SUBDIFFERENTIALLY POLYNOMIALLY BOUNDED FUNCTIONS AND GAUSSIAN SMOOTHING-BASED ZEROTH-ORDER OPTIMIZATION *

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

Abstract. We introduce 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 non-smooth 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 constrained minimization of SPB functions, and analyze their iteration complexity. An important instrument in our analysis, which could be of independent interest, is the quantification of Goldstein stationarity via the GS gradient.

Key words. Gaussian smoothing, Zeroth-order optimization, Subdifferentially polynomially bounded functions, Goldstein stationarity

MSC codes. 49J52, 90C30, 90C56

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 typically arises in situations where the objective function is given as the output of simulations of complex physical systems, and has attracted intense research over the last few decades. We refer the readers to the expositions [10, 16] and references therein for classic works and recent developments on zeroth-order optimization.

A prominent zeroth-order optimization algorithm is the Nesterov and Spokoiny's random search method [16, 24, 25, 28] developed based on the concept of Gaussian smoothing (GS), whose definition is recalled here for convenience.

DEFINITION 1.1 ([25, section 2]). *Let $\sigma > 0$ and $f : \mathbb{R}^d \to \mathbb{R}$ 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.

---

*Submitted to the editors May 7, 2024.

Funding: Ting Kei Pong is supported in part by a Research Scheme of the Research Grants Council of Hong Kong SAR, China (project T22-504/21R). Shuqin Sun is supported in part by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (project 2023QYJ08). Man-Chung Yue is supported in part by the Hong Kong Research Grants Council under the GRF project 15304422.

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

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

§School of Mathematics Education, China West Normal University, Nanchong, People's Republic of China (sunshuqinsusan@163.com).

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

As a convolution of f with the Gaussian kernel, the GS f_σ enjoys many desirable properties. For example, it was shown in [25, section 2] to inherit convexity and Lipschitz continuity from f. Moreover, it was shown that ∇f_σ is Lipschitz continuous whenever f is globally Lipschitz. This latter fact was leveraged in [25, 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 [25] has stimulated a surge of studies on GS-based zeroth-order optimization algorithms; see, e.g, [1, 2, 11, 15, 17, 21, 26, 31, 32].

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 distributionally robust optimization [14, 33] and the training of large language models [34]. 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 [6, 14, 22, 34] 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 tempted 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. In particular, the GS of a generalized smooth function in the sense of [14] is not well-defined in general.

In this paper, we identify a new and rich class of locally Lipschitz functions whose GS are well-defined and possess useful properties for algorithmic analysis, and develop new GS-based zeroth-order optimization algorithms for minimizing this class of functions. Our main contributions are threefold.

1. We introduce 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 all functions that are Lipschitz continuous, or have Lipschitz continuous gradients or Hessians. More interestingly, the SPB class includes even certain nonsmooth locally Lipschitz functions, such as functions arising from neural networks; see Examples 3.1(v)-(vi).
2. We show that if f is SPB, then its GS f_σ is well-defined and continuously differentiable; moreover, f_σ and its partial derivatives are SPB too. We also establish a relationship between ∇f_σ and the Goldstein δ-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 δ-subdifferential is a commonly used subdifferential for studying stationarity of nonsmooth functions [13, 35], and our result can be viewed as an extension of [25, Theorem 2] and [17, Theorem 3.1] from globally Lipschitz to SPB functions.
3. We propose a GS-based zeroth-order algorithm for minimizing an SPB function f over a closed convex set $\mathcal{D}$. Our algorithm updates 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 a sample average of the random vector $(f(x^k+\sigma u)-f(x^k))u/\sigma$ with $u \sim \mathcal{N}(0, I)$, where the sample size depends on the desired accuracy and a certain local variance parameter. When f is additionally convex or when $\mathcal{D} = \mathbb{R}^d$ (i.e., the minimization problem is unconstrained), we develop variants of our algorithm where the required sample size at each iteration for forming the approximate gradient is reduced to *one*. We analyze the iteration complexity of the proposed algorithms. The crux of our analysis is a novel descent lemma for f_σ analogous to the standard descent lemma for Lipschitz differentiable functions, where the Lipschitz constant is replaced by a polynomial 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 δ-stationarity. In section 4, we prove the descent lemma and develop our GS-based zeroth-order algorithms for constrained minimization of 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 $I = [e_1, e_2, \cdots, e_d]$ denote the $d \times d$ identity matrix, where $e_i \in \mathbb{R}^d$ is the i-th canonical basis vector for $i = 1, \ldots, d$, 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, ∂D and $\operatorname{conv}(D)$ denote its complement, boundary and convex hull, respectively; we also denote the characteristic function of D by

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

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

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

For a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, the Clarke directional derivative of f (see [7, Page 25]) at any $x \in \mathbb{R}^d$ in the direction $v \in \mathbb{R}^d$ is defined as

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

and the Clarke subdifferential of f (see [7, 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 Ω_f be the set of points at which f is not differentiable, we have

$$\partial_C f(x) = \operatorname{conv}\Big(\big\{s \in \mathbb{R}^d : \exists \{x^k\} \subset \mathbb{R}^d \setminus \Omega_f \text{ with } x^k \to x \text{ and } \nabla f(x^k) \to s\big\}\Big); \tag{2.1}$$

see [7, Propositions 2.1.2(b)] and [7, Theorem 2.5.1].

Next, for any $\delta > 0$, the Goldstein δ-subdifferential [13] of f at $x \in \mathbb{R}^d$ is the set

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

Note that at any $x \in \mathbb{R}^d$, both the Clarke subdifferential and Goldstein δ-subdifferential are compact convex sets. Finally, following [35, Definition 4], an $x \in \mathbb{R}^d$ is said to be a (δ, ϵ)-stationary point of f if $\operatorname{dist}(0, \partial_G^\delta f(x)) \le \epsilon$.

3. Subdifferentially polynomially bounded functions.

The purpose of this section is to introduce the class of subdifferentially polynomially bounded (SPB) functions and study their properties.

Definition 3.1 (Subdifferentially polynomially bounded functions). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous. We say that f is subdifferentially polynomially bounded (SPB) if there exist $\mathrm{R}_1 \ge 0$, $\mathrm{R}_2 > 0$ and an integer $m \ge 1$ such that*

$$\sup_{\zeta\in\partial_C f(x)} \|\zeta\| \le \mathrm{R}_1\|x\|^m + \mathrm{R}_2 \qquad \forall x \in \mathbb{R}^d. \tag{3.1}$$

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

Note that using calculus rules for Clarke subdifferential (see Corollary 2 of [7, 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 $\mathrm{R}_1 = 0$ 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 $f : \mathbb{R}^d \to \mathbb{R}$ is globally Lipschitz continuous with Lipschitz continuity modulus $L > 0$, then we have from [7, Proposition 2.1.2(a)] that $\sup_{u\in\partial_C f(x)} \|u\| \le L$ 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, there exists $L > 0$ such that*

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

showing that g is SPB.

(iv) *Let $f = g \circ h$, where $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}^n$. Assume that g and all component functions of h are SPB. Then one can deduce from [7, 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*

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

where for $\ell = 1, \ldots, L$, $\varrho_\ell : \mathbb{R} \to \mathbb{R}$ is the activation function for the ℓ-th layer (for any vector z, the notation $\varrho_\ell(z)$ is understood as the vector obtained by applying the activation function ϱ_ℓ entrywise to z), $W_\ell : \mathbb{R}^{p_\ell} \to \mathbb{R}^{p_{\ell+1}}$ is an affine mapping for some positive integers p_ℓ 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* $W_1, \ldots, W_L$*; see [3, 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) [16, 24, 25, 28]. 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*

$$|f(x) - f(y)| \leq (2^{m-1}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y - x\|^m + \mathrm{R}_2)\|x - y\| \qquad \forall x, y \in \mathbb{R}^d.$$

Proof. From [7, 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

$$\begin{aligned} |f(x) - f(y)| &\leq \sup_{\alpha \in (0,1)} \{\mathrm{R}_1\|x + \alpha(y - x)\|^m + \mathrm{R}_2\}\|x - y\| \\ &\leq (2^{m-1}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y - x\|^m + \mathrm{R}_2)\|x - y\|, \end{aligned}$$

where the second inequality follows from the convexity of the function $\|\cdot\|^m$ (thanks to $m \geq 1$) and $\alpha \in (0, 1)$. ∎

The theorem below asserts that for any SPB function f, the GS f_σ and its gradient ∇f_σ are both well-defined.

THEOREM 3.3 (Well-definedness of GS and its gradient). *Let* $f \in \mathcal{SPB}(\mathbb{R}^d)$*. Then its GS* f_σ*, given in Definition 1.1, is well-defined. Moreover, the gradient of* f_σ *is given by*

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

and is well-defined and continuous.

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

$$\begin{aligned} &\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[|f(x+\sigma u)|\right] \le \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[|f(x+\sigma u)-f(x)|\right] + |f(x)| \\ &\le \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[(2^{m-1}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\sigma^m\|u\|^m + \mathrm{R}_2)\cdot\sigma\|u\|\right] + |f(x)| < \infty, \end{aligned} \tag{3.4}$$

where the second inequality follows from Lemma 3.2 with R_1, R_2 and m as in (3.1). Therefore, f_σ is well-defined.

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

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

Note that $e^{-\frac{\|\cdot\|^2}{2\sigma^2}}$ is a Schwartz function on $\mathbb{R}^d$ according to [8, Example 2.2.2]. On the other hand, we have the following inequality for each $r > 0$ based on Lemma 3.2:

$$\begin{aligned} \int_{\mathbb{B}_r} |f(x)|dx &\le r^d\alpha(d)\cdot(|f(0)| + \max_{x\in\mathbb{B}_r}|f(x)-f(0)|) \\ &\le r^d\alpha(d)\cdot(|f(0)| + [2^{m-1}\mathrm{R}_1 r^m + \mathrm{R}_2]r) =: \mathfrak{U}(r), \end{aligned}$$

where $\alpha(d)$ is the volume of the d-dimensional unit ball $\mathbb{B}$. Since $\mathfrak{U}(r) = O(r^{m+d+1})$ as $r\to\infty$, in view of [9],[1] 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 space (i.e., a tempered distribution).

The next part follows closely from the proof of [8, Theorem 2.3.20], which is included for self-containedness. Specifically, from (3.5), we see that for any $h \in \mathbb{R}\backslash\{0\}$ and any $i \in \{1,\ldots,d\}$,

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

Since $(e^{-\frac{\|x+he_i-y\|^2}{2\sigma^2}} - e^{-\frac{\|x-y\|^2}{2\sigma^2}})/h \to -\frac{x_i-y_i}{\sigma^2}e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ as $h\to 0$ in Schwartz space according to [8, Exercise 2.3.5(a)] and $\mathfrak{F}$ is a tempered distribution, we conclude upon passing to the limit as $h\to 0$ in the above display that

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

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

Finally, the continuity of ∇f_σ 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). ∎

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.

[1]Specifically, see the last example on page 106.

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_σ be defined in Definition 1.1. Then the following statements hold.*

(i) *It holds that*

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

where $\mathfrak{A} = 2^{2m-2}\mathrm{R}_1\sigma^m(m+d)^{\frac{m}{2}} + \mathrm{R}_2$, $\mathfrak{B} = 2^{2m-2}\mathrm{R}_1$ and $\mathfrak{C} = 2^{m-1}\mathrm{R}_1$. In particular, f_σ is SPB.

(ii) *It holds that*

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

where $\mathcal{A} = 2^{2m-2}\mathrm{R}_1\sigma^{m-1}\left(m+1+d\right)^{\frac{m+1}{2}} + \frac{1}{\sigma}\mathrm{R}_2\sqrt{d}$, $\mathcal{B} = \frac{1}{\sigma}2^{2m-2}\mathrm{R}_1\sqrt{d}$ and $\mathcal{C} = \frac{1}{\sigma}2^{m-1}\mathrm{R}_1\sqrt{d}$. In particular, $\frac{\partial f_\sigma}{\partial x_i}$ is SPB for any i.

Proof. We first observe from Lemma 3.2 that for every x, y and $u \in \mathbb{R}^d$,

$$\begin{aligned}
&|f(x+\sigma u) - f(y+\sigma u)| \\
&\le (2^{m-1}\mathrm{R}_1\|x+\sigma u\|^m + 2^{m-1}\mathrm{R}_1\|y-x\|^m + \mathrm{R}_2)\|y-x\| \\
&\le \left(2^{2m-2}\mathrm{R}_1\sigma^m\|u\|^m + 2^{2m-2}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y-x\|^m + \mathrm{R}_2\right)\|x-y\|,
\end{aligned} \tag{3.8}$$

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

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

$$\begin{aligned}
&\frac{|f_\sigma(x) - f_\sigma(y)|}{\|x-y\|} \le \frac{\mathbb{E}_{u\sim\mathcal{N}(0,I)}[|f(x+\sigma u) - f(y+\sigma u)|]}{\|x-y\|} \\
&\le \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(2^{2m-2}\mathrm{R}_1\sigma^m\|u\|^m + 2^{2m-2}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y-x\|^m + \mathrm{R}_2\right)\right] \\
&= (2^{2m-2}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y-x\|^m + \mathrm{R}_2) + 2^{2m-2}\mathrm{R}_1\sigma^m\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|^m] \\
&\overset{\text{(a)}}{\le} (2^{2m-2}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|y-x\|^m + \mathrm{R}_2) + 2^{2m-2}\mathrm{R}_1\sigma^m(m+d)^{\frac{m}{2}},
\end{aligned}$$

where (a) follows from [25, 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

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

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

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

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

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

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

where (a) follows from [25, 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). ∎

3.2. Approximate Goldstein stationarity. In this subsection, we explore the relationship between the GS gradient and the Goldstein δ-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., [4, 12, 27]); 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 $M \geq \left[-d\, W_{-1}\left(-\nu^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$, 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 $M \geq \wp := \left[-d\, W_{-1}\left(-\nu^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$. 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

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

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 $M \geq \wp > \sqrt{d}$.[2] Then, from [29, Lemma 2.2] (see also [30, Proposition 5.3.1]), we have

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

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

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

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 $M \geq \wp$. Thus, the desired conclusion follows from (3.11) and (3.10). ∎

In the next theorem, we show that for all sufficiently small $\sigma > 0$, the Goldstein δ-subgradients can be approximated by the GS gradient ∇f_σ. Specifically, we derive a sufficient condition on σ, in the form of an *explicit* upper bound depending on δ and ϵ, for the GS gradient to reside in an ϵ neighborhood of the Goldstein δ-subdifferential. Similar results have been derived under a globally Lipschitz continuity assumption on f; see, e.g., [25, Theorem 2] and [17, Theorem 3.1]. In particular, the proof of [17, 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 [18, section 4.2.5] and [19, Lemma 9.1]. Here we include an elementary proof to highlight the role of polynomial boundedness of the subdifferential.

[2]The second inequality holds because $W_{-1}(t) < -1$ when $t \in (-1/e, 0)$; see [20, Page 2].

THEOREM 3.6 (GS gradient as approximate Goldstein δ-subgradient). *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1), and $\mathcal{X}$ be a compact set. Let ∇f_σ and $\partial_G^\delta f$ be given in (3.3) and (2.2), respectively. Then the following hold.*

(i) *For every $x \in \mathcal{X}$, it holds that*

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

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

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

$$\nabla f_\sigma(x) \in \partial_G^\delta f(x) + \varepsilon\cdot\mathbb{B} \qquad \forall\, \sigma\in(0,\bar\sigma] \ \text{ and } \ \forall\, x\in\mathcal{X},$$

where

$$\bar\sigma = \min\left\{\left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}, 1, \frac{\delta}{H}\right\}, \quad C_1 = \max_{w\in\mathcal{X}}\|w\|, \tag{3.13}$$

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

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

$$\mathcal{P} = 4\left(2^{m-1}\mathrm{R}_1 C_1^m + \mathrm{R}_2\right) + 2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}, \tag{3.16}$$

and W_{-1} is the negative real branch of the Lambert W function.[3]

Proof. Fix any $x \in \mathcal{X}$. From Theorem 3.3, the GS f_σ and it gradient ∇f_σ are well-defined at x. For notational simplicity, for any $\sigma > 0$, let $\mathfrak{D}_\sigma = \{u \in \mathbb{R}^d : f$ is differentiable at $x + \sigma u\}$. Then it follows from the 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 $h \in \mathbb{R}^d$

$$\begin{aligned} |f(x+h+\sigma u) - f(x+\sigma u)| &\le [2^{m-1}\mathrm{R}_1\|x+\sigma u\|^m + 2^{m-1}\mathrm{R}_1\|h\|^m + \mathrm{R}_2]\|h\| \\ &\le (2^{2m-2}\mathrm{R}_1\|x\|^m + 2^{2m-2}\mathrm{R}_1\sigma^m\|u\|^m + 2^{m-1}\mathrm{R}_1\|h\|^m + \mathrm{R}_2)\|h\|, \end{aligned} \tag{3.17}$$

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

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

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

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

$$\begin{aligned} 0 &= \lim_{t\to0}\frac{f_\sigma(x+th_x) - f_\sigma(x) - \langle\nabla f_\sigma(x), th_x\rangle}{\|th_x\|} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\lim_{t\to0}\int_{\mathbb{R}^d}\frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle\nabla f_\sigma(x), th_x\rangle}{\|th_x\|}\cdot e^{-\frac{\|u\|^2}{2}}du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\lim_{t\to0}\int_{\mathfrak{D}_\sigma}\frac{f(x+th_x+\sigma u) - f(x+\sigma u) - \langle\nabla f_\sigma(x), th_x\rangle}{\|th_x\|}\cdot e^{-\frac{\|u\|^2}{2}}du \\ &\overset{(a)}{=} \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\langle\nabla f(x+\sigma u) - \nabla f_\sigma(x), h_x\rangle}{\|h_x\|}\cdot\mathbb{1}_{\mathfrak{D}_\sigma}(u)\right] \overset{(b)}{=} \|h_x\|, \end{aligned}$$

[3]Since $\eta_1 \le (2\pi)^{\frac{d}{2}} - \frac{1}{2}$, we have $\eta_1^{2/d}/(2\pi e) < 1/e$ and hence $W_{-1}(-\eta_1^{2/d}/(2\pi e)) < -1$. Thus, $H \in (\sqrt{d}, \infty)$.

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

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

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

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

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

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

$$\begin{aligned}&\left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\nabla f(x+\sigma u)\cdot\mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M^c}(u)]\right\|\\ &\overset{\text{(a)}}{\le}\mathbb{E}_{u\sim\mathcal{N}(0,I)}[(\mathrm{R}_1\|x+\sigma u\|^m+\mathrm{R}_2)\cdot\mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M^c}(u)]\overset{\text{(b)}}{\le}\Xi_1+\Xi_2\end{aligned}\tag{3.19}$$

with

$$\begin{aligned}\Xi_1&=(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2)\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)],\\ \Xi_2&=(2^{m-1}\mathrm{R}_1\sigma^m)\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|^m\cdot\mathbb{1}_{\mathbb{B}_M^c}(u)],\end{aligned}$$

where we invoked (3.1) in (a), and used the convexity of $\|\cdot\|^m$ and the definition of C_1 in (3.13) for (b). Now, observe that

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

where the second inequality follows from [25, Lemma 1]. Thus, if we choose a finite positive σ such that $\sigma\le\left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}$, then

$$\Xi_2=(2^{m-1}\mathrm{R}_1\sigma^m)\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|^m\cdot\mathbb{1}_{\mathbb{B}_M^c}(u)]\le\frac{\varepsilon}{4}.\tag{3.20}$$

Next, choose $\eta=\varepsilon(2\pi)^{\frac{d}{2}}\left[4\left(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2\right)\right]^{-1}$. Then, by setting $\nu=\eta$ in Lemma 3.5, we see that whenever $M\ge\left[-d\,W_{-1}\left(-\eta^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$,

$$\Xi_1=(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2)\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M^c}(u)]\le\frac{\varepsilon}{4}.\tag{3.21}$$

Additionally, for the above η and the η_1 in (3.15), we have

$$\eta_1\le\varepsilon\left[4\left(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2\right)+2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}\right]^{-1}\le\eta.$$

Combining this conclusion with (3.19), (3.20) and (3.21), we know that for any finite positive $\sigma \leq \left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}$ and $M \geq \left[-d\,W_{-1}\left(-\eta_1^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$ with η_1 as in (3.15), we have

$$\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\nabla f(x+\sigma u)\cdot \mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M^c}(u)]\| \leq \frac{\varepsilon}{2}. \tag{3.22}$$

We next estimate the second term on the second line of (3.18). We first notice that $\int_{\mathbb{B}_M} e^{-\frac{\|u\|^2}{2}}du = (2\pi)^{\frac{d}{2}} - \int_{\mathbb{B}_M^c} e^{-\frac{\|u\|^2}{2}}du$. So, if $\int_{\mathbb{B}_M^c} e^{-\frac{\|u\|^2}{2}}du \leq (2\pi)^{\frac{d}{2}} - \frac{1}{2}$, then $\int_{\mathbb{B}_M} e^{-\frac{\|u\|^2}{2}}du \geq \frac{1}{2}$. In view of Lemma 3.5 and the definition of η_1, this happens when we choose $M \geq \left[-d\,W_{-1}\left(-\eta_1^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$, since $\eta_1 \leq (2\pi)^{\frac{d}{2}} - \frac{1}{2}$.

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:

$$\begin{aligned}
&\left\|\left(1-\frac{1}{\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\mathbb{1}_{\mathbb{B}_M}(u)]}\right)\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\nabla f(x+\sigma u)\cdot\mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M}(u)\right]\right\| \\
&\leq \frac{\widetilde{K}}{\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\mathbb{1}_{\mathbb{B}_M}(u)\right]}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\nabla f(x+\sigma u)\|\cdot\mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M}(u)\right] \\
&\overset{\text{(a)}}{\leq} 2(2\pi)^{\frac{d}{2}}\widetilde{K}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\nabla f(x+\sigma u)\|\cdot\mathbb{1}_{\mathfrak{D}_\sigma\cap\mathbb{B}_M}(u)\right] \\
&\overset{\text{(b)}}{\leq} 2(2\pi)^{\frac{d}{2}}\widetilde{K}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\mathrm{R}_1\|x+\sigma u\|^m+\mathrm{R}_2\right] \\
&\overset{\text{(c)}}{\leq} 2(2\pi)^{\frac{d}{2}}\widetilde{K}\left[(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2)+2^{m-1}\mathrm{R}_1\sigma^m\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|^m]\right] \\
&\overset{\text{(d)}}{\leq} 2(2\pi)^{\frac{d}{2}}\widetilde{K}\left[\left(2^{m-1}\mathrm{R}_1C_1^m+\mathrm{R}_2\right)+2^{m-1}\mathrm{R}_1(m+d)^{\frac{m}{2}}\right],
\end{aligned} \tag{3.23}$$

where $\widetilde{K} := \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\mathbb{1}_{\mathbb{B}_M^c}(u)\right]$, and (a) holds because $(2\pi)^{\frac{d}{2}}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\mathbb{1}_{\mathbb{B}_M}(u)\right] = \int_{\mathbb{B}_M} e^{-\frac{\|u\|^2}{2}}du \geq \frac{1}{2}$, (b) holds upon using (3.1) and enlarging the domain of integration, (c) follows from the convexity of $\|\cdot\|^m$ and the definition of C_1 in (3.13), (d) follows from [25, 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 $M \geq \left[-d\,W_{-1}\left(-\eta_1^{\frac{2}{d}}/(2\pi e)\right)\right]^{\frac{1}{2}}$ with η_1 as in (3.15) and $\sigma \leq 1$, then

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

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 $\sigma \leq \widetilde{\sigma} = \min\left\{\left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}, 1\right\}$. Hence, we have

$$\|\Delta_h\| \leq \epsilon \quad \text{whenever} \quad M \geq H \text{ and } \sigma \leq \widetilde{\sigma}, \tag{3.25}$$

where Δ_h is given in (3.18).

Finally, for $M = H$ and any $\sigma \leq \bar{\sigma} = \min\left\{\left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}, 1, \frac{\delta}{H}\right\}$, we have $\sigma M \leq \delta$. Hence, by (2.1) and the definition of Goldstein δ-subdifferential,

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

The desired conclusion follows from (3.12), (3.26), (3.25) and the definition of Δ_h in (3.18). ∎

Remark 3.7 (Simplified expressions for the choice of σ). We present more explicit upper bounds for σ in Theorem 3.6(ii).

(i) (For a general $f \in \mathcal{SPB}(\mathbb{R}^d)$) Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1) and $\mathcal{X}$ be a compact set. Let

$$0 < \varepsilon < \min\{5\mathrm{R}_2, 1\} \quad \text{and} \quad 0 < \delta < 1. \tag{3.27}$$

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

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

where $\mathcal{P}$ is given in (3.16). Since $\mathcal{P} \geq 4\mathrm{R}_2$, we can deduce that

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

This means that for the ε and δ chosen as in (3.27), we indeed have

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

Using the definitions of $\mathfrak{M}_1$ and $\mathfrak{M}_2$ above and the definition of η_1 in (3.15), we can then deduce that[4]

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

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 [20, Eq. (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),

$$H \leq \sqrt{d\pi e}\,\eta_1^{-\frac{1}{d}}. \tag{3.31}$$

Finally, since $H > \sqrt{d}$ (see footnote 3) and we chose $0 < \delta < 1$ as stated in (3.27), it follows from (3.13) that $\bar{\sigma} = \min\left\{\left[\frac{\varepsilon}{2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}}\right]^{\frac{1}{m}}, \frac{\delta}{H}\right\}$. Therefore, for the ε and δ chosen as in (3.27), we have upon combining (3.29), (3.30) and (3.31) that the inclusion in Theorem 3.6(ii) holds when

$$\sigma \leq \min\left\{\left[2^{m+1}\mathrm{R}_1(m+d)^{\frac{m}{2}}\right]^{-\frac{1}{m}}, \delta\mathcal{P}^{-1/d}/\sqrt{d\pi e}\right\} \cdot \varepsilon^{\max\{\frac{1}{m},\frac{1}{d}\}}. \tag{3.32}$$

(ii) (For a globally Lipschitz f) Suppose that f is globally Lipschitz continuous and $\mathcal{X}$ is compact. Then one can choose $\mathrm{R}_1 = 0$, $\mathrm{R}_2 > 0$ and $m = d$ in (3.1). Thus, when ε and δ are chosen as in (3.27), we deduce from (3.32) that the inclusion in Theorem 3.6(ii) holds whenever $\sigma \leq \frac{\delta\mathcal{P}^{-1/d}}{\sqrt{d\pi e}}\varepsilon^{\frac{1}{d}}$.

The following corollary concerning gradient consistency is immediate.

Corollary 3.8 (GS gradient consistency). *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ and let ∇f_σ and $\partial_C f$ be given 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 $\sigma \to 0^+$ belongs to $\partial_C f(x)$.*

[4] Specifically, we deduce from $\varepsilon < \mathfrak{M}_1$ and the definition of η_1 that $\eta_1 = \varepsilon\mathcal{P}^{-1}$, and then from $\varepsilon < \mathfrak{M}_2$ and the definition of η_1 that $\eta_1^{\frac{2}{d}}/(2\pi e) < 1/10$.

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

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

where $\mathcal{D} \subseteq \mathbb{R}^d$ 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., [10, 16] 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 σ 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 tempted 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, 15, 21, 25]. We refer to them as GS-based algorithms. Most existing works rely on the global Lipschitz continuity of ∇f_σ, 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 ∇f_σ is locally Lipschitz with a polynomially bounded Lipschitz modulus of order $O(\|x\|^m + 1)$ for some $m \geq 1$. 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.[5]

Before presenting our settings and the corresponding algorithms, we first establish some auxiliary lemmas, which will be useful for the algorithmic analysis.

4.1. Auxiliary lemmas for complexity analysis. The first lemma quantifies the approximation error of f_σ.

LEMMA 4.1. *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1) and let f_σ be defined in Definition 1.1. Then it holds that*

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

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

$$\mathcal{M}(x) := (2^{m-1}\mathrm{R}_1\|x\|^m + \mathrm{R}_2)\sqrt{d} + 2^{m-1}\mathrm{R}_1\sigma^m(m+1+d)^{\frac{m+1}{2}}.$$

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

$$\begin{aligned} &|f_\sigma(x) - f(x)| \leq \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[|f(x+\sigma u) - f(x)|\right] \\ &\leq \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(2^{m-1}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\sigma^m\|u\|^m + \mathrm{R}_2\right)\cdot\sigma\|u\|\right] \\ &\leq \left[(2^{m-1}\mathrm{R}_1\|x\|^m + \mathrm{R}_2)\sqrt{d} + 2^{m-1}\mathrm{R}_1\sigma^m(m+1+d)^{\frac{m+1}{2}}\right]\cdot\sigma = \mathcal{M}(x)\cdot\sigma, \end{aligned}$$

[5]This section focuses on those SPB functions with $\mathrm{R}_1 > 0$. While our convergence results are still valid for globally Lipschitz functions (i.e., SPB functions with $\mathrm{R}_1 = 0$), we expect that in this case our results will be worse than existing ones specialized for globally Lipschitz functions (e.g., [17, 25]).

where the second inequality follows from Lemma 3.2 and the last inequality follows from [25, Lemma 1]. ◻

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., [23, Theorem 2.1.5]. We present a descent lemma of f_σ for SPB functions f, whose gradients ∇f_σ are not necessarily globally Lipschitz.

LEMMA 4.2 (Descent lemma). *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1), f_σ be defined in Definition 1.1, and $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ be given in (3.7). Then*

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

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

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

where (a) follows from Theorem 3.4(ii). ◻

We remark that another descent lemma for functions without a Lipschitz gradient has been formulated and studied in a very recent work [22] (see [22, 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 size 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. Our Lemma 4.2 is analogous to their descent lemma, but the adaptive stepsize derived from our descent lemma (see Theorems 4.6, 4.7 and 4.8 below) only depends on the current iterate and is given by an *explicit* formula.

The next two lemmas concern the random vector $\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)u$ 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), $x \in \mathbb{R}^d$, $\sigma > 0$ and p be a positive integer. Define $F(u) = \frac{1}{\sigma}[f(x+\sigma u) - f(x)]u$. Then*

$$\begin{aligned}
\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u)\|^p\right] \le{}& \left[3^{p-1}2^{(m-1)p}\mathrm{R}_1^p\|x\|^{mp} + 3^{p-1}\mathrm{R}_2^p\right](2p+d)^p \\
&+ 3^{p-1}2^{(m-1)p}\mathrm{R}_1^p\sigma^{mp}[(m+2)p+d]^{(m+2)p/2}.
\end{aligned} \tag{4.2}$$

Proof. First, from Lemma 3.2, for any $u \in \mathbb{R}^d$, we have

$$|f(x+\sigma u) - f(x)| \le (2^{m-1}\mathrm{R}_1\|x\|^m + 2^{m-1}\mathrm{R}_1\|\sigma u\|^m + \mathrm{R}_2)\|\sigma u\|.$$

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

$$\begin{aligned}
&\left|\frac{f(x+\sigma u) - f(x)}{\sigma}\right|^p \|u\|^p \\
&\le \left[3^{p-1}2^{(m-1)p}\mathrm{R}_1^p\|x\|^{mp} + 3^{p-1}\mathrm{R}_2^p\right]\|u\|^{2p} + 3^{p-1}2^{(m-1)p}\mathrm{R}_1^p\sigma^{mp}\|u\|^{(m+2)p}.
\end{aligned}$$

Taking the expectation on both sides of the above display with respect to $u \sim \mathcal{N}(0, I)$ and invoking [25, Lemma 1] for upper bounding moments of the form $\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|^k]$ for $k \geq 1$, we obtain the desired result. ∎

We have the following immediate corollary.

Corollary 4.4. *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1), $x \in \mathbb{R}^d$ and $\sigma > 0$. Define $F(u) = \frac{1}{\sigma}[f(x + \sigma u) - f(x)]u$. Then it holds that*

$$\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u)\|^{m+2}\right] \leq C_b(\|x\|^{m(m+2)} + 1) \tag{4.3}$$

and

$$\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u)\|^{2}\right] \leq \mathcal{H}_a(\|x\|^{2m} + 1), \tag{4.4}$$

where $C_b = \max\{\mathcal{H}_1, \mathcal{H}_2\}$, $\mathcal{H}_1 = 3^{m+1}(2m + 4 + d)^{m+2}2^{(m-1)(m+2)}\mathrm{R}_1^{m+2}$,

$$\begin{aligned} \mathcal{H}_a &= 3\max\{4^{m-1}\mathrm{R}_1^2(4+d)^2, \mathrm{R}_2^2(4+d)^2 + 4^{m-1}\mathrm{R}_1^2\sigma^{2m}(2m+4+d)^{m+2}\}, \\ \mathcal{H}_2 &= 3^{m+1}(2m+4+d)^{m+2}\mathrm{R}_2^{m+2} \\ &\quad + 3^{m+1}2^{(m-1)(m+2)}\mathrm{R}_1^{m+2}\sigma^{m(m+2)}((m+2)^2+d)^{(m+2)^2/2}. \end{aligned} \tag{4.5}$$

Proof. Set $p = m + 2$ (resp., $p = 2$) in (4.2) and apply the inequality $at + b \leq \max\{a, b\}(t + 1)$ (which holds for any a, b, $t \geq 0$) to obtain (4.3) (resp., (4.4)). ∎

Lemma 4.5. *Let $f \in \mathcal{SPB}(\mathbb{R}^d)$ with parameters R_1, R_2 and m as in (3.1), ∇f_σ be given in (3.3), and C_b be defined in Corollary 4.4. Let $x \in \mathbb{R}^d$ and define $F(u) = \frac{1}{\sigma}[f(x + \sigma u) - f(x)]u$. Then the following statements hold.*

(i) *Let $V(x) = \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u) - \nabla f_\sigma(x)\|^2\right]$, $\epsilon_v > 0$ and $\{u^1, \ldots, u^M\}$ be i.i.d. samples of $\mathcal{N}(0, I)$, where sample size $M \geq \frac{V(x)}{\epsilon_v^2}$. Then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} F(u^i) - \nabla f_\sigma(x)\right\|^2\right] \leq \epsilon_v^2,$$

where $\mathbb{E}$ is the expectation over $\{u^1, \ldots, u^M\}$.

(ii) *Let M be a positive integer and $\{u^1, \ldots, u^M\}$ be i.i.d. samples of $\mathcal{N}(0, I)$. Then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} F(u^i)\right\|^{m+2}\right] \leq C_b(\|x\|^m + 1)^{m+2},$$

where $\mathbb{E}$ is also the expectation over $\{u^1, \ldots, u^M\}$.

Proof. To prove (i), we see from (3.3) that $\nabla f_\sigma(x) = \mathbb{E}_{u\sim\mathcal{N}(0,I)}[F(u)]$. Let $\{u^1, \ldots, u^M\}$ be i.i.d. and $u^i \sim \mathcal{N}(0, I)$. Then

$$\begin{aligned} &\mathbb{E}\left[\left\|\frac{\sum_{i=1}^M F(u^i)}{M} - \nabla f_\sigma(x)\right\|^2\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{i=1}^{M}\left(F(u^i) - \nabla f_\sigma(x)\right)\right\|^2\right] \\ &= \frac{1}{M^2}\mathbb{E}\left[\sum_{i=1}^{M}\|F(u^i) - \nabla f_\sigma(x)\|^2\right] = \frac{1}{M^2}\sum_{i=1}^{M}\mathbb{E}_{u^i\sim\mathcal{N}(0,I)}\left[\|F(u^i) - \nabla f_\sigma(x)\|^2\right] \\ &= \frac{1}{M}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u) - \nabla f_\sigma(x)\|^2\right] = \frac{1}{M}V(x). \end{aligned}$$

The desired conclusion now follows immediately.

To prove (ii), by direct computation, we have

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}F(u^i)\right\|^{m+2}\right]\le\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{u^i\sim\mathcal{N}(0,I)}\left[\left\|F(u^i)\right\|^{m+2}\right]=\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|F(u)\|^{m+2}\right]$$
$$\le C_b(\|x\|^{m(m+2)}+1)\le C_b(\|x\|^m+1)^{(m+2)},$$

where the first inequality follows from the convexity of $\|\cdot\|^{m+2}$, and the second inequality follows from Corollary 4.4. ∎

4.2. A GS-based zeroth-order algorithm for problem (4.1). We now introduce our first algorithm for solving (4.1), whose objective function is SPB and not necessarily globally Lipschitz. The algorithm is presented in Algorithm 4.1 below. Contrary to Algorithms 4.2 and 4.3 presented in later subsections (they will be introduced in section 4.3 and cater to specific subclasses of problem (4.1)), Algorithm 4.1 is applicable to generic instances of problem (4.1). As explained in the discussions preceding section 4.1, the key idea of GS-based zeroth-order optimization algorithms is to approximate the gradient $\nabla f(x)$ by the random gradient estimator

$$v(x,M)=\frac{1}{M\sigma}\sum_{i=1}^{M}(f(x+\sigma u^i)-f(x))u^i, \tag{4.6}$$

where $u^1,\ldots,u^M$ are M independent copies of standard Gaussian random vectors. From Theorem 3.3, we see that $v(x,M)$ is an unbiased estimator of $\nabla f_\sigma(x)$.

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

1: **Input:** Initial point $x^0\in\mathcal{D}$, $\{\tau_k\}\subset(0,1]$, $\sigma>0$ and $\epsilon_v>0$. Let m be defined as in (3.1) corresponding to our $f\in\mathcal{SPB}(\mathbb{R}^d)$.
2: **for** $k=0,1,2,\ldots$ **do**
3: Generate M_k i.i.d. $u_1^k,\ldots,u_{M_k}^k\sim\mathcal{N}(0,I)$, with[6] $M_k\ge V(x^k)/\epsilon_v^2$, and form

$$v^k=\frac{1}{M_k\sigma}\sum_{i=1}^{M_k}(f(x^k+\sigma u_i^k)-f(x^k))u_i^k.$$

4: Compute

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

5: **end for**

Compared to the GS-based zeroth-order optimization algorithms developed under global Lipschitz assumption on f or its gradient (e.g., [25]), a distinctive characteristic of Algorithm 4.1 is the extra rescaling of the stepsize by the factor $\|x^k\|^m+1$.

The following result establishes the convergence rate of Algorithm 4.1 in terms of the complexity measure ω_k^2 in (4.8). Note that $\{x^k\}$ can be unbounded in general.

[6] Recall that $V(\cdot)$ is defined in Lemma 4.5(i). In situations where $V(\cdot)$ is not easy to obtain, one may invoke (4.4) to obtain an upper bound of $V(x^k)$: $V(x^k)=\mathbb{E}[\|F(u)\|^2]-\|\mathbb{E}[F(u)]\|^2\le\mathbb{E}[\|F(u)\|^2]\le\mathcal{H}_a(\|x^k\|^{2m}+1)$ (where $F(u):=(f(x^k+\sigma u)-f(x^k))/\sigma$). Then any $M_k\ge\mathcal{H}_a(\|x^k\|^{2m}+1)/\epsilon_v^2$ satisfies the condition in line 3 of Algorithm 4.1.

Therefore, the quantity ω_k^2, which is rescaled by $\|x^k\|^m + 1$, may be interpreted as a *relative* stationarity measure for the problem $\min_{x\in\mathcal{D}} f_\sigma(x)$.[7] In section 4.4, we will discuss how explicit complexity bounds for (δ,ϵ)-stationarity can be obtained using Theorem 4.6.

THEOREM 4.6 (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). Let ∇f_σ be given in (3.3) and $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be given in (3.7). Suppose that there exists $\theta_0 \in \mathbb{R}$ such that $\inf_{x\in\mathbb{R}^d} f(x) \geq \theta_0$ and that $0 < \tau_k \leq (\max\{2\mathcal{A}, 2\mathcal{B}, 1\})^{-1}$ for all k. Then the sequence $\{x^k\}$ generated by Algorithm 4.1 satisfies that for any $T > 0$,*

$$\min_{0\leq k\leq T} \omega_k^2 \leq \frac{W_T}{\sum_{k=0}^T \tau_k}, \tag{4.7}$$

where

$$W_T = 8f(x^0) + 8\mathcal{M}(x^0)\cdot\sigma - 8\theta_0 + 12\epsilon_v^2 \sum_{k=0}^T \tau_k + \frac{8C_b\cdot\mathcal{C}}{m+2}\sum_{k=0}^T \tau_k^{m+2},$$

$$\omega_k^2 = \mathbb{E}\left[\frac{\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2}{\|x^k\|^m + 1}\right], \tag{4.8}$$

$\mathcal{M}(\cdot)$ is defined in Lemma 4.1, and C_b is defined in Corollary 4.4.

Proof. For notational simplicity, write

$$\alpha_k := \tau_k/(\|x^k\|^m + 1) \in (0, 1]. \tag{4.9}$$

Then we can obtain from Lemma 4.2 that

$$\begin{aligned} f_\sigma(x^{k+1}) &\leq f_\sigma(x^k) - \langle \nabla f_\sigma(x^k), x^k - x^{k+1}\rangle \\ &\quad + \frac{1}{2}\|x^k - x^{k+1}\|^2\left[\mathcal{A} + \mathcal{B}\|x^k\|^m + \frac{2\mathcal{C}\|x^{k+1} - x^k\|^m}{m+2}\right] \\ &= f_\sigma(x^k) - \frac{1}{\alpha_k}\langle \alpha_k v^k, x^k - x^{k+1}\rangle - \langle \nabla f_\sigma(x^k) - v^k, x^k - x^{k+1}\rangle \\ &\quad + \frac{1}{2}\|x^k - x^{k+1}\|^2\left[\mathcal{A} + \mathcal{B}\|x^k\|^m + \frac{2\mathcal{C}\|x^{k+1} - x^k\|^m}{m+2}\right], \end{aligned} \tag{4.10}$$

Now, using the definition of x^{k+1} as a projection, we have from [5, Lemma 2.2] and [5, Lemma 2.1] that

$$\alpha_k^2\|x^k - P_{\mathcal{D}}(x^k - v^k)\|^2 \leq \|x^k - x^{k+1}\|^2 \leq \langle \alpha_k v^k, x^k - x^{k+1}\rangle. \tag{4.11}$$

Combining the second inequality in (4.11) with (4.10), we obtain

$$\begin{aligned} f_\sigma(x^{k+1}) - f_\sigma(x^k) &\leq -\frac{1}{\alpha_k}\|x^k - x^{k+1}\|^2 - \langle \nabla f_\sigma(x^k) - v^k, x^k - x^{k+1}\rangle \\ &\quad + \frac{1}{2}\|x^k - x^{k+1}\|^2\left[\mathcal{A} + \mathcal{B}\|x^k\|^m + \frac{2\mathcal{C}\|x^{k+1} - x^k\|^m}{m+2}\right]. \end{aligned} \tag{4.12}$$

[7] Recall that x^* is stationary for $\min_{x\in\mathcal{D}} f_\sigma(x)$ if and only if $x^* = P_{\mathcal{D}}(x^* - \nabla f_\sigma(x^*))$.

To bound the right-hand side of (4.12), we will bound $-\langle \nabla f_\sigma(x^k) - v^k, x^k - x^{k+1}\rangle$ and $\|x^{k+1} - x^k\|^{m+2}$. We first deduce that

$$\begin{aligned}-\langle \nabla f_\sigma(x^k) - v^k, x^k - x^{k+1}\rangle &\le \|v^k - \nabla f_\sigma(x^k)\| \cdot \|x^k - x^{k+1}\| \\ &\le \frac{\alpha_k}{2}\|v^k - \nabla f_\sigma(x^k)\|^2 + \frac{1}{2\alpha_k}\|x^k - x^{k+1}\|^2. \end{aligned} \tag{4.13}$$

Next, observing that $x^k \in \mathcal{D}$ and $P_\mathcal{D}$ is nonexpansive, we have

$$\|x^{k+1} - x^k\|^{m+2} = \|P_\mathcal{D}(x^k - \alpha_k v^k) - P_\mathcal{D}(x^k)\|^{m+2} \le \alpha_k^{m+2}\|v^k\|^{m+2}.$$

Using the above bounds on $\|x^{k+1} - x^k\|^{m+2}$ and $-\langle \nabla f_\sigma(x^k) - v^k, x^k - x^{k+1}\rangle$ (see (4.13)), we can then deduce the following inequality from (4.12):

$$\begin{aligned}&f_\sigma(x^{k+1}) - f_\sigma(x^k) \\ &\le -\frac{1}{\alpha_k}\|x^k - x^{k+1}\|^2 + \frac{\alpha_k}{2}\|v^k - \nabla f_\sigma(x^k)\|^2 + \frac{1}{2\alpha_k}\|x^k - x^{k+1}\|^2 \\ &\quad + \frac{1}{2}\|x^k - x^{k+1}\|^2(\mathcal{A} + \mathcal{B}\|x^k\|^m) + \frac{\mathcal{C}}{m+2}\alpha_k^{m+2}\|v^k\|^{m+2} \\ &\le -S_k\|x^k - x^{k+1}\|^2 + D_k + \beta B_k, \end{aligned} \tag{4.14}$$

where

$$S_k = \frac{1}{2\alpha_k} - \frac{1}{2}\left(\mathcal{B}\|x^k\|^m + \mathcal{A}\right), \quad D_k = \frac{\alpha_k}{2}\|v^k - \nabla f_\sigma(x^k)\|^2, \tag{4.15}$$

$$B_k = \alpha_k^{m+2}\|v^k\|^{m+2} \quad \text{and} \quad \beta = \frac{\mathcal{C}}{m+2}. \tag{4.16}$$

Now, upon rearranging terms in (4.14), we see that

$$S_k\|x^k - x^{k+1}\|^2 \le f_\sigma(x^k) - f_\sigma(x^{k+1}) + D_k + \beta B_k. \tag{4.17}$$

Combining (4.17) with the first inequality in (4.11), we obtain that

$$S_k\alpha_k^2\|x^k - P_\mathcal{D}(x^k - v^k)\|^2 \le f_\sigma(x^k) - f_\sigma(x^{k+1}) + D_k + \beta B_k. \tag{4.18}$$

We next analyze the terms $S_k\alpha_k^2$ and $\|x^k - P_\mathcal{D}(x^k - v^k)\|^2$ in (4.18).

We start with $S_k\alpha_k^2$. Recall that $0 < \tau_k \le 1/\max\{2\mathcal{A}, 2\mathcal{B}\}$. Then we have from the definitions of S_k in (4.15) and α_k in (4.9) that

$$\frac{1}{4}\frac{\tau_k}{\|x^k\|^m + 1} = \frac{1}{4}\alpha_k \le \underbrace{\left[\frac{1}{2} - \frac{\alpha_k}{2}\left(\mathcal{B}\|x^k\|^m + \mathcal{A}\right)\right]\alpha_k}_{S_k\alpha_k^2} \le \alpha_k \le \tau_k. \tag{4.19}$$

Next, for $\|x^k - P_\mathcal{D}(x^k - v^k)\|^2$, we consider the following relations:

$$\begin{aligned}&\|x^k - P_\mathcal{D}(x^k - \nabla f_\sigma(x^k))\|^2 \\ &= \|x^k - P_\mathcal{D}(x^k - \nabla f_\sigma(x^k)) - [x^k - P_\mathcal{D}(x^k - v^k)] + x^k - P_\mathcal{D}(x^k - v^k)\|^2 \\ &\le \left(\|P_\mathcal{D}(x^k - v^k) - P_\mathcal{D}(x^k - \nabla f_\sigma(x^k))\| + \|x^k - P_\mathcal{D}(x^k - v^k)\|\right)^2 \\ &\le 2\left(\|x^k - P_\mathcal{D}(x^k - v^k)\|^2 + \|\nabla f_\sigma(x^k) - v^k\|^2\right), \end{aligned} \tag{4.20}$$

where the second inequality follows from the nonexpansiveness of $P_{\mathcal{D}}$ and the elementary relation $(a+b)^2 \le 2a^2 + 2b^2$ for any a, $b \in \mathbb{R}$.

Based on the estimates (4.19) and (4.20) as well as the inequality in (4.18), we can deduce that

$$
\begin{aligned}
&\frac{\alpha_k}{4}\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2 \overset{\text{(a)}}{\le} S_k\alpha_k^2\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2\\
&\overset{\text{(b)}}{\le} 2S_k\alpha_k^2\left(\|x^k - P_{\mathcal{D}}(x^k - v^k)\|^2 + \|\nabla f_\sigma(x^k) - v^k\|^2\right)\\
&\overset{\text{(c)}}{\le} 2f_\sigma(x^k) - 2f_\sigma(x^{k+1}) + 2D_k + 2S_k\alpha_k^2\|\nabla f_\sigma(x^k) - v^k\|^2 + 2\beta B_k\\
&\overset{\text{(d)}}{\le} 2f_\sigma(x^k) - 2f_\sigma(x^{k+1}) + 2D_k + 2\tau_k\|\nabla f_\sigma(x^k) - v^k\|^2 + 2\beta B_k\\
&\le 2f_\sigma(x^k) - 2f_\sigma(x^{k+1}) + 3\tau_k\|\nabla f_\sigma(x^k) - v^k\|^2 + 2\beta B_k,
\end{aligned}
\tag{4.21}
$$

where (a) and (d) follow from the first and third inequalities in (4.19), respectively, (b) follows from (4.20), (c) follows from (4.18), and the last inequality follows from (4.15) and (4.19). We now take expectation on both sides of (4.21) with respect to the random variable v^k to obtain

$$
\begin{aligned}
&0.25\alpha_k\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2\\
&\le 2f_\sigma(x^k) - 2\mathbb{E}_{v^k}[f_\sigma(x^{k+1})] + 3\tau_k\mathbb{E}_{v^k}[\|\nabla f_\sigma(x^k) - v^k\|^2] + 2\beta\mathbb{E}_{v^k}[B_k]\\
&\overset{\text{(a)}}{\le} 2f_\sigma(x^k) - 2\mathbb{E}_{v^k}[f_\sigma(x^{k+1})] + 3\tau_k\epsilon_v^2 + 2\beta\mathbb{E}_{v^k}[B_k],
\end{aligned}
\tag{4.22}
$$

where we use Lemma 4.5(i) in (a), thanks to the choice of M_k in the algorithm.

To further upper bound the right hand side of (4.22), we notice upon invoking the definition of B_k in (4.16), the definition of α_k in (4.9) and Lemma 4.5(ii) that

$$
\mathbb{E}_{v^k}[B_k] \le \tau_k^{m+2}C_b. \tag{4.23}
$$

Combining (4.22) and (4.23), we obtain that

$$
\alpha_k\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2 \le 8f_\sigma(x^k) - 8\mathbb{E}_{v^k}[f_\sigma(x^{k+1})] + 12\tau_k\epsilon_v^2 + 8\beta C_b\tau_k^{m+2},
$$

which, upon taking expectations on both sides, yields that

$$
\begin{aligned}
&\mathbb{E}\big[\alpha_k\|x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k))\|^2\big]\\
&\le 8\mathbb{E}\left[f_\sigma(x^k)\right] - 8\mathbb{E}\left[f_\sigma(x^{k+1})\right] + 12\tau_k\epsilon_v^2 + 8\beta C_b\tau_k^{m+2}.
\end{aligned}
$$

Summing the above display on both sides from $k = 0$ to T and recalling the definitions of ω_k^2 in (4.8) and α_k in (4.9), we arrive at

$$
\sum_{k=0}^{T}\tau_k\omega_k^2 \le 8f_\sigma(x^0) - 8\mathbb{E}\left[f_\sigma(x^{T+1})\right] + 12\epsilon_v^2\sum_{k=0}^{T}\tau_k + 8\beta C_b\sum_{k=0}^{T}\tau_k^{m+2}. \tag{4.24}
$$

To obtain the desired conclusion, it remains to estimate the terms involving f_σ on the right hand side of (4.24). To this end, notice that for all $x \in \mathcal{D}$, we have from Lemma 4.1 that $|f_\sigma(x) - f(x)| \le \mathcal{M}(x)\cdot\sigma$. In addition, $\inf f \ge \theta_0$ implies that $\inf f_\sigma \ge \theta_0$. Applying these two observations to upper bound the $f_\sigma(x^0)$ in (4.24) by $f(x^0) + \mathcal{M}(x^0)\sigma$ and lower bound the $f_\sigma(x^{T+1})$ in (4.24) by θ_0, we obtain further that

$$
\sum_{k=0}^{T}\tau_k\omega_k^2 \le 8f(x^0) + 8\mathcal{M}(x^0)\sigma - 8\theta_0 + 12\epsilon_v^2\sum_{k=0}^{T}\tau_k + 8\beta C_b\sum_{k=0}^{T}\tau_k^{m+2},
$$

which yields the desired result upon recalling the definition of β in (4.16). □

4.3. Single-sample variants for special cases. Although Algorithm 4.1 can be applied to generic instances of problem (4.1) with $\inf f > -\infty$ and enjoys the complexity bound in Theorem 4.6, the required number $V(x^k)/\epsilon_v^2$ of function evaluations is large when $\epsilon_v > 0$ is small. A natural question is then whether one can develop an algorithm with a smaller number of function evaluations. When f is globally Lipschitz or has Lipschitz gradient, together with some additional assumptions such as the convexity of f or $\mathcal{D} = \mathbb{R}^d$, the answer is positive. In fact, under such assumptions, the number of function evaluations at each iteration can be reduced to a constant, often only one or two; see, e.g., [17,25]. In view of this, below we study two specific situations where a certain variant of Algorithm 4.1 requires only two function evaluations and yet retains a similar (iteration) complexity guarantee.

4.3.1. When f is convex. 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 4.2, which in contrast to Algorithm 4.1, generates only one Gaussian vector at each iteration. In other words, $M_k = 1$ for all $k \geq 0$. Algorithm 4.2 should also be compared with the one developed in [25] for convex and globally Lipschitz objective functions f. Our algorithm differs from theirs in that the stepsize is rescaled by $\|x^k\|^m + 1$ to account for the lack of Lipschitzness, see [25, Eq. (50)].

Algorithm 4.2 GS-based zeroth-order algorithm for convex problem (4.1)

1: **Input:** Initial point $x^0 \in \mathcal{D}$, $\{\tau_k\} \subset (0, 1]$ and $\sigma > 0$. Let m be defined as in (3.1) corresponding to our $f \in \mathcal{SPB}(\mathbb{R}^d)$.
2: **for** $k = 0, 1, 2, \ldots$ **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
$$x^{k+1} = P_{\mathcal{D}}\left(x^k - \tau_k \cdot \frac{v^k}{\|x^k\|^m + 1}\right).$$
5: **end for**

THEOREM 4.7 (Complexity bound for Algorithm 4.2). *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 4.2 satisfies that for any $T > 0$,*

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

where $\mathcal{H}_a$ is given in (4.5) and $\mathcal{M}(\cdot)$ is defined in Lemma 4.1.

Proof. Since the projection operator is nonexpansive, we see that

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

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

$$\begin{aligned}&2\alpha_k\langle \nabla f_\sigma(x^k), x^k - x^*\rangle \\ &\leq \|x^k - x^*\|^2 - \mathbb{E}_{u^k\sim\mathcal{N}(0,I)}[\|x^{k+1} - x^*\|^2] + \alpha_k^2\mathbb{E}_{u^k\sim\mathcal{N}(0,I)}[\|v^k\|^2].\end{aligned} \tag{4.25}$$

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

$$f_\sigma(x^k) - f_\sigma(x^*) \le \langle \nabla f_\sigma(x^k), x^k - x^* \rangle. \tag{4.26}$$

In addition, we have from Corollary 4.4 that

$$\mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^2\right] \le \mathcal{H}_a(\|x^k\|^{2m} + 1) \le \mathcal{H}_a(\|x^k\|^m + 1)^2, \tag{4.27}$$

where $\mathcal{H}_a$ is given in (4.5). Combining (4.25), (4.26) and (4.27) yields

$$2\alpha_k[f_\sigma(x^k) - f_\sigma(x^*)] \le \|x^k - x^*\|^2 - \mathbb{E}_{u^k \sim \mathcal{N}(0,I)}[\|x^{k+1} - x^*\|^2] + \mathcal{H}_a \tau_k^2. \tag{4.28}$$

We now lower bound the left-hand side of (4.28). To this end, we first note from [25, Eq. (11)] that

$$f_\sigma(x^k) \ge f(x^k). \tag{4.29}$$

Also, we have

$$f_\sigma(x^*) = f_\sigma(x^*) - f(x^*) + f(x^*) \le \mathcal{M}(x^*) \cdot \sigma + f(x^*), \tag{4.30}$$

where the inequality follows from Lemma 4.1. Using (4.29) and (4.30), we can lower bound the left hand side of (4.28), which in turn yields

$$2\alpha_k[f(x^k) - f(x^*)] \le \|x^k - x^*\|^2 - \mathbb{E}_{u^k \sim \mathcal{N}(0,I)}[\|x^{k+1} - x^*\|^2] + 2\mathcal{M}(x^*)\sigma \cdot \alpha_k + \mathcal{H}_a \tau_k^2.$$

Taking the expectation on both sides of the above display and invoking the definition of α_k (and noting also that $\alpha_k \le \tau_k$), we deduce further that

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

which upon summing over k, completes the proof. ∎

4.3.2. Unconstrained minimization. Here, we consider problem (4.1) with $\mathcal{D} = \mathbb{R}^d$. The specific algorithm is presented in Algorithm 4.3 below. Notice that the update rule for x^k differs from that of Algorithm 4.1 in two aspects. First, the v^k is chosen as in (4.6) with $M = 1$ and hence *does not* enjoy the bound in Lemma 4.5(i). Second, since we do not have the bound in Lemma 4.5(i) for v^k or the convexity for f, 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) different than the one in [17, Algorithm 1], which is designed for f being globally Lipschitz.

THEOREM 4.8 (Complexity bound for Algorithm 4.3). *Consider* (4.1), *where* $f \in \mathcal{SPB}(\mathbb{R}^d)$ *with parameters* R_1, R_2 *and* m *as in* (3.1). *Let* ∇f_σ *be given in* (3.3) *and* $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ *be given in* (3.7). *Assume in addition that* $\mathcal{D} = \mathbb{R}^d$ *and there exists* $\theta_0 \in \mathbb{R}$ *such that* $\inf_{x \in \mathbb{R}^d} f(x) \ge \theta_0$. *Then the sequence* $\{x^k\}$ *be generated by Algorithm 4.3 satisfies that for any* $T > 0$,

$$\begin{aligned} &\min_{0 \le k \le T} \mathbb{E}\left[\left\|\frac{\nabla f_\sigma(x^k)}{\|x^k\|^m + 1}\right\|^2\right] \\ &\le \frac{1}{\sum_{k=0}^T \tau_k}\left[f(x^0) - \theta_0 + \mathcal{M}(x^0) \cdot \sigma + \frac{\mathcal{H}_a \cdot (\mathcal{A} + \mathcal{B})}{2} \sum_{k=0}^T \tau_k^2 + \frac{C_b \cdot \mathcal{C}}{m+2} \sum_{k=0}^T \tau_k^{m+2}\right], \end{aligned}$$

where $\mathcal{H}_a$ *is given in* (4.5), C_b *is defined in Corollary 4.4, and* $\mathcal{M}(\cdot)$ *is defined in Lemma 4.1.*

Algorithm 4.3 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** $k = 0, 1, 2, \ldots$ **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**

Proof. From Lemma 4.2, we know that for any $k \geq 0$,

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

This implies that

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

where

$$\widetilde{\alpha}_k = \tau_k / (\|x^k\|^{2m} + 1). \tag{4.31}$$

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

$$\begin{aligned} \widetilde{\alpha}_k \|\nabla f_\sigma(x^k)\|^2 &\leq f_\sigma(x^k) - \mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[f_\sigma(x^{k+1})\right] \\ &\quad + \frac{(\mathcal{A} + \mathcal{B}\|x^k\|^m)\widetilde{\alpha}_k^2}{2}\mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^2\right] + \frac{\mathcal{C}\widetilde{\alpha}_k^{m+2}}{m+2}\mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^{m+2}\right]. \end{aligned} \tag{4.32}$$

We now upper bound the two terms $(\mathcal{A} + \mathcal{B}\|x^k\|^m)\widetilde{\alpha}_k^2 \mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^2\right]$ and $\widetilde{\alpha}_k^{m+2}\mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^{m+2}\right]$ in (4.32). For the former term, we have

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

where (a) follows from Corollary 4.4 with $\mathcal{H}_a$ given in (4.5), and we used the definition of $\widetilde{\alpha}_k$ in (4.31) for the equality. As for the latter term (i.e., $\widetilde{\alpha}_k^{m+2}\mathbb{E}_{u^k \sim \mathcal{N}(0,I)}\left[\|v^k\|^{m+2}\right]$), we can deduce from Corollary 4.4 that

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

Combining (4.32) with the above two displays, one has

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

where (a) follows from (4.31). Summing both sides of the above display from $k = 0$ to T, we obtain further that

$$\sum_{k=0}^{T} \tau_k \mathbb{E}\left[\left\|\frac{\nabla f_\sigma(x^k)}{\|x^k\|^m + 1}\right\|^2\right]$$
$$\leq f_\sigma(x^0) - \mathbb{E}\left[f_\sigma(x^{T+1})\right] + 0.5\mathcal{H}_a \cdot (\mathcal{A} + \mathcal{B}) \sum_{k=0}^{T} \tau_k^2 + \frac{C_b \cdot \mathcal{C}}{m+2} \sum_{k=0}^{T} \tau_k^{m+2}.$$

Finally, we observe that $f_\sigma(x^0) - \mathbb{E}\left[f_\sigma(x^{T+1})\right] \leq f_\sigma(x^0) - \theta_0 \leq \mathcal{M}(x^0) \cdot \sigma + f(x^0) - \theta_0$, where we used Lemma 4.1 for the last inequality. The desired result now follows immediately upon combining this last observation with the above display. ◻

4.4. Explicit complexity and (relaxed) (δ, ϵ)-stationarity. When $\mathcal{D}$ in (4.1) is compact, we interpret the bound (4.7) in terms of an approximate optimality condition of (4.1) based on Goldstein subdifferential, leveraging our consistency results in section 3.2. Using [7, Proposition 2.4.4], the corollary of [7, Proposition 2.4.3], and the definition of Goldstein δ-subdifferential, one can readily show that if x^* is locally optimal for (4.1), then for any $\delta > 0$, we have

$$0 \in \partial_C f(x^*) + N_{\mathcal{D}}(x^*) \subseteq \partial_G^\delta f(x^*) + N_{\mathcal{D}}(x^*).$$

In what follows, we outline how the bound on ω_k^2 in (4.7) allows us to derive the complexity for obtaining a (relaxed) (δ, ϵ)-stationary point $\hat{x}$ in the sense that $\mathrm{dist}(0, \partial_G^\delta f(\hat{x}) + N_{\mathcal{D}}(\hat{x})) < \epsilon$ when $\mathcal{D}$ in (4.1) is in addition compact.

To this end, consider (4.1) and suppose there exists θ_0 satisfying $\inf_{x\in\mathbb{R}^d} f(x) \geq \theta_0$. Since $\mathcal{D}$ is compact, according to Theorem 3.6(ii), for given $\delta > 0$ and $\epsilon > 0$, there exists $\bar{\sigma} > 0$ such that

$$\nabla f_\sigma(x) \in \partial_G^\delta f(x) + (\epsilon/3)\mathbb{B} \quad \forall x \in \mathcal{D} \text{ and } \forall \sigma \in (0, \bar{\sigma}]. \tag{4.33}$$

Fix any such σ (see also Remark 3.7). Then we see from Theorem 3.4(ii) that

$$\|\nabla f_\sigma(x) - \nabla f_\sigma(y)\| \leq L_1\|x - y\| \quad \forall x, y \in \mathcal{D}, \tag{4.34}$$

where $L_1 = \mathcal{A} + \mathcal{B}C_1^m + 2^m C_1^m \mathcal{C}$ with $C_1 = \sup_{x\in\mathcal{D}} \|x\|$.

Let $\{x^k\}$ and $\{v^k\}$ be generated by Algorithm 4.1 with σ, $\epsilon_v > 0$ and $\{\tau_k\}$ satisfying the assumptions in Theorem 4.6. Let

$$\tilde{x}^k = P_{\mathcal{D}}(x^k - v^k) \text{ and } y^k = x^k - P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k)).$$

Then we have

$$\left\|x^k - \tilde{x}^k\right\| = \left\|x^k - P_{\mathcal{D}}(x^k - v^k)\right\| = \left\|y^k + P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k)) - P_{\mathcal{D}}(x^k - v^k)\right\|$$
$$\leq \left\|y^k\right\| + \left\|P_{\mathcal{D}}(x^k - \nabla f_\sigma(x^k)) - P_{\mathcal{D}}(x^k - v^k)\right\| \leq \left\|y^k\right\| + \left\|\nabla f_\sigma(x^k) - v^k\right\|, \tag{4.35}$$

where the last inequality follows from the nonexpansiveness of $P_{\mathcal{D}}$. In addition, we can rewrite $\tilde{x}^k$ as

$$\tilde{x}^k = P_{\mathcal{D}}(\tilde{x}^k + G(x^k)),$$

where $G(x^k) = -\nabla f_\sigma(\tilde{x}^k) + (x^k - \tilde{x}^k) + \left[\nabla f_\sigma(\tilde{x}^k) - \nabla f_\sigma(x^k)\right] + \left[\nabla f_\sigma(x^k) - v^k\right]$. The above display implies that $G(x^k) \in N_{\mathcal{D}}(\tilde{x}^k)$. Therefore, one has

$$(x^k - \tilde{x}^k) + \left[\nabla f_\sigma(\tilde{x}^k) - \nabla f_\sigma(x^k)\right] + \left[\nabla f_\sigma(x^k) - v^k\right] \in \nabla f_\sigma(\tilde{x}^k) + N_{\mathcal{D}}(\tilde{x}^k). \tag{4.36}$$

From our choice of σ (so that (4.33) holds) and (4.36), we have

$$\begin{aligned}&\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\\&\le \|x^k-\tilde{x}^k\|+\|\nabla f_\sigma(\tilde{x}^k)-\nabla f_\sigma(x^k)\|+\|\nabla f_\sigma(x^k)-v^k\|+\epsilon/3\\&\le (1+L_1)\|x^k-\tilde{x}^k\|+\|\nabla f_\sigma(x^k)-v^k\|+\epsilon/3,\end{aligned}\tag{4.37}$$

where the last inequality follows from (4.34). Combining (4.35) and (4.37) yields

$$\begin{aligned}&\left[\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\right]^2\le\left[(1+L_1)\|y^k\|+(2+L_1)\|\nabla f_\sigma(x^k)-v^k\|+\epsilon/3\right]^2\\&\overset{\text{(a)}}{\le} 3(1+L_1)^2\|y^k\|^2+3(2+L_1)^2\|\nabla f_\sigma(x^k)-v^k\|^2+\epsilon^2/3,\end{aligned}$$

where (a) follows from the convexity of $(\cdot)^2$. Taking the expectation with respect to the random variable v^k on both sides of the above inequality, we can obtain

$$\begin{aligned}&\mathbb{E}_{v^k}\left[\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\right]^2\\&\le 3(1+L_1)^2\|y^k\|^2+3(2+L_1)^2\mathbb{E}_{v^k}\left[\|\nabla f_\sigma(x^k)-v^k\|^2\right]+\epsilon^2/3\\&\le 3(1+L_1)^2\|y^k\|^2+3(2+L_1)^2\epsilon_v^2+\epsilon^2/3,\end{aligned}$$

where the second inequality follows from the choice of M_k in Algorithm 4.1 and Lemma 4.5(i). Finally, taking the expectation over the trajectory $\{x^0,\dots,x^k\}$ and taking the min over $k\in\{0,\dots,T\}$ on both sides of the above display, we obtain that

$$\begin{aligned}&\min_{0\le k\le T}\mathbb{E}\left[\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\right]^2\\&\le 3(1+L_1)^2\min_{0\le k\le T}\mathbb{E}\left[\|y^k\|^2\right]+3(2+L_1)^2\epsilon_v^2+\epsilon^2/3\\&\overset{\text{(a)}}{\le}\frac{3(1+L_1)^2}{\sum_{k=0}^T\tau_k}\left[8f(x^0)+8\mathcal{M}(x^0)\sigma-8\theta_0+\frac{8C_b\cdot\mathcal{C}}{m+2}\sum_{k=0}^T\tau_k^{m+2}\right](C_1^m+1)\\&\quad+36(1+L_1)^2\epsilon_v^2(C_1^m+1)+3(2+L_1)^2\epsilon_v^2+\epsilon^2/3,\end{aligned}\tag{4.38}$$

where (a) follows from (4.7) and the definition $C_1=\sup_{x\in\mathcal{D}}\|x\|$.

Now, suppose that for these T iterations, we set $\tau_k=\left(\frac{\gamma}{T+1}\right)^{\frac{1}{m+2}}$ for some $0<\gamma\le(\max\{2\mathcal{A},2\mathcal{B},1\})^{-m-2}$. Then we deduce from (4.38) that

$$\begin{aligned}&\min_{0\le k\le T}\mathbb{E}\left[\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\right]^2\\&\le\frac{3(1+L_1)^2}{\gamma^{\frac{1}{m+2}}(T+1)^{\frac{m+1}{m+2}}}\left[8f(x^0)+8\mathcal{M}(x^0)\sigma-8\theta_0+\frac{8C_b\cdot\mathcal{C}}{m+2}\gamma\right](C_1^m+1)\\&\quad+\left[36(1+L_1)^2(C_1^m+1)+3(2+L_1)^2\right]\epsilon_v^2+\epsilon^2/3.\end{aligned}\tag{4.39}$$

Consequently, to obtain an $\tilde{x}^k$ with $\min_{0\le k\le T}\mathbb{E}\left[\mathrm{dist}(0,\partial_G^\delta f(\tilde{x}^k)+N_{\mathcal{D}}(\tilde{x}^k))\right]<\epsilon$, it suffices to choose σ as described above, and then choose ϵ_v in Algorithm 4.1 such that

$$\epsilon_v<\epsilon/\sqrt{3\left[36(1+L_1)^2(C_1^m+1)+3(2+L_1)^2\right]},$$

and finally choose T such that the first term in (4.39) is below $\epsilon^2/3$.

Acknowledgements. The authors would like to thank Wenqing Ouyang for discussions on the tail bound and tempered distribution used in section 3.

REFERENCES

[1] K. Balasubramanian and S. Ghadimi. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics* 22, pp. 35–76 (2022).

[2] A. Berahas, L. Cao, K. Choromanski and K. Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics* 22, pp. 507–560 (2022).

[3] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming* 188, pp. 19–51 (2021).

[4] J. M. Borwein and S. B. Lindstrom. Meetings with Lambert W and other special functions in optimization and analysis. *Pure and Applied Functional Analysis* 1, pp. 361–396 (2016).

[5] P. H. Calamai and J. J. Moré. Projected gradient methods for linearly constrained problems. *Mathematical Programming* 39, pp. 93–116 (1987).

[6] 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, Hawaii, USA. PMLR 202 (2023).

[7] F. H. Clarke. *Optimization and Nonsmooth Analysis.* Wiley, New York, (1983).

[8] L. Grafakos. *Classical Fourier Analysis.* Springer, New York, (2008).

[9] E. M. Stein, R. Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis.* Princeton University Press, Princeton, (2011).

[10] A. R. Conn, K. Scheinberg and L. N. Vicente. *Introduction to Derivative-free Optimization.* MOS-SIAM Series on Optimization, SIAM (2009).

[11] A. Dereventsov, C. G. Webster and J. Daws. An adaptive stochastic gradient-free approach for high-dimensional blackbox optimization. In: Tiwari, R., Mishra, A., Yadav, N., Pavone, M. (eds) *Proceedings of International Conference on Computational Intelligence. Algorithms for Intelligent Systems.* Springer, Singapore (2022).

[12] F. N. Fritsch, R. E. Shafer and W. P. Crowly Solution of the transcendental equation $we^w = x$. *Communications of the ACM* 16, pp. 123–124 (1973).

[13] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming* 13, pp. 14–22 (1977).

[14] J. Jin, B. Zhang, H. Wang and L. Wang. Non-convex distributionally robust optimization: Nonasymptotic analysis. *Advances in Neural Information Processing Systems* 34, pp.2771–2782 (2021).

[15] W. Jongeneel, M.-C. Yue and D. Kuhn. Small errors in random zeroth-order optimization are imaginary. Preprint (2021). Available at https://arxiv.org/abs/2103.05478.

[16] J. Larson, M. Menickelly and S. M. Wild. Derivative-free optimization methods. *Acta Numerica* 28, pp. 287–404 (2019).

[17] T. Lin, Z. Zheng and M. I. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. Preprint (2022). Available at https://arxiv.org/abs/2209.05045.

[18] V. S. Vladimirov. *Methods of the Theory of Generalized Functions.* Taylor & Francis, London, (2002).

[19] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, (2011).

[20] L. Lóczi. Guaranteed- and high-precision evaluation of the Lambert W function. *Applied Mathematics and Computation* 433, (2022).

[21] 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 Journal on Optimization* 28, pp. 1478–1507 (2018).

[22] A. Mishkin, A. Khaled, Y. Wang, A. Defazio and R. M. Gower. Directional smoothness and gradient methods: convergence and adaptivity. Preprint (2024). Available at https://arxiv.org/abs/2403.04081.

[23] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Springer (2004).

[24] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley (1983).

[25] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17, pp. 527–566 (2017).

[26] S. Osher, B. Wang, P. Yin, X. Luo, F. Barekat, M. Pham and A. Lin. Laplacian smoothing gradient descent. *Research in the Mathematical Sciences* 9, (2022).

[27] F. D. Parker. Integrals of inverse functions. *American Mathematical Monthly* 62, pp. 439–440 (1955).

[28] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc. (1983).
[29] S. Dasgupta, A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures Algorithms* 22, pp. 60–65 (2003).
[30] A. M.-C. So. *A Semidefinite Programming Approach to the Graph Realization Problem: Theory, Applications and Extensions*. Ph.D. thesis, Stanford University (2007).
[31] A. Starnes, A. Dereventsov and C. Webster. Gaussian smoothing gradient descent for minimizing functions (GSmoothGD). Preprint (2023). Available at https://arxiv.org/abs/2311.00521v2.
[32] A. Starnes and C. Webster. Gaussian smoothing stochastic gradient descent (GSmoothSGD). Preprint (2023). Available at https://arxiv.org/abs/2311.00531.
[33] 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 pp. 10162–10172 (2021).
[34] 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).
[35] J. Zhang, H. Lin and 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).