (\|x^k\|^m + 1)^2.$$

Combining (4.7), (4.8), and (4.9) yields

$$(4.10) \quad 2\alpha_k [f_\sigma(x^k) - f_\sigma(z)] \leq \|x^k - z\|^2 - \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|x^{k+1} - z\|^2] + \mathcal{H}_{(2)} \tau_k^2,$$

which gives (4.5) upon taking expectation on both sides.

We now lower bound the left-hand side of (4.10) for  $z = x^*$ . To this end, we first note from [30, eqn. (11)] that

$$(4.11) \quad f_\sigma(x^k) \geq f(x^k).$$

Also, we have

$$(4.12) \quad f_\sigma(x^*) = f_\sigma(x^*) - f(x^*) + f(x^*) \leq \mathcal{M}(x^*) \cdot \sigma + f(x^*),$$

where the inequality follows from Lemma 4.2. Using (4.11) and (4.12), we can lower bound the left-hand side of (4.5), which in turn yields

$$\mathbb{E}[2\alpha_k (f(x^k) - f(x^*))] \leq \mathbb{E}[\|x^k - x^*\|^2] - \mathbb{E}[\|x^{k+1} - x^*\|^2] + 2\mathcal{M}(x^*)\sigma \cdot \alpha_k + \mathcal{H}_{(2)} \tau_k^2.$$

Finally, invoking the definition of  $\alpha_k$  (and noting also that  $\alpha_k \leq \tau_k$ ), we deduce from the above display that

$$2\tau_k \mathbb{E} \left[ \frac{f(x^k) - f(x^*)}{\|x^k\|^m + 1} \right] \leq \mathbb{E}[\|x^k - x^*\|^2] - \mathbb{E}[\|x^{k+1} - x^*\|^2] + 2\mathcal{M}(x^*)\sigma \cdot \tau_k + \mathcal{H}_{(2)} \tau_k^2,$$

which, upon summing over  $k$ , completes the proof.  $\square$

*Remark 4.6* (comparing existing complexity results for (4.1) with a convex  $f$ ). When  $m = 0$ ,  $R_1 = 0$ . We then see from Lemma 4.3 that  $\mathcal{H}_{(2)} = 3R_2^2(4 + d)^2$  and from Lemma 4.2 that  $\mathcal{M}(x^*) = R_2\sqrt{d}$ . These together with (4.6) give a bound that matches the one obtained in [30, Theorem 6] up to a constant scaling factor.

Theorem 4.5 gives a bound on  $\mathbb{E} \left[ \frac{f(x^k) - f(x^*)}{\|x^k\|^m + 1} \right]$ , which can be regarded as a relative optimality measure when  $m \geq 1$ . The next corollary shows that under suitable assumptions on  $f$ , one can derive a bound on an *absolute* optimality measure.

COROLLARY 4.7. Consider problem (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1, R_2$ , and  $m \geq 1$  as in (3.1). Assume that  $f$  is convex and level-bounded, and let  $x^*$  be a minimizer for (4.1). Let  $\gamma \in (0, 1]$  and  $T$  be a positive integer. If  $\tau_k = \gamma/\sqrt{T+1}$  for  $k = 0, \dots, T$ , then the sequence  $\{x^k\}$  generated by Algorithm 4.1 satisfies

$$\min_{0 \leq k \leq T} \mathbb{E}[(f(x^k) - f(x^*))^{(\frac{1}{2\lceil m/2 \rceil})}] \leq (1 + \sqrt{M_{\text{bd}}})(2C_{\text{bd}}T^{-1/2} + 2\mathcal{M}(x^*)\sigma)^{\frac{1}{(2\lceil m/2 \rceil)}},$$

where

$$(4.13) \quad M_{\text{bd}} = 4\|x^0\|^2 + 6C_{\text{lev}}^2 + 2\mathcal{H}_{(2)}\gamma^2, \quad C_{\text{bd}} = \frac{1}{2\gamma}(\|x^0 - x^*\|^2 + \mathcal{H}_{(2)}\gamma^2),$$

$$(4.14) \quad C_{\text{lev}} = \sup\{\|x\| : f(x) \leq f(x^*) + \mathcal{M}(x^*)\sigma\} < \infty,$$

$\mathcal{M}(\cdot)$  is defined as in Lemma 4.2, and  $\mathcal{H}_{(\cdot)}$  is defined as in Lemma 4.3.<sup>7</sup>

*Proof.* Let  $k_* \in \arg \min_{0 \leq k \leq T} \mathbb{E}[\frac{f(x^k) - f(x^*)}{\|x^k\|^{m+1}}]$ . From Theorem 4.5, we have

$$(4.15) \quad \begin{aligned} \mathbb{E}\left[\frac{f(x^{k_*}) - f(x^*)}{\|x^{k_*}\|^{m+1}}\right] &\leq \frac{1}{2\sum_{k=0}^T \tau_k} \left[ \|x^0 - x^*\|^2 + \mathcal{H}_{(2)} \sum_{k=0}^T \tau_k^2 \right] + \mathcal{M}(x^*)\sigma \\ &\stackrel{(a)}{\leq} \frac{1}{2\gamma\sqrt{T}} (\|x^0 - x^*\|^2 + \mathcal{H}_{(2)}\gamma^2) + \mathcal{M}(x^*)\sigma \leq \frac{C_{\text{bd}}}{\sqrt{T}} + \mathcal{M}(x^*)\sigma, \end{aligned}$$

where we used the fact that  $\tau_k = \gamma/\sqrt{T+1}$  in (a), and  $C_{\text{bd}}$  is defined as in (4.13).

Next, pick any  $z^* \in \arg \min_{x \in \Omega} f_\sigma(x)$ .<sup>8</sup> Then we see from (4.5) that

$$(4.16) \quad 0 \leq \mathbb{E}[\|x^k - z^*\|^2] - \mathbb{E}[\|x^{k+1} - z^*\|^2] + \mathcal{H}_{(2)}\tau_k^2.$$

When  $k_* \geq 1$ , we can sum both sides of (4.16) from  $k = 0$  to  $k_* - 1$  to obtain

$$\mathbb{E}[\|x^{k_*} - z^*\|^2] \leq \|x^0 - z^*\|^2 + \mathcal{H}_{(2)} \sum_{i=0}^{k_*-1} \tau_i^2 \leq 2\|x^0\|^2 + 2\|z^*\|^2 + \mathcal{H}_{(2)}\gamma^2,$$

where we used the fact that  $\tau_k = \gamma/\sqrt{T+1}$  for the last inequality. The above display further implies that

$$(4.17) \quad \mathbb{E}[\|x^{k_*}\|^2] \leq 2\mathbb{E}[\|x^{k_*} - z^*\|^2] + 2\|z^*\|^2 \leq 4\|x^0\|^2 + 6\|z^*\|^2 + 2\mathcal{H}_{(2)}\gamma^2.$$

Note that the above inequality also holds when  $k_* = 0$ .

Now, since  $z^* \in \arg \min_{x \in \Omega} f_\sigma(x)$ , we have

$$f(z^*) \leq f_\sigma(z^*) \leq f_\sigma(x^*) \leq f(x^*) + \mathcal{M}(x^*)\sigma,$$

where the first inequality follows from [30, eqn. (11)] and the last inequality follows from Lemma 4.2. Thus, it holds that  $\sup\{\|z\| : z \in \arg \min_{x \in \Omega} f_\sigma(x)\} \leq C_{\text{lev}} < +\infty$ , where  $C_{\text{lev}}$  is defined as in (4.14). This observation together with (4.17) implies that

$$(4.18) \quad \mathbb{E}[\|x^{k_*}\|^2] \leq M_{\text{bd}},$$

where  $M_{\text{bd}}$  is defined as in (4.13). The desired conclusion now follows immediately upon combining (4.18) with (4.15) and Lemma 4.4.  $\square$

<sup>7</sup>The finiteness of  $C_{\text{lev}}$  follows from the assumption that  $f$  is level-bounded.

<sup>8</sup>Notice that  $\arg \min_{x \in \Omega} f_\sigma(x)$  is nonempty because the level-boundedness of  $f$  and [30, eqn. (11)] imply the level-boundedness of  $f_\sigma$ .

---

**Algorithm 4.2** GS-based zeroth-order algorithm for unconstrained problem (4.1).

---

- 1: **Input:** Initial point  $x^0 \in \mathbb{R}^d$ ,  $\{\tau_k\} \subset (0, 1]$  and  $\sigma > 0$ . Let  $m$  be defined as in (3.1) corresponding to our  $f \in \mathcal{SPB}(\mathbb{R}^d)$ .
- 2: **for**  $k = 0, 1, 2, \dots$  **do**
- 3:     Generate  $u^k \sim \mathcal{N}(0, I)$  and form  $v^k = \frac{1}{\sigma}[f(x^k + \sigma u^k) - f(x^k)]u^k$ .
- 4:     Compute

$$x^{k+1} = x^k - \tau_k \cdot \frac{v^k}{\|x^k\|^{2m} + 1}.$$

5: **end for**

---

**4.3. Unconstrained SPB minimization.** Here, we consider problem (4.1) with  $\Omega = \mathbb{R}^d$ . The specific algorithm is presented in Algorithm 4.2 below. Notice that the update rule for  $x^k$  differs from that of Algorithm 4.1 in that the stepsize has to be rescaled by  $\|x^k\|^{2m} + 1$  instead of  $\|x^k\|^m + 1$ . This rescaling also makes our algorithm (for SPB functions) a natural extension of the ones in [30, eqn. (66)] and [24, Algorithm 1], which are designed for  $f$  being globally Lipschitz. Indeed, when  $m = 0$ , our algorithm essentially reduces to [30, eqn. (66)] (with an extra factor of 0.5 in our stepsize).

**THEOREM 4.8** (complexity bound for Algorithm 4.2). *Consider (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1, R_2$ , and  $m$  as in (3.1). Let  $\nabla f_\sigma$  be given as in (3.3) and  $\mathcal{A}, \mathcal{B}, \mathcal{C}$  be given as in (3.7). Assume in addition that  $\Omega = \mathbb{R}^d$  and  $\inf f > -\infty$ . Then the sequence  $\{x^k\}$  generated by Algorithm 4.2 satisfies that*

$$(4.19) \quad \mathbb{E}[f_\sigma(x^{k+1})] \leq \mathbb{E}[f_\sigma(x^k)] + \frac{\mathcal{H}_{(2)}\tau_k^2(\mathcal{A} + \mathcal{B})}{2} + \frac{\mathcal{C}}{m+2}\mathcal{H}_{(m+2)}\tau_k^{m+2},$$

and for any  $T > 0$ ,

$$\min_{0 \leq k \leq T} \tilde{w}_k^2 \leq \frac{\mathcal{M}(x^0)\sigma + f(x^0) - \inf f + 0.5\mathcal{H}_{(2)}(\mathcal{A} + \mathcal{B}) \sum_{k=0}^T \tau_k^2 + \frac{\mathcal{H}_{(m+2)}\mathcal{C}}{m+2} \sum_{k=0}^T \tau_k^{m+2}}{\sum_{k=0}^T \tau_k},$$

where  $\tilde{w}_k^2 = \mathbb{E}[\|\frac{\nabla f_\sigma(x^k)}{\|x^k\|^{m+1}}\|^2]$ ,  $\mathcal{M}(\cdot)$  is defined as in Lemma 4.2, and  $\mathcal{H}_{(\cdot)}$  is defined as in Lemma 4.3.

*Proof.* From Lemma 4.1, we know that for any  $k \geq 0$ ,

$$\begin{aligned} f_\sigma(x^{k+1}) &\leq f_\sigma(x^k) - \langle \nabla f_\sigma(x^k), x^k - x^{k+1} \rangle \\ &\quad + \frac{1}{2}\|x^k - x^{k+1}\|^2 \left[ \mathcal{A} + \mathcal{B}\|x^k\|^m + \frac{2\mathcal{C}\|x^{k+1} - x^k\|^m}{m+2} \right]. \end{aligned}$$

This implies that

$$\tilde{\alpha}_k \langle \nabla f_\sigma(x^k), v^k \rangle \leq f_\sigma(x^k) - f_\sigma(x^{k+1}) + \frac{(\mathcal{A} + \mathcal{B}\|x^k\|^m)\tilde{\alpha}_k^2}{2}\|v^k\|^2 + \frac{\mathcal{C}\tilde{\alpha}_k^{m+2}}{m+2}\|v^k\|^{m+2},$$

where

$$(4.20) \quad \tilde{\alpha}_k = \tau_k / (\|x^k\|^{2m} + 1).$$

By taking the expectation on both sides of the above inequality, we can obtain from (3.3) that

$$(4.21) \quad \begin{aligned} \tilde{\alpha}_k \|\nabla f_\sigma(x^k)\|^2 &\leq f_\sigma(x^k) - \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [f_\sigma(x^{k+1})] \\ &\quad + \frac{(\mathcal{A} + \mathcal{B}\|x^k\|^m)\tilde{\alpha}_k^2}{2} \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^2] + \frac{\mathcal{C}\tilde{\alpha}_k^{m+2}}{m+2} \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^{m+2}]. \end{aligned}$$

We now upper bound the two terms  $(\mathcal{A} + \mathcal{B}\|x^k\|^m)\tilde{\alpha}_k^2 \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^2]$  and  $\tilde{\alpha}_k^{m+2} \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^{m+2}]$  in (4.21). For the former term, we have

$$\begin{aligned} (\mathcal{A} + \mathcal{B}\|x^k\|^m)\tilde{\alpha}_k^2 \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^2] &\leq (\mathcal{A} + \mathcal{B}\|x^k\|^m)\tilde{\alpha}_k^2 \mathcal{H}_{(2)}(\|x^k\|^{2m} + 1) \\ &\leq (\mathcal{A} + \mathcal{B} + \mathcal{B}\|x^k\|^{2m})\tilde{\alpha}_k^2 \mathcal{H}_{(2)}(\|x^k\|^{2m} + 1) \\ &\leq \mathcal{H}_{(2)} \cdot (\mathcal{A} + \mathcal{B}) \cdot (\|x^k\|^{2m} + 1)^2 \tilde{\alpha}_k^2 = \mathcal{H}_{(2)} \tau_k^2 (\mathcal{A} + \mathcal{B}), \end{aligned}$$

where the first inequality follows from Lemma 4.3 with  $p = 2$ , and we used the definition of  $\tilde{\alpha}_k$  in (4.20) for the equality. As for the latter term (i.e.,  $\tilde{\alpha}_k^{m+2} \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^{m+2}]$ ), we can deduce from Lemma 4.3 with  $p = m + 2$  that

$$\begin{aligned} \tilde{\alpha}_k^{m+2} \mathbb{E}_{u^k \sim \mathcal{N}(0, I)} [\|v^k\|^{m+2}] &\leq \tilde{\alpha}_k^{m+2} \mathcal{H}_{(m+2)} [\|x^k\|^{m(m+2)} + 1] \\ &\leq \tilde{\alpha}_k^{m+2} \mathcal{H}_{(m+2)} [\|x^k\|^{2m} + 1]^{(m+2)/2} \leq \mathcal{H}_{(m+2)} \tau_k^{m+2}. \end{aligned}$$

Combining (4.21) with the above two displays, one has

$$\begin{aligned} \tau_k \mathbb{E} \left[ \left\| \frac{\nabla f_\sigma(x^k)}{\|x^k\|^m + 1} \right\|^2 \right] &\stackrel{(a)}{\leq} \mathbb{E} [\tilde{\alpha}_k \|\nabla f_\sigma(x^k)\|^2] \\ &\leq \mathbb{E} [f_\sigma(x^k)] - \mathbb{E} [f_\sigma(x^{k+1})] + \frac{\mathcal{H}_{(2)} \tau_k^2 (\mathcal{A} + \mathcal{B})}{2} + \frac{\mathcal{C}}{m+2} \mathcal{H}_{(m+2)} \tau_k^{m+2}, \end{aligned}$$

where (a) follows from (4.20). This proves (4.19). Summing both sides of the above display from  $k = 0$  to  $T$ , we obtain further that

$$\sum_{k=0}^T \tau_k \tilde{w}_k^2 \leq f_\sigma(x^0) - \mathbb{E} [f_\sigma(x^{T+1})] + 0.5 \mathcal{H}_{(2)} \cdot (\mathcal{A} + \mathcal{B}) \sum_{k=0}^T \tau_k^2 + \frac{\mathcal{H}_{(m+2)} \cdot \mathcal{C}}{m+2} \sum_{k=0}^T \tau_k^{m+2}.$$

Finally, we have  $f_\sigma(x^0) \leq \mathcal{M}(x^0) \cdot \sigma + f(x^0)$  from Lemma 4.2. The desired result now follows immediately upon combining this last observation with the above display.  $\square$

*Remark 4.9* (comparing existing complexity results for (4.1) with  $\Omega = \mathbb{R}^d$ ). When  $m = 0$ ,  $R_1 = 0$ . We then see from (3.7) that  $\mathcal{A} = R_2 \sqrt{d}/\sigma$ ,  $\mathcal{B} = \mathcal{C} = 0$ . Moreover, we have  $\mathcal{M}(x^0) = R_2 \sqrt{d}$  (see Lemma 4.2),  $\mathcal{H}_{(m+2)} = \mathcal{H}_{(2)} = 3 R_2^2 (4 + d)^2$  (see Lemma 4.3). If we let  $\tau_k \equiv \tau$  for some  $\tau \in (0, 1]$ , we see from Theorem 4.8 that

$$(4.22) \quad \min_{0 \leq k \leq T} \tilde{w}_k^2 \leq \frac{1}{\tau(T+1)} \left[ \Delta + R_2 \sqrt{d} \sigma + \frac{3R_2^3}{2\sigma} (4 + d)^2 \sqrt{d} (T+1) \tau^2 \right],$$

where  $\Delta := f(x^0) - \inf f$ . Let  $\delta_a = R_2 \sqrt{d} \sigma$ .<sup>9</sup> Then (4.22) gives

$$(4.23) \quad \min_{0 \leq k \leq T} \mathbb{E} [\|\nabla f_\sigma(x^k)\|^2] \leq 4 \left[ \frac{1}{\tau(T+1)} (\Delta + \delta_a) + \frac{3R_2^4}{2\delta_a} (4 + d)^2 d \tau \right],$$

which matches the bound in [30, section 7] (up to a constant scaling factor).

<sup>9</sup>Note that in view of Lemma 4.2 and [30, Theorem 1], our definition of  $\delta_a$  corresponds to the  $\epsilon$  defined three lines below [30, eqn. (69)].

COROLLARY 4.10. Consider problem (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m \geq 1$  as in (3.1), and assume that  $\Omega = \mathbb{R}^d$ . Let  $\nabla f_\sigma$  be given as in (3.3) and let  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  be given as in (3.7). Assume that  $\mathcal{S} := \arg \min_{u \in \mathbb{R}^d} f(u)$  is nonempty and bounded, and there exists  $\mu > 0$  such that

$$(4.24) \quad f(x) - \inf_{u \in \mathbb{R}^d} f(u) \geq \frac{\mu}{2} \text{dist}(x, \mathcal{S})^2 \quad \forall x \in \mathbb{R}^d.$$

Let  $\gamma \in (0, 1]$  and  $T$  be a positive integer. If  $\tau_k = \gamma/\sqrt{T+1}$  for  $k = 0, \dots, T$ , then the sequence  $\{x^k\}$  generated by Algorithm 4.2 satisfies  $\mathbb{E}[\|x^{k_*}\|^2] \leq \widetilde{M}_\Omega$  for any  $k_* \in \arg \min_{0 \leq k \leq T} \widetilde{w}_k^2$  and

$$(4.25) \quad \min_{0 \leq k \leq T} \mathbb{E}[\|\nabla f_\sigma(x^k)\|^{\frac{1}{2\lceil m/2 \rceil}}] \leq 2^{\frac{1}{2\lceil m/2 \rceil}} \left(1 + \sqrt{\widetilde{M}_\Omega}\right) \widetilde{C}_\Omega^{\frac{1}{4\lceil m/2 \rceil}} T^{-\frac{1}{8\lceil m/2 \rceil}},$$

where

$$(4.26) \quad \widetilde{C}_\Omega = \frac{1}{\gamma} \left[ f(x^0) - \inf_{u \in \mathbb{R}^d} f(u) + \mathcal{M}(x^0)\sigma + \frac{\mathcal{H}_{(2)}(\mathcal{A} + \mathcal{B})}{2} \gamma^2 + \frac{\mathcal{H}_{(m+2)}\mathcal{C}}{m+2} \gamma^{m+2} \right],$$

$$(4.27) \quad \widetilde{M}_\Omega = 8\mu^{-1} [\gamma \widetilde{C}_\Omega + 0.5\mu\sigma^2 d] + 2 \sup_{w \in \mathcal{S}} \|w\|^2,$$

$\widetilde{w}_k^2$  is defined as in Theorem 4.8,  $\mathcal{M}(\cdot)$  is defined as in Lemma 4.2, and  $\mathcal{H}_{(\cdot)}$  is defined as in Lemma 4.3.

Remark 4.11. The condition (4.24) is known as the second-order growth condition for  $f$ , and is a commonly used condition for deriving (global) asymptotic convergence rates of first-order methods; see, e.g., [14]. It is known to hold if  $f$  is strongly convex, and we refer the reader to [14] and references therein for more concrete examples.

Proof of Corollary 4.10. We can deduce from Theorem 4.8 and  $\tau_k \equiv \gamma/\sqrt{T+1}$  that  $\widetilde{w}_{k_*}^2 \leq \widetilde{C}_\Omega/\sqrt{T}$ , where  $\widetilde{C}_\Omega$  is defined as in (4.26), which implies

$$(4.28) \quad \mathbb{E} \left[ \left\| \frac{\nabla f_\sigma(x^{k_*})}{1 + \|x^{k_*}\|^m} \right\|^2 \right] \leq \frac{\widetilde{C}_\Omega}{\sqrt{T}}, \quad \text{and hence} \quad \mathbb{E} \left[ \frac{\|\nabla f_\sigma(x^{k_*})\|}{1 + \|x^{k_*}\|^m} \right] \leq \widetilde{C}_\Omega^{\frac{1}{2}} T^{-\frac{1}{4}}.$$

Now, in view of (4.28) and Lemma 4.4, the conclusion follows once we show

$$(4.29) \quad \mathbb{E}[\|x^{k_*}\|^2] \leq \widetilde{M}_\Omega,$$

where  $\widetilde{M}_\Omega$  is defined as in (4.27). Thus, in what follows, we will prove (4.29).

When  $k_* \geq 1$ , we can sum both sides of (4.19) from  $k = 0$  to  $k_* - 1$  to obtain

$$(4.30) \quad \begin{aligned} \mathbb{E}[f_\sigma(x^{k_*})] &\leq \mathbb{E}[f_\sigma(x^0)] + 0.5\mathcal{H}_{(2)}(\mathcal{A} + \mathcal{B}) \sum_{k=0}^{k_*-1} \tau_k^2 + \frac{\mathcal{H}_{(m+2)}\mathcal{C}}{m+2} \sum_{k=0}^{k_*-1} \tau_k^{m+2} \\ &\leq f(x^0) + \mathcal{M}(x^0)\sigma + 0.5\mathcal{H}_{(2)}(\mathcal{A} + \mathcal{B})\gamma^2 + \frac{\mathcal{H}_{(m+2)}\mathcal{C}}{m+2} \gamma^{m+2}, \end{aligned}$$

where the last inequality follows from Lemma 4.2 and the fact that  $\tau_k = \gamma/\sqrt{T+1}$ . Notice that (4.30) also holds when  $k_* = 0$ .

On the other hand, according to (4.24), we can obtain

$$\begin{aligned} f_\sigma(x^{k_*}) - \inf f &= \mathbb{E}_{u \sim \mathcal{N}(0, I)} [f(x^{k_*} + \sigma u) - \inf f] \geq \frac{\mu}{2} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [\text{dist}(x^{k_*} + \sigma u, \mathcal{S})^2] \\ &= \frac{\mu}{4} \mathbb{E}_{u \sim \mathcal{N}(0, I)} [2\text{dist}(x^{k_*} + \sigma u, \mathcal{S})^2 + 2\|\sigma u\|^2 - 2\|\sigma u\|^2] \geq \frac{\mu}{4} \text{dist}(x^{k_*}, \mathcal{S})^2 - \frac{1}{2} \mu d \sigma^2, \end{aligned}$$

where we used [30, Lemma 1] in the last inequality. This implies that

$$(4.31) \quad \mathbb{E}[f_\sigma(x^{k*})] \geq \frac{\mu}{4} \mathbb{E}[\text{dist}(x^{k*}, \mathcal{S})^2] - \frac{1}{2} \mu d \sigma^2 + \inf f.$$

Combining (4.30) and (4.31), one has

$$\frac{\mu}{4} \mathbb{E}[\text{dist}(x^{k*}, \mathcal{S})^2] \leq f(x^0) + \mathcal{M}(x^0) \sigma + \frac{\mu d \sigma^2}{2} - \inf f + \frac{\mathcal{H}_{(2)}(\mathcal{A} + \mathcal{B})}{2} \gamma^2 + \frac{\mathcal{H}_{(m+2)} \mathcal{C}}{m+2} \gamma^{m+2}.$$

Thus, we can deduce from the above display and the definition of  $\widetilde{M}_\Omega$  in (4.27) that

$$\mathbb{E}[\|x^{k*}\|^2] \leq 2\mathbb{E}[\text{dist}(x^{k*}, \mathcal{S})^2] + 2 \sup_{w \in \mathcal{S}} \|w\|^2 \leq \widetilde{M}_\Omega. \quad \square$$

**5. Explicit complexity and  $(\delta, \epsilon)$ -stationarity.** Note that the results Theorem 4.8 and Corollary 4.10 for the unconstrained nonconvex setting are both with respect to the GS  $f_\sigma$  but not the objective function  $f$ . In this section, we study the iteration complexity of Algorithm 4.2 for achieving a  $(\delta, \epsilon)$ -stationary point of an SPB function  $f$ .

We start by defining analogues of  $\mathcal{H}_{(p)}$  in Lemma 4.3 and  $\mathcal{M}(\cdot)$  in Lemma 4.2:

$$(5.1) \quad \check{\mathcal{H}}_{(p)} = \begin{cases} 3^{p-1} (\mathbf{R}_2^p (2p+d)^p + 2^{(m-1)p} \mathbf{R}_1^p [(m+2)p+d]^{\frac{(m+2)p}{2}}) & \text{if } p \geq 1, \\ 1 & \text{if } p = 0, \end{cases}$$

$$(5.2) \quad \check{\mathcal{M}}(x) = (2^{m-1} \mathbf{R}_1 \|x\|^m + \mathbf{R}_2) \sqrt{d} + 2^{m-1} \mathbf{R}_1 (m+1+d)^{\frac{m+1}{2}}.$$

Notice that if  $\sigma$  is bounded by 1, then it holds that

$$(5.3) \quad \mathcal{H}_{(p)} \leq \check{\mathcal{H}}_{(p)} \quad \forall p \in \mathbb{N} \cup \{0\} \quad \text{and} \quad \mathcal{M}(x) \leq \check{\mathcal{M}}(x) \quad \forall x \in \mathbb{R}^d.$$

In the next auxiliary lemma, we derive bounds that *explicitly* depend on  $\sigma$  for the  $\widetilde{C}_\Omega$  in (4.26) that appeared in Corollary 4.10.

**LEMMA 5.1.** *Consider problem (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $\mathbf{R}_1$ ,  $\mathbf{R}_2$ , and  $m$  as in (3.1). Suppose that  $\inf_{u \in \mathbb{R}^d} f(u) > -\infty$  and let  $\sigma \in (0, 1]$  and  $\gamma = \sigma$ . Then  $\widetilde{C}_\Omega \leq K\sigma^{-1}$ , where*

$$K = f(x^0) - \inf f + \check{\mathcal{M}}(x^0) + 0.5 \check{\mathcal{H}}_{(2)} (2^{2m-2} \mathbf{R}_1 (m+1+d)^{\frac{m+1}{2}} + \mathbf{R}_2 \sqrt{d}) \\ + 2^{2m-3} \check{\mathcal{H}}_{(2)} \mathbf{R}_1 \sqrt{d} + 2^{m-1} \mathbf{R}_1 \check{\mathcal{H}}_{(m+2)} (m+2)^{-1} \sqrt{d},$$

$\check{\mathcal{M}}(\cdot)$  is defined as in (5.2),  $\check{\mathcal{H}}_{(\cdot)}$  is defined as in (5.1), and  $\widetilde{C}_\Omega$  is defined as in (4.26).

*Proof.* First, from the definitions of  $\mathcal{A}$  and  $\mathcal{B}$ , we have that

$$(\mathcal{A} + \mathcal{B}) \gamma^2 = \left[ 2^{2m-2} \mathbf{R}_1 \sigma^{m-1} (m+1+d)^{\frac{m+1}{2}} + \sigma^{-1} \mathbf{R}_2 \sqrt{d} + 2^{2m-2} \sigma^{-1} \mathbf{R}_1 \sqrt{d} \right] \sigma^2 \\ \leq 2^{2m-2} \mathbf{R}_1 (m+1+d)^{\frac{m+1}{2}} + \mathbf{R}_2 \sqrt{d} + 2^{2m-2} \mathbf{R}_1 \sqrt{d}.$$

Next, one can deduce upon invoking the definition of  $\mathcal{C}$  that

$$\mathcal{C} \gamma^{m+2} = 2^{m-1} \sigma^{-1} \mathbf{R}_1 \sqrt{d} \sigma^{m+2} \leq 2^{m-1} \mathbf{R}_1 \sqrt{d},$$

which yields the desired conclusion upon invoking the definition of  $\widetilde{C}_\Omega$  and (5.3).  $\square$

THEOREM 5.2 (complexity bound for approximate Goldstein stationarity). *Consider problem (4.1), where  $f \in \mathcal{SPB}(\mathbb{R}^d)$  with parameters  $R_1$ ,  $R_2$ , and  $m$  as in (3.1). Let  $\inf_{u \in \mathbb{R}^d} f(u) > -\infty$ , let  $\delta \in (0, 1)$ , let  $\mathcal{P}$  be defined as in (3.16), and define*

$$(5.4) \quad \check{\mathcal{N}}(m) = \begin{cases} \max\{\min\{5R_2, 1\}^{-\frac{1}{\min\{m, d\}}}, \kappa_1\}^{4\min\{m, d\}+2} + 1 & \text{if } m \geq 1, \\ \max\{\min\{5R_2, 1\}^{-\frac{1}{d}}, \kappa_2\}^{4d+2} + 1 & \text{if } m = 0, \end{cases}$$

where

$$(5.5) \quad \kappa_1 = \min \left\{ [2^{m+1}R_1(m+d)^{\frac{m}{2}}]^{-\frac{1}{m}}, \frac{\delta \mathcal{P}^{-1/d}}{\sqrt{d\pi e}} \right\}, \quad \kappa_2 = \frac{\delta \mathcal{P}^{-1/d}}{\sqrt{d\pi e}}.$$

Let  $T \geq \check{\mathcal{N}}(m)$  be a positive integer and  $\gamma = \sigma = \check{\sigma}(m)$ , where

$$\check{\sigma}(m) := \begin{cases} \kappa_1 T^{-\frac{1}{4\min\{m, d\}+2}} & \text{if } m \geq 1, \\ \kappa_2 T^{-\frac{1}{4d+2}} & \text{if } m = 0. \end{cases}$$

Let  $\tau_k = \gamma/\sqrt{T+1}$  for  $k = 0, \dots, T$ . Then the following statements hold:

- (i) For  $m \geq 1$ , under the conditions of Corollary 4.10, the sequence  $\{x^k\}$  generated by Algorithm 4.2 satisfies that

$$(5.6) \quad \min_{0 \leq k \leq T} \mathbb{E} [\text{dist}(0, \partial_G^\delta f(x^k))^{\frac{1}{2\lceil m/2 \rceil}}] \leq \frac{(1 + \sqrt{\tilde{K}_\Omega})(2K^{\frac{1}{2}}\kappa_1^{-\frac{1}{2}} + 2)^{\frac{1}{2\lceil m/2 \rceil}}}{T^{(\frac{1}{4} - \frac{1}{8\min\{m, d\}+4})\frac{1}{2\lceil m/2 \rceil}}},$$

where  $K$  is defined as in Lemma 5.1 and  $\tilde{K}_\Omega = 8\mu^{-1}K + 4d + 2\sup_{w \in \mathcal{S}} \|w\|^2$ .

- (ii) For  $m = 0$ , the sequence  $\{x^k\}$  generated by Algorithm 4.2 satisfies that

$$\min_{0 \leq k \leq T} \mathbb{E} [\text{dist}(0, \partial_G^\delta f(x^k))] \leq (2K^{\frac{1}{2}}\kappa_2^{-\frac{1}{2}} + 2)T^{-(\frac{1}{4} - \frac{1}{8d+4})},$$

where  $K$  is defined as in Lemma 5.1.

*Proof.* Since  $T \geq \check{\mathcal{N}}(m)$ , we see that  $\check{\sigma}(m) \leq 1$ , which means  $\gamma \in (0, 1]$ . We can now deduce from Theorem 4.8 and  $\tau_k \equiv \gamma/\sqrt{T+1}$  that  $\tilde{w}_{k_*}^2 \leq \tilde{C}_\Omega/\sqrt{T}$ , where  $\tilde{C}_\Omega$  is defined as in (4.26), which implies

$$(5.7) \quad \mathbb{E} \left[ \left\| \frac{\nabla f_\sigma(x^{k_*})}{1 + \|x^{k_*}\|^m} \right\|^2 \right] \leq \frac{\tilde{C}_\Omega}{\sqrt{T}}, \quad \text{and hence} \quad \mathbb{E} \left[ \frac{\|\nabla f_\sigma(x^{k_*})\|}{1 + \|x^{k_*}\|^m} \right] \leq \tilde{C}_\Omega^{\frac{1}{2}} T^{-\frac{1}{4}}.$$

On the other hand, if we define

$$\check{\epsilon}(m) = \begin{cases} T^{-\frac{\min\{m, d\}}{4\min\{m, d\}+2}} & \text{if } m \geq 1, \\ T^{-\frac{d}{4d+2}} & \text{if } m = 0, \end{cases}$$

then for the  $T \geq \check{\mathcal{N}}(m)$ ,  $\delta \in (0, 1)$ ,  $\epsilon = \check{\epsilon}(m)$ , and  $\sigma = \check{\sigma}(m)$ , one has  $0 < \epsilon < \min\{5R_2, 1\}$  and  $\sigma$  satisfies (3.32). Consequently, by Remark 3.7, one has

$$\nabla f_\sigma(x^{k_*}) \in \partial_G^\delta f(x^{k_*}) + (1 + \|x^{k_*}\|^m)\epsilon \cdot \mathbb{B}.$$

This implies that

$$\text{dist}(0, \partial_G^\delta f(x^{k_*})) \leq \|\nabla f_\sigma(x^{k_*})\| + (1 + \|x^{k_*}\|^m)\epsilon.$$

Rearranging the above display and then taking expectation, one can obtain

$$\mathbb{E} \left[ \frac{\text{dist}(0, \partial_G^\delta f(x^{k_*}))}{1 + \|x^{k_*}\|^m} \right] \leq \mathbb{E} \left[ \frac{\|\nabla f_\sigma(x^{k_*})\|}{1 + \|x^{k_*}\|^m} \right] + \epsilon.$$

This together with (5.7) implies that

$$(5.8) \quad \mathbb{E} \left[ \frac{\text{dist}(0, \partial_G^\delta f(x^{k_*}))}{1 + \|x^{k_*}\|^m} \right] \leq \tilde{C}_\Omega^{\frac{1}{2}} T^{-\frac{1}{4}} + \epsilon.$$

We now prove (i). Notice from Corollary 4.10 that

$$(5.9) \quad \mathbb{E}[\|x^{k_*}\|^2] \leq \tilde{M}_\Omega,$$

where  $\tilde{M}_\Omega$  is defined as in (4.27). Combining (5.9), (5.8), and Lemma 4.4, one has

$$(5.10) \quad \mathbb{E}[\text{dist}(0, \partial_G^\delta f(x^{k_*}))^{\frac{1}{2\lceil m/2 \rceil}}] \leq \left(1 + \sqrt{\tilde{M}_\Omega}\right) (2\tilde{C}_\Omega^{\frac{1}{2}} T^{-\frac{1}{4}} + 2\epsilon)^{\frac{1}{2\lceil m/2 \rceil}}.$$

Next, we utilize Lemma 5.1 and the above display to obtain the desired result for  $m \geq 1$ . To this end, notice that  $\sigma = \check{\sigma}(m) \in (0, 1]$  and  $\gamma = \sigma$ . Combining Lemma 5.1 and the definitions of  $\tilde{M}_\Omega$  and  $\tilde{K}_\Omega$ , one has

$$\tilde{M}_\Omega \leq 8\mu^{-1}[\gamma\tilde{C}_\Omega + 0.5\mu d] + 2 \sup_{w \in \mathcal{S}} \|w\|^2 \leq 8\mu^{-1}(K + 0.5\mu d) + 2 \sup_{w \in \mathcal{S}} \|w\|^2 = \tilde{K}_\Omega.$$

Thus, for  $m \geq 1$ , Lemma 5.1 and (5.10) yield that

$$\begin{aligned} \mathbb{E}[\text{dist}(0, \partial_G^\delta f(x^{k_*}))^{\frac{1}{2\lceil m/2 \rceil}}] &\leq \left(1 + \sqrt{\tilde{K}_\Omega}\right) (2K^{\frac{1}{2}} \sigma^{-\frac{1}{2}} T^{-\frac{1}{4}} + 2\epsilon)^{\frac{1}{2\lceil m/2 \rceil}} \\ &= \left(1 + \sqrt{\tilde{K}_\Omega}\right) \left[2K^{\frac{1}{2}} (\kappa_1 T^{-\frac{1}{4\min\{m,d\}+2}})^{-\frac{1}{2}} T^{-\frac{1}{4}} + 2T^{-\frac{\min\{m,d\}}{4\min\{m,d\}+2}}\right]^{\frac{1}{2\lceil m/2 \rceil}} \\ &= \left(1 + \sqrt{\tilde{K}_\Omega}\right) (2K^{\frac{1}{2}} \kappa_1^{-\frac{1}{2}} + 2)^{\frac{1}{2\lceil m/2 \rceil}} T^{-(\frac{1}{4} - \frac{1}{8\min\{m,d\}+4})^{\frac{1}{2\lceil m/2 \rceil}}}. \end{aligned}$$

Finally, to prove (ii), we deduce from Lemma 5.1 and (5.8) that

$$\begin{aligned} \mathbb{E}[\text{dist}(0, \partial_G^\delta f(x^{k_*}))] &\leq 2K^{\frac{1}{2}} \sigma^{-\frac{1}{2}} T^{-\frac{1}{4}} + 2\epsilon \\ &= 2K^{\frac{1}{2}} (\kappa_2 T^{-\frac{1}{4d+2}})^{-\frac{1}{2}} T^{-\frac{1}{4}} + 2T^{-\frac{d}{4d+2}} = (2K^{\frac{1}{2}} \kappa_2^{-\frac{1}{2}} + 2) T^{-(\frac{1}{4} - \frac{1}{8d+4})}. \quad \square \end{aligned}$$

*Remark 5.3* (explicit bound on  $\check{\mathcal{N}}(m)$ ). Note that the result in Theorem 5.2 requires that  $T \geq \check{\mathcal{N}}(m)$ . Here, we derive simpler bounds for  $\check{\mathcal{N}}(m)$  that are *independent* of  $d$  when  $d$  is large, under the assumptions of Theorem 5.2. We first consider the case  $m \geq 1$ . Assume in addition that  $d \geq m$ . Then

$$\kappa_1 \stackrel{(a)}{\leq} (2^{m+1} R_1)^{-\frac{1}{m}} (m+d)^{-\frac{1}{2}} \leq R_1^{-\frac{1}{m}},$$

where we used the definition of  $\kappa_1$  (see (5.5)) in (a). From the above display and the definition of  $\check{\mathcal{N}}(m)$ , we obtain

$$\check{\mathcal{N}}(m) \leq \max\{\min\{5R_2, 1\}^{-1}, R_1^{-1}\}^{\frac{4m+2}{m}} + 1.$$

Next, in the case  $m = 0$ , let  $d \geq \max\{2, 2^{-2}R_2^{-1}\}$ . Since  $d \geq 2$ , we have

$$(5.11) \quad \min\{5R_2, 1\}^{-\frac{4d+2}{d}} \leq \min\{5R_2, 1\}^{-5}.$$



On the other hand, we know

$$(5.12) \quad \kappa_2^{4d+2} \stackrel{(a)}{=} \delta^{4d+2} \frac{(4R_2)^{-\frac{4d+2}{d}}}{(\sqrt{d\pi e})^{4d+2}} \stackrel{(b)}{\leq} \frac{1}{(4R_2)^{\frac{4d+2}{d}} d^{2d+1}} \stackrel{(c)}{\leq} \frac{1}{(4R_2 d)^{\frac{4d+2}{d}}} \stackrel{(d)}{\leq} 1,$$

where (a) follows from the definition of  $\kappa_2$  in (5.5), (b) holds because  $\sqrt{\pi e} > 1$  and  $\delta \in (0, 1)$ , (c) follows from  $d \geq 2$ , and (d) follows from the fact  $d \geq 2^{-2}R_2^{-1}$ . Now, combining (5.11), (5.12), and the definition of  $\check{\mathcal{N}}(m)$ , one has

$$\check{\mathcal{N}}(m) \leq \min\{5R_2, 1\}^{-5} + 1.$$

**Acknowledgments.** The authors would like to thank Wenqing Ouyang for discussions on the tail bound and tempered distribution used in section 3. We would also like to thank the referees for their comments and suggestions, which significantly improved the manuscript.

#### REFERENCES

- [1] K. BALASUBRAMANIAN AND S. GHADIMI, *Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points*, Found. Comput. Math., 22 (2022), pp. 35–76, <https://doi.org/10.1007/s10208-021-09499-8>.
- [2] A. BERAHAS, L. CAO, K. CHOROMANSKI, AND K. SCHEINBERG, *A theoretical and empirical comparison of gradient approximations in derivative-free optimization*, Found. Comput. Math., 22 (2022), pp. 507–560, <https://doi.org/10.1007/s10208-021-09513-z>.
- [3] J. BOLTE, T. LE, AND E. PAUWELS, *Subgradient sampling for nonsmooth nonconvex minimization*, SIAM J. Optim., 33 (2023), pp. 2542–2569, <https://doi.org/10.1137/22M1479178>.
- [4] J. BOLTE AND E. PAUWELS, *Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning*, Math. Program., 188 (2021), pp. 19–51, <https://doi.org/10.1007/s10107-020-01501-5>.
- [5] J. M. BORWEIN AND S. B. LINDSTROM, *Meetings with Lambert W and other special functions in optimization and analysis*, Pure Appl. Funct. Anal., 1 (2016), pp. 361–396.
- [6] H. BREZIS, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, 2011.
- [7] P. H. CALAMAI AND J. J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Program., 39 (1987), pp. 93–116, <https://doi.org/10.1007/BF02592073>.
- [8] P. CHEN, H. ZHANG, Y. SHARMA, J. YI, AND C.-J. HSIEH, *ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models*, in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 15–26.
- [9] Z. CHEN, Y. ZHOU, Y. LIANG, AND Z. LU, *Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization*, in Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, PMLR 202, 2023.
- [10] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
- [11] A. R. CONN, K. SCHEINBERG, AND L. N. VICENTE, *Introduction to Derivative-Free Optimization*, MOS-SIAM Ser. Optim. 8, SIAM, 2009, <https://doi.org/10.1137/1.9780898718768>.
- [12] S. DASGUPTA AND A. GUPTA, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms, 22 (2003), pp. 60–65, <https://doi.org/10.1002/rsa.10073>.
- [13] A. DEREVENTSOV, C. G. WEBSTER, AND J. DAWS, *An adaptive stochastic gradient-free approach for high-dimensional blackbox optimization*, in Proceedings of the International Conference on Computational Intelligence: Algorithms for Intelligent Systems, R. Tiwari, A. Mishra, N. Yadav, and M. Pavone, eds., Springer, Singapore, 2022.
- [14] D. DRUSVYATSKIY AND A. S. LEWIS, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2021), pp. 919–948, <https://doi.org/10.1287/moor.2017.0889>.
- [15] M. J. EHRHARDT AND L. ROBERTS, *Efficient Hyperparameter Tuning with Dynamic Accuracy Derivative-Free Optimization*, preprint, <https://arxiv.org/abs/2011.03151>, 2020.
- [16] F. N. FRITSCH, R. E. SHAFER, AND W. P. CROWLY, *Algorithm 443: Solution of the transcendental equation  $we^w = x$* , Commun. ACM, 16 (1973), pp. 123–124.
- [17] A. A. GOLDSTEIN, *Optimization of Lipschitz continuous functions*, Math. Program., 13 (1977), pp. 14–22, <https://doi.org/10.1007/BF01584320>.

- [18] L. GRAFAKOS, *Classical Fourier Analysis*, Springer, New York, 2008.
- [19] X. GUO, D. KEIVAN, G. DULLERUD, P. SEILER, AND B. HU, *Complexity of derivative-free policy optimization for structured  $\mathcal{H}_\infty$  control*, Adv. Neural Inf. Process. Syst., 36 (2023), pp. 5050–5078.
- [20] J. JIN, B. ZHANG, H. WANG, AND L. WANG, *Non-convex distributionally robust optimization: Nonasymptotic analysis*, Adv. Neural Inf. Process. Syst., 34 (2021), pp. 2771–2782.
- [21] W. JONGENEEL, M.-C. YUE, AND D. KUHN, *Small errors in random zeroth-order optimization are imaginary*, SIAM J. Optim., 34 (2024), pp. 2638–2670, <https://doi.org/10.1137/22M1510261>.
- [22] K. SCHEINBERG, *Finite difference gradient approximation: To randomize or not?*, Inform. J. Comput., 34 (2022), pp. 2384–2388, <https://doi.org/10.1287/ijoc.2022.1218>.
- [23] J. LARSON, M. MENICKELLY, AND S. M. WILD, *Derivative-free optimization methods*, Acta Numer., 28 (2019), pp. 287–404, <https://doi.org/10.1017/S0962492919000060>.
- [24] T. LIN, Z. ZHENG, AND M. I. JORDAN, *Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization*, Adv. Neural Inf. Process. Syst., 35 (2022), pp. 26160–26175.
- [25] L. LÓCZI, *Guaranteed- and high-precision evaluation of the Lambert W function*, Appl. Math. Comput., 433 (2022).
- [26] A. MAGGIAR, A. WÄCHTER, I. S. DOLINSKAYA, AND J. STAUM, *A derivative-free trust region algorithm for the optimization of functions smoothed via Gaussian convolution using adaptive multiple importance sampling*, SIAM J. Optim., 28 (2018), pp. 1478–1507, <https://doi.org/10.1137/15M1031679>.
- [27] A. MISHKIN, A. KHALED, Y. WANG, A. DEFAZIO, AND R. M. GOWER, *Directional smoothness and gradient methods: Convergence and adaptivity*, Adv. Neural Inf. Process. Syst., 37 (2024), pp. 14810–14848.
- [28] Y. NESTEROV, *Introductory Lectures on Convex Optimization*, Springer, 2004.
- [29] A. S. NEMIROVSKY AND D. B. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, Wiley, 1983.
- [30] Y. NESTEROV AND V. SPOKOINY, *Random gradient-free minimization of convex functions*, Found. Comput. Math., 17 (2017), pp. 527–566, <https://doi.org/10.1007/s10208-015-9296-2>.
- [31] S. OSHER, B. WANG, P. YIN, X. LUO, F. BAREKAT, M. PHAM, AND A. LIN, *Laplacian smoothing gradient descent*, Res. Math. Sci., 9 (2022), <https://doi.org/10.1007/s40687-022-00351-1>.
- [32] F. D. PARKER, *Integrals of inverse functions*, Amer. Math. Monthly, 62 (1955), pp. 439–440, <https://doi.org/10.2307/2307006>.
- [33] B. T. POLYAK, *Introduction to Optimization*, Optimization Software, 1983.
- [34] A. M.-C. SO, *A Semidefinite Programming Approach to the Graph Realization Problem: Theory, Applications and Extensions*, Ph.D. thesis, Stanford University, 2007.
- [35] A. STARNES, A. DEREVENTSOV, AND C. WEBSTER, *Gaussian Smoothing Gradient Descent for Minimizing Functions (GSmoothGD)*, preprint, <https://arxiv.org/abs/2311.00521v2>, 2023.
- [36] A. STARNES AND C. WEBSTER, *Gaussian Smoothing Stochastic Gradient Descent (GSmoothSGD)*, preprint, <https://arxiv.org/abs/2311.00531>, 2023.
- [37] E. M. STEIN AND R. SHAKARCHI, *Functional Analysis: Introduction to Further Topics in Analysis*, Princeton University Press, Princeton, NJ, 2011.
- [38] B. TASKESEN, M.-C. YUE, J. BLANCHET, D. KUHN, AND V. A. NGUYEN, *Sequential domain adaptation by synthesizing distributionally robust experts*, in Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021, pp. 10162–10172.
- [39] V. S. VLADIMIROV, *Methods of the Theory of Generalized Functions*, Taylor & Francis, London, 2002.
- [40] J. ZHANG, T. HE, S. SRA, AND A. JADBABAIE, *Why gradient clipping accelerates training: A theoretical justification for adaptivity*, in International Conference on Learning Representations, 2020.
- [41] J. ZHANG, H. LIN, S. JEGELKA, S. SRA, AND A. JADBABAIE, *Complexity of finding stationary points of nonsmooth nonconvex functions*, in Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.