A Theoretical and Empirical Comparison of Gradient Approximations in Derivative-Free Optimization

A. S. Berahas* L. Cao* K. Choromanski† K. Scheinberg‡§

April 1, 2020

Abstract

In this paper, we analyze several methods for approximating gradients of noisy functions using only function values. These methods include finite differences, linear interpolation, Gaussian smoothing and smoothing on a unit sphere. The methods differ in the number of functions sampled, the choice of the sample points, and the way in which the gradient approximations are derived. For each method, we derive bounds on the number of samples and the sampling radius which guarantee favorable convergence properties for a line search or fixed step size descent method. To this end, we use the results in [3] and show how each method can satisfy the sufficient conditions. We analyze the convergence properties even when this condition is only satisfied with some sufficiently large probability at each iteration, as happens to be the case with Gaussian smoothing and smoothing on a unit sphere. Finally, we present numerical results evaluating the quality of the gradient approximations as well as their performance in conjunction with a line search derivative-free optimization algorithm.

1 Introduction

We consider an unconstrained optimization problem of the form

$$\min_{x\in\mathbb{R}^n} \phi(x),$$

where $f(x) = \phi(x) + \epsilon(x)$ is computable, while $\phi(x)$ may not be. In other words $f : \mathbb{R}^n \to \mathbb{R}$ is a possibly noisy approximation of a smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$, and the goal is to minimize ϕ. Throughout the paper we will assume that the noise is bounded uniformly, i.e., there exists a constant $\epsilon_f \geq 0$ such that $|\epsilon(x)| \leq \epsilon_f$ for all $x \in \mathbb{R}^n$. While this may seem a strong assumption, it is often satisfied in practice when $f(x)$ is the result of a computer code aimed at computing $\phi(x)$, but that has inaccuracies due to internal discretization [20, 21]. Another common setting in which the assumption is satisfied is when $f(x)$ is a nonsmooth function and $\phi(x)$ is its smooth approximation; see e.g., [19, 22].

In this paper, we do not assume that any information on $\nabla\phi(x)$ is available. Such problems arise in many fields such as Derivative-Free Optimization (DFO) [13, 17], Simulation Optimization [25, 33] and Machine Learning [8, 15, 18, 32]. There have been a number of works analyzing the case when $\epsilon(x)$ is a random function with zero mean (not necessarily bounded). The results obtained for stochastic noise and the corresponding optimization methods are different than those for bounded arbitrary noise.

One common approach to optimizing functions without derivatives is to compute an estimate of the gradient $\nabla\phi(x)$ at the point x, denoted by $g(x)$, using (noisy) function values and then apply a gradient based

*Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA; E-mails: albertberahas@gmail.com, lic314@lehigh.edu

†Google Brain, New York, NY, USA; Email: kchoro@google.com

‡Department of Operations Research and Information Engineering, Cornell University, Bethlehem, PA, USA; E-mail: katyas@cornell.edu

§Corresponding author.

method with $g(x)$. The most straightforward way to estimate $\nabla\phi(x)$ is to use forward finite differences by sampling one point near x along each of the n coordinates. Alternatively, one can estimate $\nabla\phi(x)$ via central finite differences where two points are sampled along each coordinate in both directions. As a generalization of the finite difference approach, $g(x)$ can be computed via linear interpolation. This approach also requires n sample points near x, however, the location of the sample points can be chosen arbitrarily, as long as they form a set of n linearly independent directions from x. Linear interpolation is very useful when coupled with an optimization algorithm that (potentially) reuses some of the sample function values computed at prior iterations, thus avoiding the need to compute $n+1$ new function values at each iteration. The accuracy of the resulting gradient approximation depends on the conditioning of the matrix $Q_{\mathcal{X}}$, which is the matrix whose rows are the linearly independent directions formed by the sample points. An extensive study of optimization methods based on interpolation gradients can be found in [13].

An alternative approach for estimating gradients using an arbitrary number of function value samples is based on random sample points. The essence of these methods is to compute gradient estimates as a sum of estimates of directional derivatives along random (e.g., Gaussian) directions. Such approaches have become popular in recent literature for policy optimization in reinforcement learning (RL) [8, 9, 15, 29, 30, 32] as a particular case of simulation optimization. For example, the authors of [30] used a technique arising from Gaussian smoothing (see e.g., [19, 22]), which they referred to as *evolutionary strategies*. We will not use this terminology in this paper, as it encompasses a large and different class of optimization methods than those used in [30] and those considered in this paper.

Overall, the methods we consider in this paper compute an estimate of $\nabla\phi(x)$ (denoted by $g(x)$), as follows

$$g(x) = \sum_{i=1}^{N} \frac{f(x+\sigma u_i) - f(x)}{\sigma} \tilde{u}_i, \tag{1.1}$$

or using the central (symmetric or antithetic) version

$$g(x) = \sum_{i=1}^{N} \frac{f(x+\sigma u_i) - f(x-\sigma u_i)}{2\sigma} \tilde{u}_i, \tag{1.2}$$

where $\{u_i : i = 1, \dots, N\}$[1] is a *set of directions* that depend on the method, $\tilde{u}_i$ depends on u_i, and σ is the *sampling radius*. In particular, for the finite difference methods $N = n$ and $u_i = \tilde{u}_i = e_i$, where e_i denotes the i-th column of the identity matrix. For interpolation, $N = n$, $\{u_i : i = 1, \dots, N\}$ is a set of arbitrary linearly independent vectors with $\|u_i\| \leq 1$[2] for all i, and $\tilde{u}_i$ are the columns of $Q_{\mathcal{X}}^{-1}$, where ith row of $Q_{\mathcal{X}} \in \mathbb{R}^{N\times n}$ is u_i. A special case of linear interpolation has been explored in [8, 9, 29] where the u_i's are random orthogonal directions; this approach can also be viewed as rotated finite differences. In the case of Gaussian smoothing, the directions u_i are random directions from a standard Gaussian distribution and $\tilde{u}_i = \frac{1}{N} u_i$. Finally, a variant of this method that selects the directions u_i from a uniform distribution on a unit sphere, where $\tilde{u}_i = \frac{n}{N} u_i$, has been explored in [15, 16]. As is clear from (1.2), antithetic gradient approximations require $2N$ function evaluations. Details about these methods are given in Section 2.

We are motivated by recent empirical use of these methods in the RL literature. In [30], the authors showed that the Gaussian smoothing approach is an efficient way to compute gradient estimates when $N \sim n$. In follow-up works [8, 9, 29] it was shown empirically that better gradient estimates can be obtained for the same optimization problems by using interpolation with orthogonal directions. While the numerical results in these works confirmed the feasibility of use of (1.1) and (1.2) for various RL benchmark sets and different choice of directions, there is no theoretical analysis, neither comparing the accuracy of resulting gradient estimates, nor analyzing the connection between such accuracy and downstream optimization gains. To the best of our best knowledge, there has been no systematic analysis of the accuracy of the (stochastic)

[1] Throughout the paper, N denotes the *size of the sample set* $\{u_i : i = 1, \dots, N\}$. Note that for the central versions of the gradient approximations, the number of sampled functions is equal to $2N$.

[2] The norms used in this paper are Euclidean norms.

gradient estimates used in the DFO literature (such as Gaussian smoothing and smoothing on a unit sphere) specifically in conjunction with requirements of obtaining descent directions.

In this paper, we develop theoretical bounds on the gradient approximation errors $\|g(x) - \nabla\phi(x)\|$ for all aforementioned gradient estimation methods and show their dependence on the number of samples. Another key quantity we consider is the radius of sampling σ. In the absence of noise, σ can be chosen arbitrarily small, however, when noise is present, small values of σ can lead to large inaccuracies in the gradient estimates. We derive the range for σ which ensures that the gradient estimates are sufficiently accurate and thus can be used in conjunction with efficient gradient based methods.

So far, most of the papers that employ smoothing techniques for gradient approximations do so within a simple gradient descent scheme with a *fixed* step size parameter; see e.g., [15, 16, 22, 30]. It is well established that line search methods, where the step sizes employed are adaptive, are often much more efficient than fixed step size methods. However, for line search methods to be effective the search directions employed should be descent directions, at least with high enough probability, which implies that the gradient estimates have to be sufficiently accurate (i.e., have low variance). Moreover, it has been observed that quasi-Newton methods, while extremely useful in accelerating optimization algorithms, are also only effective with sufficiently accurate gradient estimates [2, 37]. The results of this paper can be used in conjunction with the results from [3] to establish convergence rates for a line search algorithm where the gradient is estimated with any one of the method mentioned above.

In addition to the aforementioned main results, we also provide a rigorous quantitative analysis which indicates that, when using the same number of samples, methods employing orthogonal directions produce significantly better estimates of the gradient (smaller error) than those employing (1.1) and (1.2) with Gaussian directions. In fact, orthogonal directions, as well as finite difference gradient estimates, are special cases of linear interpolation. With both theoretical and empirical results, we argue that while smoothing methods (Gaussian or unit sphere), can be applied with $N << n$, the resulting estimates generally have lower accuracy that those computed via linear interpolation. We also show that using directions uniformly distributed on a sphere has some theoretical advantages over using Gaussian directions, although its practical behavior is very similar to the latter method and both of these methods are inferior to using orthogonal directions or linear interpolation.

Organization The paper is organized as follows. In the remainder of this section, we introduce the assumptions we make for our analysis, and then present the main results of the paper. We define and derive theoretical results for the gradient approximation methods in Section 2. We present a numerical comparison of the gradient approximations and illustrate the performance of a line search DFO algorithm that employs these gradient approximations in Section 3. Finally, in Section 4, we make some concluding remarks and discuss avenues for future research.

1.1 Assumptions

Throughout the paper we will assume that the noise in the function evaluations $\epsilon(x)$ is bounded for all $x \in \mathbb{R}^n$, and that ϕ is Lipschitz smooth.

Assumption 1.1. ***(Boundedness of Noise in the Function)*** *There is a constant $\epsilon_f \geq 0$ such that $|f(x) - \phi(x)| = |\epsilon(x)| \leq \epsilon_f$ for all $x \in \mathbb{R}^n$.*

Assumption 1.2. ***(Lipschitz continuity of the gradients of ϕ)*** *The function ϕ is continuously differentiable, and the gradient of ϕ is L-Lipschitz continuous for all $x \in \mathbb{R}^n$.*

In some cases, to establish better approximations of the gradient, we will assume that ϕ has Lipschitz continuous Hessians.

Assumption 1.3. ***(Lipschitz continuity of the Hessian of ϕ)*** *The function ϕ is twice continuously differentiable, and the Hessian of ϕ is M-Lipschitz continuous for all $x \in \mathbb{R}^n$.*

1.2 Summary of Results

We begin by stating a condition that is often used in the analysis of first order methods with inexact gradient computations:

$$\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|, \tag{1.3}$$

for some $\theta \in [0, 1)$. This condition is referred to as the *norm condition* and was introduced and studied in [6] in the context of trust-region methods with inexact gradients. Note, this condition implies that $g(x)$ is a descent direction for the function ϕ. In [3] the authors establish expected complexity bounds for a generic line search algorithm that uses gradient approximations in lieu of the true gradient, under the condition that the gradient estimate $g(x)$ satisfies (1.3) for sufficiently small θ and with sufficiently high probability.

Clearly, unless we know $\|\nabla\phi(x)\|$, condition (1.3) may be hard or impossible to verify or guarantee. There is significant amount of work that attempts to circumvent this difficulty; see e.g., [5, 7, 24]. In [5] a practical approach to estimate $\|\nabla\phi(x_k)\|$ is proposed and used to ensure some approximation of (1.3) holds. In [7, 24] the condition (1.3) is replaced by

$$\|g(x) - \nabla\phi(x)\| \leq \kappa\alpha_k\|g(x)\|,$$

for some $\kappa > 0$, and convergence rate analyses are derived for a line search method that has access to deterministic function values in [7] and stochastic function values (with additional assumptions) in [24]. However, for the methods studied in this paper, condition (1.3) turns out to be achievable. Therefore, in this paper, we establish conditions under which (1.3) holds either deterministically or with sufficiently high probability. Given a point x, all of these methods compute $g(x)$ using either (1.1) or (1.2), where the u_i's are the *sampling directions* and σ is the *sampling radius*. The methods vary in their selection of the *size of the sample set* N, the set $\{u_i : i = 1, \ldots, N\}$ and the corresponding set $\{\tilde{u}_i : i = 1, \ldots, N\}$, and the *sampling radius* σ.

Here, upfront, we present a simplified summary of the conditions on N, σ and $\nabla\phi(x)$ for each method that we consider in this paper to guarantee condition (1.3); see Table 1. For more detailed results see Section 2.6, Tables 2 and 3. Note that for the smoothing methods the condition (1.3) holds with probability $1 - \delta$ and the number of samples depends on δ. Moreover, the bounds on N for the smoothing methods are a simplification of the more detailed bounds derived in the paper and apply when $n \geq 4$, while for smaller n some of the constants are larger. We should note that the constants in the bound on N are somewhat smaller for larger n.

From Table 1 we note that the bounds on N for all methods have the same dependence (order of magnitude) on the dimension n; however, for the smoothing methods the constants in the bound are significantly larger than those for deterministic methods, such as finite differences. This suggests that deterministic methods may be more efficient, at least in the setting considered in this paper, when accurate gradient estimates are desired. The bounds on the sampling radius are comparable for the smoothing and deterministic methods, as we will discuss in detail later in the paper. Our numerical results support our theoretical observations.

2 Gradient Approximations and Sampling

In this section, we analyze several existing methods for constructing gradient approximations using only noisy function information. We establish conditions under which the gradient approximations constructed via these methods satisfy the bound (1.3) for any given $\theta \in [0, 1)$.

The common feature amongst these methods is that they construct approximations $g(x)$ of the gradient $\nabla\phi(x)$ using (possibly noisy) function values $f(y)$ for $y \in \mathcal{X}$, where $\mathcal{X}$ is a *sample set* centered around x. These methods differ in the way they select $\mathcal{X}$ and the manner in which the function values $f(y)$, on all sample points $y \in \mathcal{X}$, are used to construct $g(x)$. The methods have different costs in terms of number of evaluations of f, as well as other associated computations. Our goal is to compare these costs when computing gradient

Table 1: Simplified conditions under assumption that $n \geq 4$. Bounds on N, σ and $\nabla\phi(x)$ that ensure $\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|$ (* denotes result is with probability $1-\delta$).

Gradient Approximation	N	σ	$\|\nabla\phi(x)\|$
Forward Finite Differences	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{\sqrt{n}L}\right]$	$\frac{2\sqrt{nL\epsilon_f}}{\theta}$
Central Finite Differences	n	$\left[\sqrt[3]{\frac{6\epsilon_f}{M}}, \sqrt{\frac{3\theta\|\nabla\phi(x)\|}{\sqrt{n}M}}\right]$	$\frac{2\sqrt{n}\sqrt[3]{M\epsilon_f^2}}{\sqrt[3]{6}\theta}$
Linear Interpolation	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{\|Q_{\mathcal{X}}^{-1}\|\sqrt{n}L}\right]$	$\frac{2\|Q_{\mathcal{X}}^{-1}\|\sqrt{nL\epsilon_f}}{\theta}$
Gaussian Smoothed Gradients*	$\frac{12n}{\delta\theta^2} + \frac{(n+20)}{16\delta}$	$\left[\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{6nL}\right]$	$\frac{6n\sqrt{L\epsilon_f}}{\theta}$
Centered Gaussian Smoothed Gradients*	$\frac{12n}{\delta\theta^2} + \frac{(n+30)}{144\delta}$	$\left[\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}}, \sqrt{\frac{\theta\|\nabla\phi(x)\|}{12n^{3/2}M}}\right]$	$\frac{12\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta}$
Sphere Smoothed Gradients*	$\left[\frac{8n}{\theta^2} + \frac{8n}{3\theta} + \frac{(11n+104)}{24}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt{\frac{n\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}L}\right]$	$\frac{4n\sqrt{L\epsilon_f}}{\theta}$
Centered Sphere Smoothed Gradients*	$\left[\frac{8n}{\theta^2} + \frac{8n}{3\theta} + \frac{(9n+192)}{27}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt[3]{\frac{n\epsilon_f}{M}}, \sqrt{\frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}M}}\right]$	$\frac{4\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta}$

estimates that satisfy (1.3) for some $\theta \in [0, 1)$. For each method, we derive bounds on the number of samples and the sampling radius which guarantee (1.3), the sufficient condition for convergence of a line search or fixed step size descent method [3].

2.1 Gradient Estimation via Standard Finite Differences

The first method we analyze is the standard finite difference method. The forward finite difference (FFD) approximation to the gradient of ϕ at $x \in \mathbb{R}^n$ is computed using the sample set $\mathcal{X} = \{x + \sigma e_i\}_{i=1}^n \cup \{x\}$, where $\sigma > 0$ is the finite difference interval and $e_i \in \mathbb{R}^n$ is the i-th column of the identity matrix, as follows

$$[g(x)]_i = \frac{f(x + \sigma e_i) - f(x)}{\sigma}, \quad \text{for} \quad i = 1, \ldots, n.$$

Alternatively, gradient approximations can be computed using central finite differences (CFD) based on the sample set $\mathcal{X} = \{x + \sigma e_i\}_{i=1}^n \cup \{x - \sigma e_i\}_{i=1}^n$, as

$$[g(x)]_i = \frac{f(x + \sigma e_i) - f(x - \sigma e_i)}{2\sigma}, \quad \text{for} \quad i = 1, \ldots, n.$$

FFD and CFD approximations require n and $2n$ functions evaluations, respectively. CFD approximations tends to be more accurate and stable, as we show below.

We begin by stating two standard gradient approximation bounds, i.e., the error between the finite difference approximation to the gradient and the gradient of ϕ.

Theorem 2.1. *Under Assumptions 1.1 and 1.2, let $g(x)$ denote the forward finite difference (FFD) approximation to the gradient $\nabla\phi(x)$. Then, for all $x \in \mathbb{R}^n$,*

$$\|g(x) - \nabla\phi(x)\| \leq \frac{\sqrt{n}L\sigma}{2} + \frac{2\sqrt{n}\epsilon_f}{\sigma}.$$

Theorem 2.2. *Under Assumptions 1.1 and 1.3, let $g(x)$ denote the central finite difference (CFD) approximation to the gradient $\nabla\phi(x)$. Then, for all $x \in \mathbb{R}^n$,*

$$\|g(x) - \nabla\phi(x)\| \leq \frac{\sqrt{n}M\sigma^2}{6} + \frac{\sqrt{n}\epsilon_f}{\sigma}.$$

It is apparent from Theorems 2.1 and 2.2 that the finite difference interval $\sigma > 0$ should be chosen to be not too small and not too large in order to ensure $\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|$. We use the results in these theorems to derive upper and low bounds on σ.

First we consider Theorem 2.1. In order to ensure that (1.3) holds, we require

$$\frac{\sqrt{n}L\sigma}{2} + \frac{2\sqrt{n}\epsilon_f}{\sigma} \leq \theta\|\nabla\phi(x)\|, \tag{2.1}$$

which can be written as a quadratic inequality,

$$\frac{\sqrt{n}L}{2}\sigma^2 - \theta\|\nabla\phi(x)\|\sigma + 2\sqrt{n}\epsilon_f \leq 0.$$

In Theorem 2.1 we can assume without loss of generality that $L > 0$, then the interval (in terms of σ) of the the quadratic inequality is

$$\frac{\theta\|\nabla\phi(x)\| - \sqrt{\theta^2\|\nabla\phi(x)\|^2 - 4nL\epsilon_f}}{\sqrt{n}L} \leq \sigma \leq \frac{\theta\|\nabla\phi(x)\| + \sqrt{\theta^2\|\nabla\phi(x)\|^2 - 4nL\epsilon_f}}{\sqrt{n}L}. \tag{2.2}$$

This interval is nonempty when $\theta^2\|\nabla\phi(x)\|^2 \geq 4nL\epsilon_f$, which constitutes a condition on the best achievable $\|\nabla\phi(x)\|$ in terms of the level of the noise ϵ_f. When $\|\nabla\phi(x)\|$ falls below $\frac{2\sqrt{nL\epsilon_f}}{\theta}$, finite difference approximations to the gradient can no longer deliver sufficiently accurate approximations $g(x)$, and any optimization process reliant on this approximation will fail to progress.

The bound in (2.2) defines precise interval on σ which gives $g(x)$ that satisfies (2.1). However, at the expense of somewhat narrowing the range of allowable σ values, we derive here much simpler bounds. We note that (2.1) is equivalent to

$$\sigma\left(\theta\|\nabla\phi(x)\| - \frac{\sqrt{n}L\sigma}{2}\right) \geq 2\sqrt{n}\epsilon_f.$$

Let us assume that

$$\|\nabla\phi(x)\| \geq \frac{\sqrt{n}L\sigma}{\theta}, \tag{2.3}$$

then,

$$\theta\|\nabla\phi(x)\| - \frac{\sqrt{n}L\sigma}{2} \geq \frac{\sqrt{n}L\sigma}{2}.$$

Hence, if we ensure that

$$\frac{\sqrt{n}L\sigma^2}{2} \geq 2\sqrt{n}\epsilon_f,$$

then (2.1) holds, under assumption (2.3). Thus we have the following bounds that connect σ, ϕ and ϵ_f:

$$2\sqrt{\frac{\epsilon_f}{L}} \leq \sigma \leq \frac{\theta\|\nabla\phi(x)\|}{\sqrt{n}L}, \quad \|\nabla\phi(x)\| \geq \frac{2\sqrt{nL\epsilon_f}}{\theta}.$$

We note that these bounds can be imposed in an algorithm in the following way: given an upper bound estimate of the noise ϵ_f and an upper bound on the Lipschitz constant of $\nabla\phi(x)$, we set σ equal to $2\sqrt{\frac{\epsilon_f}{L}}$ for any algorithm that uses FFD to compute gradient approximations. Such an algorithm will be able to converge to a neighborhood where $\|\nabla\phi(x)\| \approx \frac{\sqrt{nL\epsilon_f}}{\theta}$. In [3] it is shown that a line search algorithm for noisy objective functions, based on gradient approximations that satisfy (1.3) enjoy fast convergence rates until it reaches a neighborhood of optimality dictated by ϵ_f.

Applying the same logic as above to Theorem 2.2, in order to ensure that (1.3) holds, we require

$$\frac{\sqrt{n}M\sigma^2}{6} + \frac{\sqrt{n}\epsilon_f}{\sigma} \leq \theta\|\nabla\phi(x)\|, \tag{2.4}$$

which is equivalent to

$$\sigma\left(\theta\|\nabla\phi(x)\| - \frac{\sqrt{n}M\sigma^2}{6}\right) \geq \sqrt{n}\epsilon_f.$$

Let us assume that

$$\|\nabla\phi(x)\| \geq \frac{\sqrt{n}M\sigma^2}{3\theta}, \tag{2.5}$$

then,

$$\theta\|\nabla\phi(x)\| - \frac{\sqrt{n}M\sigma^2}{6} \geq \frac{\sqrt{n}M\sigma^2}{6}.$$

Hence if we ensure that

$$\frac{\sqrt{n}M\sigma^3}{6} \geq \sqrt{n}\epsilon_f,$$

then (2.4) holds, under assumption (2.5). Thus we have the following bounds that connect σ, ϕ and ϵ_f:

$$\sqrt[3]{\frac{6\epsilon_f}{M}} \leq \sigma \leq \sqrt{\frac{3\theta\|\nabla\phi(x)\|}{\sqrt{n}M}}, \quad \|\nabla\phi(x)\| \geq \frac{2}{\sqrt[3]{6}}\frac{\sqrt{n}\sqrt[3]{M\epsilon_f^2}}{\theta}.$$

2.2 Gradient Estimation via Linear Interpolation

We now consider a more general method for approximating gradients using polynomial interpolation that has become a popular choice for model based trust region methods in the DFO setting [10, 11, 13, 19, 27, 28, 36]. These methods construct surrogate models of the objective function using interpolation (or regression). While typically, in the DFO setting, interpolation is used to construct quadratic models of the objective function around $x \in \mathbb{R}^n$ of the form

$$m(y) = f(x) + g(x)^\intercal(y-x) + \frac{1}{2}(y-x)^\intercal H(x)(y-x), \tag{2.6}$$

where $f \in \mathbb{R}$ and $g \in \mathbb{R}^n$, or $H \in \mathbb{R}^{n\times n}$, in this paper we focus on the simplest case of linear models,

$$m(y) = f(x) + g(x)^\intercal(y-x), \tag{2.7}$$

as the focus of this paper is on line search methods, whereas the use of (2.6) requires a trust region approach due to the general nonconvexity of $m(y)$ [13].

Let us consider the following sample set $\mathcal{X} = \{x+\sigma u_1, x+\sigma u_2, \ldots, x+\sigma u_n\}$ for some $\sigma > 0$. In other words, we have n directions denoted by $u_i \in \mathbb{R}^n$ and we sample f along those directions, around x, using a sampling radius of size σ. We assume $f(x)$ is known (function value at x). Let $F_\mathcal{X} \in \mathbb{R}^n$ be a vector whose entries are $f(x+\sigma u_i) - f(x)$, for $i = 1\ldots n$, and let $Q_\mathcal{X} \in \mathbb{R}^{n\times n}$ define a matrix whose rows are given by u_i for $i = 1\ldots n$. The model in (2.7) is constructed to satisfy the interpolation conditions,

$$f(x+\sigma u_i) = m(x+\sigma u_i), \quad \forall i = 1,\ldots,n,$$

which can be written as

$$\sigma Q_\mathcal{X} g = F_\mathcal{X}. \tag{2.8}$$

If the matrix $Q_\mathcal{X}$ is nonsingular, then $m(y) = f(x) + g(x)^\intercal(y-x)$, with $g(x) = \frac{1}{\sigma}Q_\mathcal{X}^{-1}F_\mathcal{X}$, is a linear interpolation model of $f(y)$ on the sample set $\mathcal{X}$. When $Q_\mathcal{X}$ is the identity matrix, then we recover standard

forward finite difference gradient estimation. In the specific case when $Q_{\mathcal{X}}$ is orthonormal, then $Q_{\mathcal{X}}^{-1} = Q_{\mathcal{X}}^{\intercal}$, thus $g(x)$ is written as

$$g(x) = \sum_{i=1}^{n} \frac{f(x+\sigma u_i) - f(x)}{\sigma} u_i.$$

Next we derive a bound on $\|g(x) - \nabla\phi(x)\|$. This result is an extension of the results presented in [12, 13] to account of the noise in the function evaluations.

Theorem 2.3. *Suppose that Assumptions 1.1 and 1.2 hold. Let $\mathcal{X} = \{x+\sigma u_1, \ldots, x+\sigma u_n\}$ be a set of interpolation points such that $\max_{1\le i\le n} \|u_i\| \le 1$ and $Q_{\mathcal{X}}$ be nonsingular. Then, for all $x \in \mathbb{R}^n$,*

$$\|g(x) - \nabla\phi(x)\| \le \frac{\|Q_{\mathcal{X}}^{-1}\|_2 \sqrt{n} L\sigma}{2} + \frac{2\|Q_{\mathcal{X}}^{-1}\|_2 \sqrt{n}\epsilon_f}{\sigma}.$$

Proof. From the interpolation conditions and the mean value theorem, $\forall i = 1, \ldots, n$ we have

$$\begin{aligned}\sigma g(x)^{\intercal} u_i &= f(x+\sigma u_i) - f(x) = \phi(x+\sigma u_i) - \phi(x) + \epsilon(x+\sigma u_i) - \epsilon(x) \\ &= \int_0^1 \sigma u_i^{\intercal} \nabla\phi(x + t\sigma u_i) dt + \epsilon(x+\sigma u_i) - \epsilon(x).\end{aligned}$$

From the L-smoothness of $\phi(\cdot)$ and the bound on $\epsilon(\cdot)$ we have

$$\sigma |(g(x) - \nabla\phi(x))^{\intercal} u_i| \le \frac{L\sigma^2 \|u_i\|^2}{2} + 2\epsilon_f, \quad \forall i = 1, \ldots, n$$

which in turn implies

$$\|Q_{\mathcal{X}}(g(x) - \nabla\phi(x))\| \le \frac{\sqrt{n} L\sigma}{2} + \frac{2\sqrt{n}\epsilon_f}{\sigma},$$

and the theorem statement follows. □

This result has the implication that large $\|Q_{\mathcal{X}}^{-1}\|$ can cause large deviation of $g(x)$ from $\nabla\phi(x)$. Thus, it is desirable to select $\mathcal{X}$ in such a way that the condition number of $Q_{\mathcal{X}}^{-1}$ is small, which is clearly optimized when $Q_{\mathcal{X}}$ is orthonormal. Thus, we trivially recover the theorem for FFD, and moreover, extend this result to any orthonormal set of directions $\{u_1, u_2 \ldots, u_n\}$, such as those used in [9].

Aside from the condition number, the important difference between general interpolation sets and orthonormal ones is in the computational cost of evaluating $g(x)$. In particular, $g(x)$ is obtained by solving a system of linear equations given by (2.8), which in general requires $\mathcal{O}(n^3)$ computations, but that reduces to $\mathcal{O}(n^2)$ in the case of general orthornormal matrices $Q_{\mathcal{X}}$, and further reduces to $\mathcal{O}(n)$ for $Q_{\mathcal{X}} = I$, as in the case of FFD. In [9], it is proposed to use scaled randomized Haddamard matrices as $Q_{\mathcal{X}}$. This is only possible if the problem dimension is a power of 2, but it reduces linear algebra cost of matrix-vector products from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$.

On the other hand, using general sample sets allows for greater flexibility (within an optimization algorithm), in particular enabling the re-use of sample points from prior iterations. When using FFD to compute $g(x)$, n function evaluations are always required, while when using interpolation it is possible to update the interpolation set by replacing only one (or a few) sample point(s) in the set $\mathcal{X}$. It is important to note that while $\mathcal{X}$ can be fairly general, the condition number of the matrix $Q_{\mathcal{X}}$ has to remain bounded for Theorem 2.3 to be useful. The sets with bounded condition number of $Q_{\mathcal{X}}$ are called *well-poised*; see [13] for details about the construction and maintenance of interpolation sets in model based trust region DFO methods.

The bounds of Theorem 2.3 are similar to those of Theorem 2.1, hence, if the sampling radius σ is chosen to satisfy

$$2\sqrt{\frac{\epsilon_f}{L}} \le \sigma \le \frac{\theta\|\nabla\phi(x)\|}{\|Q_{\mathcal{X}}^{-1}\|\sqrt{n}L}, \qquad \text{with} \qquad \|\nabla\phi(x)\| \ge \frac{2\|Q_{\mathcal{X}}^{-1}\|\sqrt{nL\epsilon_f}}{\theta},$$

then (1.3) holds. Since the condition number of $\|Q_{\mathcal{X}}^{-1}\|$ may not be known when σ is chosen, an upper bound on $\|Q_{\mathcal{X}}^{-1}\|$, which needs to be imposed for convergence, can used instead; see [13].

It is possible to derive an analogue of Theorem 2.2 by including n additional sample points $\{x - \sigma u_1, \ldots, x - \sigma u_n\}$ in the gradient estimation procedure. Namely, two sample sets are used, $\mathcal{X}^+ = \{x + \sigma u_1, x + \sigma u_2, \ldots x + \sigma u_n\}$ and $\mathcal{X}^- = \{x - \sigma u_1, x - \sigma u_2, \ldots x - \sigma u_n\}$, with corresponding matrices $Q_{\mathcal{X}^+}$ and $Q_{\mathcal{X}^-}$. The linear model $m(y) = f(x) + g^\intercal(y - x)$ is then computed as an average of the two interpolation models, that is

$$g = \frac{g_0^+ + g_0^-}{2} = \frac{1}{2\sigma}[Q_{\mathcal{X}^+}^{-1} F_{\mathcal{X}^+} + Q_{\mathcal{X}^-}^{-1} F_{\mathcal{X}^-}].$$

The gradient estimates are computed in this way in [8], for the case of orthonormal sets and symmetric finite difference computations. Similarly to the CFD, this results in better accuracy bounds in terms of σ; however, this requires additional n function evaluations at each iteration, which contradicts the original idea of using interpolation as a means for reducing the per-iteration function evaluation cost.

2.3 Gradient Estimation via Gaussian Smoothing

Gaussian smoothing has recently become a popular tool for building gradient approximations using only function values. This approach has been exploited in several recent papers; see e.g., [19, 22, 30, 35].

Gaussian smoothing of a given function f is obtained as follows:

$$\begin{aligned} F(x) &= \mathbb{E}_{y \sim \mathcal{N}(x, \sigma^2 I)}[f(y)] = \int_{\mathbb{R}^n} f(y)\pi(y|x, \sigma^2 I)dy \\ &= \mathbb{E}_{u \sim \mathcal{N}(0, I)}[f(x + \sigma u)] = \int_{\mathbb{R}^n} f(x + \sigma u)\pi(u|0, I)du, \end{aligned} \tag{2.9}$$

where $\mathcal{N}(x, \sigma^2 I)$ denotes the multivariate normal distribution with mean x and covariance matrix $\sigma^2 I$, $\mathcal{N}(0, I)$ denotes the standard multivariate normal distribution, and the functions $\pi(y|x, \sigma^2 I)$ and $\pi(u|0, I)$ denote the probability density functions (pdf) of $\mathcal{N}(x, \sigma^2 I)$ evaluated at y and $\mathcal{N}(0, I)$ evaluated at u, respectively. The gradient of F can be expressed as

$$\nabla F(x) = \frac{1}{\sigma}\mathbb{E}_{u \sim \mathcal{N}(0, I)}[f(x + \sigma u)u]. \tag{2.10}$$

Assume f is an approximation of ϕ with the approximation error bounded by ϵ_f uniformly, i.e., Assumption 1.1 holds. If Assumption 1.1 holds, then the following bounds hold for the error between $\nabla F(x)$ and $\nabla \phi(x)$. If ϕ has L-Lipschitz continuous gradients, that is if Assumption 1.2 holds, then

$$\|\nabla F(x) - \nabla \phi(x)\| \leq \sqrt{n} L\sigma + \frac{\sqrt{n}\epsilon_f}{\sigma}; \tag{2.11}$$

see Appendix A.1 for the proof[3]. If the function ϕ has M-Lipschitz continuous Hessians, that is if Assumption 1.3 holds, then

$$\|\nabla F(x) - \nabla \phi(x)\| \leq nM\sigma^2 + \frac{\sqrt{n}\epsilon_f}{\sigma}; \tag{2.12}$$

see Appendix A.2 for proof.

[3]The bound (2.11) was presented in [19] without proof; we would like to thank the first author of [19] for providing us with guidance of this proof.

In order to approximate $\nabla\phi(x)$ one can approximate $\nabla F(x)$, with sufficient accuracy, by sample average approximations (SAA) applied to (2.10), i.e.,

$$g(x) = \frac{1}{N\sigma}\sum_{i=1}^{N} f(x+\sigma u_i)u_i, \tag{2.13}$$

where $u_i \sim \mathcal{N}(0, I)$ for $i = 1, 2, \ldots, N$. It can be easily verified that $g(x)$ computed via (2.13) has large variance (the variance explodes as σ goes to 0). The following simple modification,

$$g(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{f(x+\sigma u_i) - f(x)}{\sigma} u_i, \tag{2.14}$$

eliminates this problem and is indeed used in practice instead of (2.13); see [8, 9, 30]. Note that the expectation of (2.14) is also $\nabla F(x)$, since $\mathbb{E}_{u_i\sim\mathcal{N}(\mathbf{0},I)}[f(x)u]$ is an all-zero vector for all i. In what follows we will refer to $g(x)$ computed via (2.14) as the Gaussian smoothed gradient (GSG). As pointed out in [22], $\frac{f(x+\sigma u_i)-f(x)}{\sigma}u_i$ can be interpreted as a forward finite difference version of the directional derivative of f at x along u_i. Moreover, one can also consider the central difference variant of (2.14)–central Gaussian smoothed gradient (cGSG)–which is computed as follows,

$$g(x) = \frac{1}{2N}\sum_{i=1}^{N} \frac{f(x+\sigma u_i) - f(x-\sigma u_i)}{\sigma} u_i. \tag{2.15}$$

The properties of (2.9) and (2.14), with $N = 1$, were analyzed in [22]. However, this analysis does not explore the effect of $N > 1$ on the variance of $g(x)$. On the other hand, in [30] the authors propose an algorithm that uses GSG estimates, (2.14) and (2.15), with large samples sizes N in a fixed step size gradient descent algorithm, but without any analysis or discussion of the choices of N, σ or α (where α is the step size). Thus, the purpose of this section is to derive bounds on the approximation error $\|g(x) - \nabla\phi(x)\|$ for GSG and cGSG, and to derive conditions on σ and N under which condition (1.3) holds (and as a result the convergence results for a line search DFO algorithm [3] based on these approximations also hold).

We first note that there are two sources of error: (i) approximation of the true function ϕ by the Gaussian smoothed function F of the noisy function f, and (ii) approximation of $\nabla F(x)$ via sample average approximations. Hence, we have that

$$\begin{aligned}\|g(x) - \nabla\phi(x)\| &= \|(\nabla F(x) - \nabla\phi(x)) + (g(x) - \nabla F(x))\| \\ &\leq \|\nabla F(x) - \nabla\phi(x)\| + \|g(x) - \nabla F(x)\|.\end{aligned} \tag{2.16}$$

The bound on the first term is given by (2.11) or (2.12). What remains is to bound the second term $\|g(x) - \nabla F(x)\|$, the error due to the sample average approximation.

Since (2.14) (and (2.15)) is a (mini-)batch stochastic gradient estimate of $\nabla F(x)$, the probabilistic bound on $\|g(x) - \nabla F(x)\|$ is derived by bounding the expectation, which is equivalent to bounding the variance of the (mini-)batch stochastic gradient. Existing bounds in the literature, see e.g., [34], are derived under the assumption that $\|g(x) - \nabla\phi(x)\|$ is uniformly bounded above almost surely, which does not hold for GSG because when u follows a Gaussian distribution, $\frac{f(x+\sigma u)-f(x)}{\sigma}u$ can be arbitrarily large with positive probability. Here, we bound $\|g(x) - \nabla F(x)\|$ only under Assumptions 1.2 or 1.3. It is shown in [22] that Assumption 1.2 implies that $\nabla F(x)$ is L-Lipschitz, and that Assumption 1.2 implies that $\nabla^2 F(x)$ is M-Lipschitz.

The variance for (2.14) can be expressed as

$$\text{Var}\left\{g(x)\right\} = \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{f(x+\sigma u) - f(x)}{\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal, \tag{2.17}$$

and the variance of (2.15) can be expressed as

$$\operatorname{Var}\{g(x)\} = \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{f(x+\sigma u)-f(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal. \tag{2.18}$$

The following properties of a normally distributed multivariate random variable $u \in \mathbb{R}^n$ will be used in our analysis and are derived in Appendix A.3. Let a be any vector in $\mathbb{R}^n$ independent from u, then

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[(a^\intercal u)^2 uu^\intercal\right] = a^\intercal a I + 2aa^\intercal \\
&\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[a^\intercal u\cdot\|u\|^k\cdot uu^\intercal\right] = 0_{n\times n} \text{ for } k = 0,1,2,... \\
&\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k uu^\intercal\right] \begin{cases} = (n+2)(n+4)\cdots(n+k)I & \text{for } k = 0,2,4,\dots \\ \preceq (n+1)(n+3)\cdots(n+k)\cdot n^{-0.5}I & \text{for } k = 1,3,5,\dots \end{cases}
\end{aligned} \tag{2.19}$$

We now derive bounds for the variances of GSG and cGSG.

Lemma 2.4. *Under Assumption 1.2, if $g(x)$ is calculated by (2.14), then, for all $x \in \mathbb{R}^n$, $\operatorname{Var}\{g(x)\} \preceq \kappa(x)I$ where*

$$\kappa(x) = \frac{1}{N}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2(n+2)L\epsilon_f\right).$$

Alternatively, under Assumption 1.3, if $g(x)$ is calculated by (2.15), then, for all $x \in \mathbb{R}^n$, $\operatorname{Var}\{g(x)\} \preceq \kappa(x)I$ where

$$\kappa(x) = \frac{1}{N}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)(n+8)M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{(n+1)(n+3)M\sigma\epsilon_f}{6\sqrt{n}}\right).$$

Proof. By (2.17), we have

$$\begin{aligned}
\operatorname{Var}\{g(x)\} &= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal \\
&\preceq \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x)-\epsilon(x)}{\sigma}\right)^2 uu^\intercal\right] \\
&= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\phi(x+\sigma u)-\phi(x)}{\sigma}\right)^2 uu^\intercal\right] + \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\epsilon(x+\sigma u)-\epsilon(x)}{\sigma}\right)^2 uu^\intercal\right] \\
&\quad + \frac{2}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{(\phi(x+\sigma u)-\phi(x))(\epsilon(x+\sigma u)-\epsilon(x))}{\sigma^2}uu^\intercal\right] \\
&\preceq \frac{1}{N\sigma^2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\nabla\phi(x)^\intercal\sigma u + \frac{1}{2}L\sigma^2 u^\intercal u\right)^2 uu^\intercal\right] + \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{2\epsilon_f}{\sigma}\right)^2 uu^\intercal\right] \\
&\quad + \frac{2}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\left(\nabla\phi(x)^\intercal\sigma u + \frac{1}{2}L\sigma^2 u^\intercal u\right)2\epsilon_f}{\sigma^2}uu^\intercal\right] \\
&\overset{(2.19)}{=} \frac{1}{N\sigma^2}\left(\sigma^2(\nabla\phi(x)^\intercal\nabla\phi(x)I + 2\nabla\phi(x)\nabla\phi(x)^\intercal) + L\sigma^3\cdot 0 + \frac{1}{4}L^2\sigma^4(n+2)(n+4)I\right) \\
&\quad + \frac{4\epsilon_f^2}{N\sigma^2}I +^\intercal \frac{2(n+2)L\epsilon_f}{N}I \\
&\preceq \frac{1}{N}\left(3\|\nabla\phi(x)\|^2 + \frac{1}{4}L^2\sigma^2(n+2)(n+4) + \frac{4\epsilon_f^2}{\sigma^2} + 2(n+2)L\epsilon_f\right)I,
\end{aligned}$$

where the first inequality comes from the fact that $-\nabla F(x)\nabla F(x)^\intercal/N \preceq 0$, the second inequality comes from the Lipschitz continuity of the gradients (Assumption 1.2) and the bound on noise, and the bound on

noise, and the last inequality is due to $\nabla\phi(x)\nabla\phi(x)^\intercal \preceq \nabla\phi(x)^\intercal\nabla\phi(x)I$.

For cGSG, by (2.18) we have

$$\begin{aligned}
\mathrm{Var}\left\{g(x)\right\} &= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{f(x+\sigma u)-f(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal \\
&\preceq \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x-\sigma u)-\epsilon(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] \\
&= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\phi(x+\sigma u)-\phi(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] + \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\epsilon(x+\sigma u)-\epsilon(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] \\
&\quad + \frac{2}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{(\phi(x+\sigma u)-\phi(x-\sigma u))(\epsilon(x+\sigma u)-\epsilon(x-\sigma u))}{4\sigma^2}uu^\intercal\right] \\
&\preceq \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\nabla\phi(x)^\intercal u + \frac{M\sigma^2}{6}\|u\|^3\right)^2 uu^\intercal\right] + \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{\epsilon_f}{\sigma}\right)^2 uu^\intercal\right] \\
&\quad + \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\nabla\phi(x)^\intercal u + \frac{M\sigma^2}{6}\|u\|^3\right)\left(\frac{\epsilon_f}{\sigma}\right)uu^\intercal\right] \\
&\overset{(2.19)}{\preceq} \frac{1}{N}\left(\nabla\phi(x)^\intercal\nabla\phi(x)I + 2\nabla\phi(x)\nabla\phi(x)^\intercal + \frac{(n+2)(n+4)(n+8)}{36}M^2\sigma^4 I\right) \\
&\quad + \frac{\epsilon_f^2}{N\sigma^2}I + \frac{(n+1)(n+3)M\sigma\epsilon_f}{6\sqrt{n}N}I \\
&\preceq \frac{1}{N}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)(n+8)}{36}M^2\sigma^4 + \frac{\epsilon_f^2}{\sigma^2} + \frac{(n+1)(n+3)}{6\sqrt{n}}M\sigma\epsilon_f\right)I,
\end{aligned}$$

where the first inequality comes from the fact that $-\nabla F(x)\nabla F(x)^\intercal/N \preceq 0$, the second inequality comes from the Lipschitz continuity of the Hessians (Assumption 1.3) and the bound on noise, and the last inequality is due to $\nabla\phi(x)\nabla\phi(x)^\intercal \preceq \nabla\phi(x)^\intercal\nabla\phi(x)I$. □

Using the results of Lemma 2.4, we can now bound the quantity $\|g(x)-\nabla F(x)\|$ in (2.16), in probability, using Chebyshev's inequality.

Lemma 2.5. *Let F be a Gaussian smoothed approximation of f (2.9). Under Assumption 1.2, if $g(x)$ is calculated via (2.14) with sample size*

$$N \geq \frac{n}{\delta r^2}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2(n+2)L\epsilon_f\right),$$

then, for all $x\in\mathbb{R}^n$, $\|g(x)-\nabla F(x)\|\leq r$ holds with probability at least $1-\delta$, for any $r>0$ and $0<\delta<1$.

Alternatively, under Assumption 1.3, if $g(x)$ is calculated via (2.15) with sample size $2N$ where

$$N \geq \frac{n}{\delta r^2}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)(n+8)M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{(n+1)(n+3)M\sigma\epsilon_f}{6\sqrt{n}}\right),$$

then, for all $x\in\mathbb{R}^n$, $\|g(x)-\nabla F(x)\|\leq r$ holds with probability at least $1-\delta$, for any $r>0$ and $0<\delta<1$.

Proof. By Chebyshev's inequality, for any $r > 0$, we have

$$\mathbb{P}\left\{\sqrt{(g(x) - \nabla F(x))^\intercal \mathrm{Var}\left\{g(x)\right\}^{-1}(g(x) - \nabla F(x))} > r\right\} \leq \frac{n}{r^2}.$$

Since by Lemma 2.4 $\mathrm{Var}\left\{g(x)\right\} \preceq \kappa(x) I$, with the appropriate $\kappa(x)$ as shown in the statement of the Lemma, we have $\mathrm{Var}\left\{g(x)\right\}^{-1} \succeq \kappa^{-1} I$ and

$$\sqrt{(g(x) - \nabla F(x))^\intercal \mathrm{Var}\left\{g(x)\right\}^{-1}(g(x) - \nabla F(x))} \geq \kappa^{-\frac{1}{2}} \|g(x) - \nabla F(x)\|.$$

Therefore, we have,

$$\mathbb{P}\left\{\kappa^{-\frac{1}{2}} \|g(x) - \nabla F(x)\| > r\right\} \leq \frac{n}{r^2} \implies \mathbb{P}\left\{\|g(x) - \nabla F(x)\| > r\right\} \leq \frac{\kappa n}{r^2}.$$

To ensure $\mathbb{P}\left\{\|g(x) - \nabla F(x)\| \leq r\right\} \geq 1 - \delta$, we choose κ such that $\frac{\kappa n}{r^2} \leq \delta$, by choosing large enough N. The exact bounds on N follow immediately from the expressions for $\kappa(x)$ in Lemma 2.4. □

Now with bounds for both terms in (2.16), we can bound $\|g(x) - \nabla f(x)\|$, in probability.

Theorem 2.6. *Suppose that Assumption 1.2 holds and $g(x)$ is calculated via (2.14). If*

$$N \geq \frac{n}{\delta r^2}\left(3\|\nabla \phi(x)\|^2 + \frac{(n+2)(n+4)L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2(n+2)L\epsilon_f\right),$$

then, for all $x \in \mathbb{R}^n$ and $r > 0$,

$$\|g(x) - \nabla f(x)\| \leq \sqrt{n}L\sigma + \frac{\sqrt{n}\epsilon_f}{\sigma} + r. \tag{2.20}$$

with probability at least $1 - \delta$.

Alternatively, suppose that Assumption 1.3 holds and $g(x)$ is calculated via (2.15). If

$$N \geq \frac{n}{\delta r^2}\left(3\|\nabla \phi(x)\|^2 + \frac{(n+2)(n+4)(n+8)M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{(n+1)(n+3)M\sigma\epsilon_f}{6\sqrt{n}}\right),$$

then, for all $x \in \mathbb{R}^n$ and $r > 0$,

$$\|g(x) - \nabla f(x)\| \leq nM\sigma^2 + \frac{\sqrt{n}\epsilon_f}{\sigma} + r. \tag{2.21}$$

with probability at least $1 - \delta$.

Proof. The proof of the first part (2.20) is a straightforward combination of the bound in (2.11) and the result of the first part of Lemma 2.5. The proof for the second part (2.21) is a straightforward combination of the bound in (2.12) and the result of the second part of Lemma 2.5. □

Using the results of Theorem 2.6, we can now derive bounds on σ and N that ensure that (1.3) holds with probability $1 - \delta$. To ensure (1.3), with probability $1 - \delta$, using Theorem 2.6 we want the following to hold

$$\sqrt{n}L\sigma + \frac{\sqrt{n}\epsilon_f}{\sigma} \leq \lambda\theta\|\nabla \phi(x)\|, \tag{2.22}$$

$$r \leq (1 - \lambda)\theta\|\nabla \phi(x)\|, \tag{2.23}$$

for some $\lambda \in (0, 1)$.

Let us first consider $g(x)$ calculated via (2.14). To ensure that (2.22) holds, we impose the same conditions as in the Forward Finite Difference case in Section 2.1, namely,

$$\sqrt{\frac{\epsilon_f}{L}} \leq \sigma \leq \frac{\lambda\theta\|\nabla\phi(x)\|}{2\sqrt{n}L}, \quad \|\nabla\phi(x)\| \geq \frac{2\sqrt{nL\epsilon_f}}{\lambda\theta}.$$

Now using these bounds and substituting $r = (1-\lambda)\theta\|\nabla\phi(x)\|$ into the first bound on N in Theorem 2.6 we have

$$\begin{aligned}\frac{n}{\delta r^2}\left(3\|\nabla\phi(x)\|^2 + \frac{1}{4}L^2\sigma^2(n+2)(n+4) + \frac{4\epsilon_f^2}{\sigma^2} + 2(n+2)L\epsilon_f\right) \\ \leq \frac{3n}{\delta\theta^2}\frac{1}{(1-\lambda)^2} + \left(\frac{(n+2)(n+4)}{16\delta} + \frac{1}{\delta} + \frac{(n+2)}{2\delta}\right)\frac{\lambda^2}{(1-\lambda)^2}.\end{aligned} \tag{2.24}$$

We are interested in making the lower bound on N as small as possible and hence we are concerned with its dependence on n, when n is relatively large. Henceforth, we assume that $n > 1$ and choose λ such that $\frac{\lambda^2}{(1-\lambda)^2} \leq \frac{1}{n+2}$ so as to reduce the scaling of the second term with n and to simplify the expression. This is always possible, because $\frac{\lambda^2}{(1-\lambda)^2}$ is monotonically increasing with λ and equals 0 for $\lambda = 0$. Specifically, we can choose $\lambda = \frac{1}{3\sqrt{n}}$, because it is easy to show that for for this value of λ, $\frac{\lambda^2}{(1-\lambda)^2} \leq \frac{1}{n+2}$ for all $n \geq 1$. In fact, for large values of n we can choose λ to be closer in value to $\frac{1}{\sqrt{n}}$, but for simplicity we will consider the choice that fits all n. Using the fact that $\lambda \leq \frac{1}{\sqrt{n}}$, and thus $\frac{1}{(1-\lambda)^2} \leq \frac{n}{(\sqrt{n}-1)^2}$, and also that $\frac{1}{n+2} \leq \frac{1}{2}$, the right hand side of (2.24) is bounded from above by

$$\frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+4)}{16\delta} + \frac{1}{\delta}. \tag{2.25}$$

This implies that by choosing N at least as large as the value of (2.25) we ensure that (2.23) holds.

We now summarize the result for the gradient approximation computed via (2.14), using the fact that $\lambda = \frac{1}{3\sqrt{n}}$.

Corollary 2.7. *Suppose that Assumption 1.2 holds, $n > 1$ and $g(x_k)$ is computed via (2.14) with N and σ satisfying*

$$\begin{aligned}N &\geq \frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+4)}{16\delta} + \frac{1}{\delta}, \\ \sqrt{\frac{\epsilon_f}{L}} \leq \sigma \leq \frac{\theta\|\nabla\phi(x)\|}{6nL}, &\quad \|\nabla\phi(x)\| \geq \frac{6n\sqrt{L\epsilon_f}}{\theta},\end{aligned}$$

then (1.3) holds with probability $1-\delta$.

Let us compare the sampling radius σ and the number of samples N with those used by forward finite differences and interpolation to guarantee (1.3). For FFD the upper bound on σ is $\frac{\theta\|\nabla f(x)\|}{\sqrt{n}L}$, which is $\Omega(\sqrt{n})$ times larger than the one derived for GSG. However, this is natural because the expected length of σu where $u \sim \mathcal{N}(0, I)$ is approximately $\sqrt{n}\sigma$; in other words, in expectation the sample points evaluated by GSG have the same distance to x as the sample points evaluated by FFD.

The bound for the number of samples for GSG is larger than those required by FFD and interpolation, since the latter are fixed at n, although both scale linearly in n. Moreover, the dependence of N on δ is high. However, this bound is derived as an upper bound, and hence in order to verify that GSG indeed requires a large number of samples to satisfy (1.3) we need to establish the lower bound on N. In what follows we show that linear scaling of N with respect to n is necessary to guarantee that (1.3) is satisfied. The dependence on δ is likely to be too pessimistic and is an artifact of using Chebychev's inequality. In the next section we

analyze a method that estimates gradients using samples uniformly distributed on a sphere, and for which we obtain better dependence on δ but still linear scaling with n. Note, also, that the dependence of the lower bound for $\|\nabla\phi(x)\|$ on ϵ_f is larger by a factor of $\sqrt{n}$ in the GSG case when compared to the FFD case, unless GSG is applied with N of the order of n^2.

We now derive the analogous bounds on N and σ for the case when $g(x)$ is calculated via (2.15). To ensure (1.3), with probability $1-\delta$, using Theorem 2.6 we want the following to hold

$$nM\sigma^2 + \frac{\sqrt{n}\epsilon_f}{\sigma} \leq \lambda\theta\|\nabla\phi(x)\|, \tag{2.26}$$

$$r \leq (1-\lambda)\theta\|\nabla\phi(x)\|, \tag{2.27}$$

for some $\lambda \in (0,1)$. In order to ensure that (2.26) holds, we use the same logic as was done for Central Finite Differences in Section 2.1. Namely, we require the following,

$$\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}} \leq \sigma \leq \sqrt{\frac{\lambda\theta\|\nabla\phi(x)\|}{2nM}}, \quad \|\nabla\phi(x)\| \geq \frac{2\sqrt[3]{n^2M\epsilon_f^2}}{\lambda\theta}.$$

Now using these bounds and setting $r = (1-\lambda)\theta\|\nabla\phi(x)\|$ into the second bound on N in Theorem 2.6 we have

$$\frac{n}{\delta r^2}\left(3\|\nabla\phi(x)\|^2 + \frac{(n+2)(n+4)(n+6)M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{(n+1)(n+3)M\sigma\epsilon_f}{6\sqrt{n}}\right)$$
$$\leq \frac{3n}{\delta\theta^2}\frac{1}{(1-\lambda)^2} + \left(\frac{(n+2)(n+4)(n+6)}{144n\delta} + \frac{1}{4\delta} + \frac{(n+1)(n+3)}{24n\delta}\right)\frac{\lambda^2}{(1-\lambda)^2}.$$

As before, we are interested in making the lower bound on N to scale at most linearly with n. Thus, to achieve this and to simplify the expression we choose λ such that $\frac{\lambda^2}{(1-\lambda)^2} \leq \frac{n}{(n+2)(n+4)}$, which. It is easy to show that $\lambda = \frac{1}{6\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ satisfies this condition. Then, using again the fact that $\frac{1}{(1-\lambda)^2} \leq \frac{n}{(\sqrt{n}-1)^2}$ and $\frac{n}{(n+2)(n+4)} \leq \frac{1}{2}$ the above expression is bounded by

$$\frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+6)}{144\delta} + \frac{1}{8\delta} + \frac{1}{24\delta}.$$

We now summarize the result for the gradient approximation computed via (2.15), using the fact that $\lambda = \frac{1}{6\sqrt{n}}$.

Corollary 2.8. *Suppose that Assumption 1.3 holds, $n > 1$ and $g(x_k)$ is computed via (2.15) with N and σ satisfying*

$$N \geq \frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+6)}{144\delta} + \frac{1}{6\delta},$$

$$\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}} \leq \sigma \leq \sqrt{\frac{\theta\|\nabla\phi(x)\|}{12n^{3/2}M}}, \quad \|\nabla\phi(x)\| \geq \frac{12\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta},$$

then (1.3) holds with probability $1-\delta$.

2.3.1 Lower bound on δ

We have demonstrated that if $N \geq \Omega(\frac{3n}{\theta^2\delta})$, then $\mathbb{P}(\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|) \geq 1-\delta$; that is, having a sufficiently large number of samples is *sufficient* to ensure accurate gradient approximations with a desired probability. A question that remains is how many samples are *necessary* to ensure that accurate gradient

approximations are obtained with high probability. Here we derive a lower bound on the probability of failure for (2.14) to satisfy condition (1.3); i.e.,

$$\mathbb{P}\left(\|g(x) - \nabla\phi(x)\| > \theta\|\nabla\phi(x)\|\right). \tag{2.28}$$

We derive the lower bound for (2.28) theoretically, and then illustrate the lower bounds via numerical simulations for the specific case of a simple linear function of the form $f(x) = \phi(x) = a^\intercal x$, where a is an arbitrary nonzero vector in $\mathbb{R}^n$. For simplicity, through this subsection we assume that $\epsilon(x) = 0$ for all $x \in \mathbb{R}^n$. In this case, for any σ, $\nabla F(x) = a$. Note also that in this case $\nabla f(x) = \nabla\phi(x) = \nabla F(x) = a$. We show that while theory gives us a weak lower bound, numerical simulations indicate that the lower bound is reasonably close to the upper bound, in terms of dependence on n.

We use the following lower bound on the tail of a random variable X, derived in [26]. For any b that satisfies $0 \leq b \leq \mathbb{E}\left[|X|\right] < \infty$,

$$\mathbb{P}(|X| > b) \geq \frac{(\mathbb{E}[|X|] - b)^2}{\mathbb{E}[|X|^2]}.$$

We apply this bound to the random variable $\|g(x) - \nabla F(x)\|$, and $b = \theta\|a\|$. We have

$$\begin{aligned}
\mathbb{P}\left(\|g(x) - \nabla\phi(x)\| > \theta\|\nabla\phi(x)\|\right) &= \mathbb{P}(\|g(x) - \nabla F(x)\| > b) \\
&= \mathbb{P}(\|g(x) - \nabla F(x)\|^2 > b^2) \\
&\geq \frac{\left(\mathbb{E}\left[\|g(x) - \nabla F(x)\|^2\right] - b^2\right)^2}{\mathbb{E}\left[\|g(x) - \nabla F(x)\|^4\right]}
\end{aligned}$$

for any b such that $0 \leq b \leq \sqrt{\mathbb{E}\left[\|g(x) - \nabla F(x)\|^2\right]}$. The reason we consider the squared version of the condition is we are unable to calculate $\mathbb{E}\left[\|g(x) - \nabla F(x)\|^k\right]$ when k is odd.

For brevity, we omit the derivations of $\mathbb{E}\left[\|g(x) - \nabla F(x)\|^2\right]$ and $\mathbb{E}\left[\|g(x) - \nabla F(x)\|^4\right]$ from the main paper, and refer the reader to Appendices A.4 and A.5, respectively. The required expressions are:

$$\mathbb{E}\left[\|g(x) - \nabla F(x)\|^2\right] = \frac{1}{N}(n+1)a^\intercal a, \tag{2.29}$$

$$\mathbb{E}\left[\|g(x) - \nabla F(x)\|^4\right] = \frac{1}{N^3}\left((N-1)(n^2+4n+7)(a^\intercal a)^2 + (3n^2+20n+37)(a^\intercal a)^2\right). \tag{2.30}$$

Thus for $\phi(x) = a^\intercal x$,

$$\begin{aligned}
\mathbb{P}\left(\|g(x) - a\| > \theta\|a\|\right) &\geq \frac{N^4\left(\frac{1}{N}(n+1)a^\intercal a - \theta^2 a^\intercal a\right)^2}{N(N-1)(n^2+4n+7)(a^\intercal a)^2 + N(3n^2+20n+37)(a^\intercal a)^2} \\
&= \frac{N\left((n+1)a^\intercal a - N\theta^2 a^\intercal a\right)^2}{(N-1)(n^2+4n+7)(a^\intercal a)^2 + (3n^2+20n+37)(a^\intercal a)^2} \\
&= \frac{N\left((n+1) - \theta^2 N\right)^2}{(N-1)(n^2+4n+7) + (3n^2+20n+37)}
\end{aligned}$$

for any θ and N such that $0 \leq \theta^2 a^\intercal a \leq \frac{1}{N}(n+1)a^\intercal a$.

Consider n large enough such that $4(n+1)^2 \geq n^2+4n+7$ which is satisfied for $n \geq \frac{\sqrt{13}-2}{3} \approx 0.54$; and $4(n+1)^2 \geq 3n^2+20n+37$ which is satisfied for $n \geq 6+\sqrt{69} \approx 14.31$ (henceforth we assume that $n \geq 15$). Then we have,

$$\frac{N((n+1) - \theta^2 N)^2}{(N-1)(n^2+4n+7) + (3n^2+20n+37)} \geq \frac{N((n+1) - \theta^2 N)^2}{(N-1)4(n+1)^2 + 4(n+1)^2} = \frac{((n+1) - \theta^2 N)^2}{4(n+1)^2}.$$

Thus, from

$$\mathbb{P}\left(\|g(x)-a\|>\theta\|a\|\right)\geq\frac{((n+1)-\theta^2N)^2}{4(n+1)^2}\geq\delta,$$

we get

$$N\leq\frac{(n+1)(1-2\sqrt{\delta})}{\theta^2}\Rightarrow\mathbb{P}(\|g(x)-\nabla\phi(x)\|>\theta\|\nabla\phi(x)\|)\geq\delta.$$

It follows that for any $0<\delta<\frac{1}{4}$, $n\geq 15$ and $N\leq\frac{1}{\theta^2}(1-2\sqrt{\delta})(n+1)$

$$\mathbb{P}(\|g(x)-\nabla\phi(x)\|>\theta\|\nabla\phi(x)\|)\geq\delta.$$

In other words, to have $\mathbb{P}(\|g(x)-\nabla\phi(x)\|\leq\theta\|\nabla\phi(x)\|)>1-\delta$, it is necessary to have $N>\frac{(1-2\sqrt{\delta})}{\theta^2}(n+1)$, which is a linear function in n.

We now show through numerical simulation of the specific case of $\phi(x)=a^\intercal x$ that in fact for much larger values of δ, $N\geq n$ is required to achieve (1.3) for any $\theta<1$. Specifically, Figure 1 shows the distribution of

$$\theta=\frac{\|g(x)-\nabla\phi(x)\|}{\|\nabla\phi(x)\|},$$

approximately computed via running 10000 experiments, where $\phi(x)=e^\intercal x$ (e is a vector of all ones) and $n=32$, for different choices of $N\in\{1,2,4,8,16,32,64,128,256,512\}$. As is clear, θ is never smaller than 1 when $N=1$. Moreover, θ is smaller than $\frac{1}{2}$, which is required by the theory in [3], only about half the time when $N=128=4n$. Figure 1(k) shows the percent of successful trials ($\theta<\frac{1}{2}$) versus the size of the sample set (N), and Table 1(l) shows statistics of the empirical experiments for different sizes of the sample set (N). As expected, as N grows, the value of θ decreases, something that is not surprising, but at the same time not captured by the derived lower bound. Thus, we conclude that the theoretical lower bound we derive here is weak and to satisfy (1.3) with $\theta<\frac{1}{2}$ and probability of at least $\frac{1}{2}$ the size of the sample set needs to be larger than n. A stronger theoretical lower bound supporting this claim remains an open question.

In Section 3 we present numerical evidence that shows that for a variety of functions choosing N to be a small constant almost always results in large values of $\frac{\|g(x)-\nabla\phi(x)\|}{\|\nabla\phi(x)\|}$ with probability close to 1.

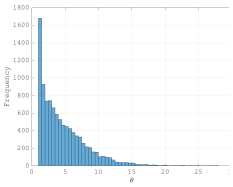
2.4 Gradient Estimation via Smoothing on a Sphere

Similar to the Gaussian smoothing technique, one can also smooth the function f with a uniform distribution on a ball, i.e.,

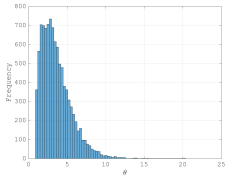
$$\begin{aligned}F(x)&=\mathbb{E}_{y\sim\mathcal{U}(\mathcal{B}(x,\sigma))}[f(y)]=\int_{\mathcal{B}(x,\sigma)}f(y)\frac{1}{V_n(\sigma)}dy\\&=\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}[f(x+\sigma u)]=\int_{\mathcal{B}(0,1)}f(x+\sigma u)\frac{1}{V_n(1)}du,\end{aligned}\tag{2.31}$$

where $\mathcal{U}(\mathcal{B}(x,\sigma))$ denotes the multivariate uniform distribution on a ball of radius σ centered at x and $\mathcal{U}(\mathcal{B}(0,1))$ denotes the multivariate uniform distribution on a ball of radius 1 centered at 0. The function $V_n(\sigma)$ represents the volume of a ball in $\mathbb{R}^n$ of radius σ. It was shown in [16] that the gradient of F can be expressed as

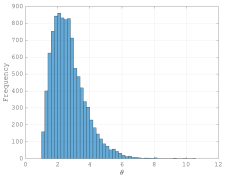
$$\nabla F(x)=\frac{n}{\sigma}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}[f(x+\sigma u)u],$$



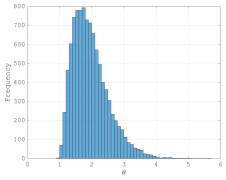
(a) $N = 1$ $(n/32)$



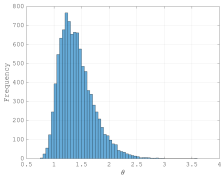
(b) $N = 2$ $(n/16)$



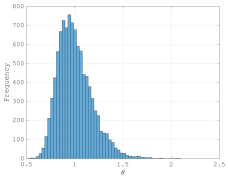
(c) $N = 4$ $(n/8)$



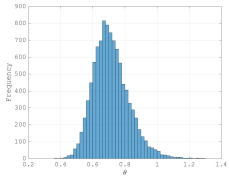
(d) $N = 8$ $(n/4)$



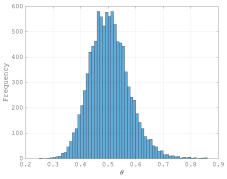
(e) $N = 16$ $(n/2)$



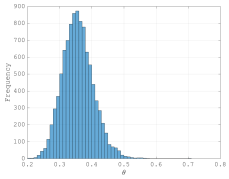
(f) $N = 32$ (n)



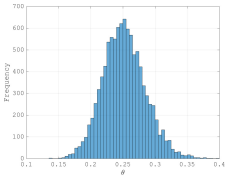
(g) $N = 64$ $(2n)$



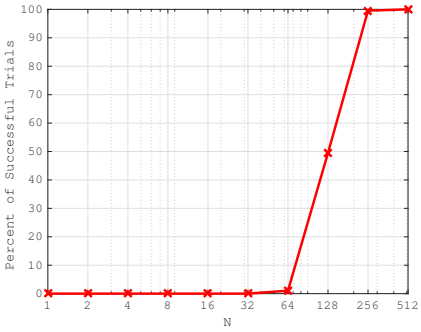
(h) $N = 128$ $(4n)$



(i) $N = 256$ $(8n)$



(j) $N = 512$ $(16n)$



(k) Percent of successful trials vs. N

N	Mean (θ)	Median (θ)	Variance (θ)	Percent of Successful Trials
1 $(n/32)$	4.62	3.69	11.44	0
2 $(n/16)$	3.62	3.23	3.67	0
4 $(n/8)$	2.70	2.53	1.07	0
8 $(n/4)$	1.96	1.87	0.31	0
16 $(n/2)$	1.41	1.36	0.093	0
32 (n)	1.00	0.98	0.032	0
64 $(2n)$	0.71	0.70	0.012	1.04
128 $(4n)$	0.50	0.50	0.0051	49.53
256 $(8n)$	0.36	0.35	0.0023	99.56
512 $(16n)$	0.25	0.25	0.0011	100

(l) Summary of Results

Figure 1: Distribution of θ for $\phi(x) = e^\intercal x$, at $x = e$, where e is a vector of all ones, and $n = 32$.

where $\mathcal{S}(0, 1)$ represents a unit sphere of radius 1 centered at 0. This leads to three ways of approximating the gradient with only function evaluations using sample average approximations (SAA)

$$g(x) = \frac{n}{N\sigma} \sum_{i=1}^{N} f(x + \sigma u_i) u_i, \tag{2.32}$$

$$g(x) = \frac{n}{N} \sum_{i=1}^{N} \frac{f(x + \sigma u_i) - f(x)}{\sigma} u_i, \tag{2.33}$$

$$g(x) = \frac{n}{N} \sum_{i=1}^{N} \frac{f(x + \sigma u_i) - f(x - \sigma u_i)}{2\sigma} u_i, \tag{2.34}$$

with N independently and identically distributed random vectors $\{u_i\}_{i=1}^N$ following a uniform distribution on the unit sphere. Similar to the case with Gaussian smoothing, the variance of (2.32) explodes when σ goes to zero, and thus we do not consider this formula. We analyze (2.33), which we refer to as ball smoothed gradient (BSG) and (2.34) which we refer to as central BSG (cBSG).

Again, as in the Gaussian smoothed case, there are two sources of error in the gradient approximations, and namely,

$$\|g(x) - \nabla\phi(x)\| \leq \|\nabla F(x) - \nabla\phi(x)\| + \|g(x) - \nabla F(x)\|. \tag{2.35}$$

Let Assumption 1.1 hold. One can bound the first term as follows; if the function ϕ has L-Lipschitz continuous gradients, that is if Assumption 1.2 holds, then

$$\|\nabla F(x) - \nabla\phi(x)\| \leq L\sigma + \frac{n\epsilon_f}{\sigma}, \tag{2.36}$$

and if the function ϕ has M-Lipschitz continuous Hessians, that is if Assumption 1.3 holds, then

$$\|\nabla F(x) - \nabla\phi(x)\| \leq M\sigma^2 + \frac{n\epsilon_f}{\sigma}. \tag{2.37}$$

The proofs are given in Appendices A.6 and A.7, respectively.

For the second error term in (2.35), similar to the case of Gaussian smoothing, we begin with the variance of $g(x)$. The variance of (2.33) can be expressed as

$$\text{Var}\left\{g(x)\right\} = \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{f(x+\sigma u) - f(x)}{\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal, \tag{2.38}$$

and the variance of (2.34) can be expressed as

$$\text{Var}\left\{g(x)\right\} = \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{f(x+\sigma u) - f(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal. \tag{2.39}$$

For a random variable $u \in \mathbb{R}^n$ that is uniformly distributed on the unit sphere $\mathcal{S}(0,1) \subset \mathbb{R}^n$, we have

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[(a^\intercal u)^2 uu^\intercal\right] = \frac{a^\intercal a I + 2aa^\intercal}{n(n+2)} \\
&\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[a^\intercal u\|u\|^k uu^\intercal\right] = 0_{n\times n} \text{ for } k = 0,1,2,... \\
&\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\|u\|^k uu^\intercal\right] = \frac{1}{n}I \text{ for } k = 0,1,2,...,
\end{aligned} \tag{2.40}$$

where a is any vector in $\mathbb{R}^n$ independent from u; see Appendix A.8 for derivations. We now provide bounds for the variances of BSG and cBSG under the assumption of Lipschitz continuous gradients and Hessians, respectively.

Lemma 2.9. *Under Assumption 1.2, if $g(x)$ is calculated by (2.33), then, for all $x \in \mathbb{R}^n$, $\text{Var}\left\{g(x)\right\} \preceq \kappa(x)I$ where*

$$\kappa(x) = \frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2 + \frac{nL^2\sigma^2}{4} + \frac{4n\epsilon_f^2}{\sigma^2} + 2nL\epsilon_f\right).$$

Alternatively, under Assumption 1.3, if $g(x)$ is calculated by (2.34), then, for all $x \in \mathbb{R}^n$, $\text{Var}\left\{g(x)\right\} \preceq \kappa(x)I$ where

$$\kappa(x) = \frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2 + \frac{nM^2\sigma^4}{36} + \frac{n\epsilon_f^2}{\sigma^2} + \frac{nM\sigma\epsilon_f}{3}\right).$$

Proof. By (2.38), we have

$$\begin{aligned}
\operatorname{Var}\{g(x)\} &= \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal \\
&\preceq \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x)-\epsilon(x)}{\sigma}\right)^2 uu^\intercal\right] \\
&\preceq \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{\nabla\phi(x)^\intercal\sigma u+\frac{1}{2}L\sigma^2 u^\intercal u+2\epsilon_f}{\sigma}\right)^2 uu^\intercal\right] \\
&\overset{(2.40)}{=} \frac{n^2}{N}\left(\frac{\nabla\phi(x)^\intercal\nabla\phi(x)I+2\nabla\phi(x)\nabla\phi(x)^\intercal}{n(n+2)}+\frac{L^2\sigma^4}{4n}I+\frac{4\epsilon_f^2}{n\sigma^2}I+\frac{2L\epsilon_f}{n}I\right) \\
&\preceq \frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2+\frac{nL^2\sigma^2}{4}+\frac{4n\epsilon_f^2}{\sigma^2}+2nL\epsilon_f\right)I,
\end{aligned}$$

where the first inequality comes from $-\nabla F(x)\nabla F(x)^\intercal/N \preceq 0$, the second inequality comes from the L-Lipschitz continuity of the gradients (Assumption 1.2) and the bound on the noise, and the last inequality is due to $\nabla\phi(x)\nabla\phi(x)^\intercal \preceq \nabla\phi(x)^\intercal\nabla\phi(x)I$.

For cBSG, by (2.39) we have

$$\begin{aligned}
\operatorname{Var}\{g(x)\} &= \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{f(x+\sigma u)-f(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] - \frac{1}{N}\nabla F(x)\nabla F(x)^\intercal \\
&\preceq \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x-\sigma u)-\epsilon(x-\sigma u)}{2\sigma}\right)^2 uu^\intercal\right] \\
&\preceq \frac{n^2}{N}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left(\frac{\nabla\phi(x)^\intercal 2\sigma u+2\frac{M\sigma^3}{6}\|u\|^3+2\epsilon_f}{2\sigma}\right)^2 uu^\intercal\right] \\
&\overset{(2.40)}{=} \frac{n^2}{N}\left(\frac{\nabla\phi(x)^\intercal\nabla\phi(x)I+2\nabla\phi(x)\nabla\phi(x)^\intercal}{n(n+2)}+\frac{M^2\sigma^4}{36n}I+\frac{\epsilon_f^2}{n\sigma^2}I+\frac{M\sigma\epsilon_f}{3n}I\right) \\
&\preceq \frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2+\frac{nM^2\sigma^4}{36}+\frac{n\epsilon_f^2}{\sigma^2}+\frac{nM\sigma\epsilon_f}{3}\right)I,
\end{aligned}$$

where the first inequality comes from $-\nabla F(x)\nabla F(x)^\intercal/N \preceq 0$, the second inequality comes from the Lipschitz continuity of the Hessians (Assumption 1.3) and the bound on noise, and the last inequality is due to $\nabla\phi(x)\nabla\phi(x)^\intercal \preceq \nabla\phi(x)^\intercal\nabla\phi(x)I$. □

Using the results of Lemma 2.9, we can bound the quantity $\|g(x)-\nabla F(x)\|$ in (2.35), with probability $1-\delta$, using Chebyshev's inequality, just as we did in the case of GSG. However, ball smoothed gradient approach has a significant advantage over Gaussian smoothing in that it allows the use of Bernstein's inequality [34] instead of Chebychev's and the resulting bound on N has a significantly improved dependence on δ. Bernstein's inequality applies here because, unlike GSG (and cGSG), BSG (and cBSG) enjoys a deterministic bound on the error term $n\frac{f(x+\sigma u)-f(x)}{\sigma}u - F(x)$; see proof of Lemma 2.10.

Lemma 2.10. *Let F be a ball smoothed approximation of f (2.31). Under Assumption 1.2, if $g(x)$ is calculated via (2.33) with sample size*

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n}+\frac{L^2\sigma^2}{4}+\frac{4\epsilon_f^2}{\sigma^2}+2L\epsilon_f\right)+\frac{2n}{3r}\left(2\|\nabla f(x)\|+L\sigma+\frac{4\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta},$$

then, for all $x\in\mathbb{R}^n$, $\|g(x)-\nabla F(x)\|\leq r$ holds with probability at least $1-\delta$, for any $r>0$ and $0<\delta<1$.

Alternatively, under Assumption 1.3 if $g(x)$ is calculated via (2.34) with sample size $2N$ where

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta},$$

then, for all $x \in \mathbb{R}^n$, $\|g(x) - \nabla F(x)\| \leq r$ holds with probability at least $1-\delta$, for any $r > 0$ and $0 < \delta < 1$.

Proof. We first note that

$$\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{N}\frac{f(x+\sigma u) - f(x)}{\sigma}u - \frac{1}{N}\nabla F(x)\right] = 0,$$

and

$$\begin{aligned}
&\left\|\frac{n}{N}\frac{f(x+\sigma u) - f(x)}{\sigma}u - \frac{1}{N}\nabla F(x)\right\| \\
&\quad\leq \frac{n}{N\sigma}\left|f(x+\sigma u) - f(x)\right|\|u\| + \frac{n}{N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left|f(x+\sigma v) - f(x)\right|\|v\|\right] \\
&\quad= \frac{n}{N\sigma}\left|\phi(x+\sigma u) + \epsilon(x+\sigma u) - \phi(x) - \epsilon(x)\right| \\
&\qquad+ \frac{n}{N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left|\phi(x+\sigma v) + \epsilon(x+\sigma v) - \phi(x) - \epsilon(x)\right|\right] \\
&\quad\leq \frac{n}{N\sigma}\left(\nabla\phi(x)^\intercal\sigma u + \frac{L\|\sigma u\|^2}{2} + 2\epsilon_f\right) + \frac{n}{N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\nabla\phi(x)^\intercal\sigma v + \frac{L\|\sigma v\|^2}{2} + 2\epsilon_f\right] \\
&\quad\leq \frac{n}{N}(2\|\nabla f(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma}),
\end{aligned}$$

for any $u \sim \mathcal{U}(\mathcal{S}(0,1))$. The *matrix variance statistic* of $g(x) - \nabla F(x)$ is

$$\begin{aligned}
&v(g(x) - \nabla F(x)) \\
&\quad= \max\left\{\left\|\mathbb{E}\left[(g(x) - \nabla F(x))(g(x) - \nabla F(x))^\intercal\right]\right\|, \mathbb{E}\left[(g(x) - \nabla F(x))^\intercal(g(x) - \nabla F(x))\right]\right\} \\
&\quad\leq \max\left\{\frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2 + \frac{nL^2\sigma^2}{4} + \frac{4n\epsilon_f^2}{\sigma^2} + 2nL\epsilon_f\right), \frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right)\right\} \\
&\quad= \frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right).
\end{aligned}$$

By Bernstein's inequality, we have

$$\begin{aligned}
&\mathbb{P}(\|g(x) - \nabla F(x)\| \geq r) \\
&\quad\leq (n+1)\exp\left(\frac{-r^2/2}{v(g(x) - \nabla F(x)) + \frac{nr}{3N}(2\|\nabla f(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma})}\right) \\
&\quad\leq (n+1)\exp\left(\frac{-r^2/2}{\frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right) + \frac{nr}{3N}(2\|\nabla f(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma})}\right).
\end{aligned}$$

In order to ensure that $\mathbb{P}(\|g(x) - \nabla F(x)\| \geq r) \leq \delta$, for some $\delta \in (0,1)$, we require that

$$(n+1)\exp\left(\frac{-r^2/2}{\frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right) + \frac{nr}{3N}(2\|\nabla f(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma})}\right) \leq \delta,$$

from which we conclude that

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right) + \frac{2n}{3r}\left(2\|\nabla f(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta}.$$

For the cBSG case, note that

$$\mathbb{E}_{u_i\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{N}\frac{f(x+\sigma u_i)-f(x-\sigma u_i)}{2\sigma}u_i - \frac{1}{N}\nabla F(x)\right] = 0,$$

and

$$\begin{aligned}
&\left\|\frac{n}{N}\frac{f(x+\sigma u)-f(x-\sigma u)}{2\sigma}u - \frac{1}{N}\nabla F(x)\right\| \\
&\quad\leq \frac{n}{2N\sigma}\left|f(x+\sigma u)-f(x-\sigma u)\right|\|u\| + \frac{n}{2N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left|f(x+\sigma v)-f(x-\sigma v)\right|\|v\|\right] \\
&\quad= \frac{n}{2N\sigma}\left|\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x-\sigma u)-\epsilon(x-\sigma u)\right| \\
&\qquad\quad + \frac{n}{2N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left|\phi(x+\sigma v)+\epsilon(x+\sigma v)-\phi(x)-\epsilon(x)\right|\right] \\
&\quad\leq \frac{n}{2N\sigma}\left(2\nabla\phi(x)^\intercal\sigma u + \frac{M\|\sigma u\|^3}{3} + 2\epsilon_f\right) + \frac{n}{2N\sigma}\mathbb{E}_{v\sim\mathcal{U}(\mathcal{S}(0,1))}\left[2\nabla\phi(x)^\intercal\sigma u + \frac{M\|\sigma u\|^3}{3} + 2\epsilon_f\right] \\
&\quad\leq \frac{n}{N}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right),
\end{aligned}$$

for any $u\sim\mathcal{U}(\mathcal{S}(0,1))$. The *matrix variance statistic* of $g(x)-\nabla F(x)$ is

$$\begin{aligned}
&v(g(x)-\nabla F(x)) \\
&\quad= \max\left\{\left\|\mathbb{E}\left[(g(x)-\nabla F(x))(g(x)-\nabla F(x))^\intercal\right]\right\|, \mathbb{E}\left[(g(x)-\nabla F(x))^\intercal(g(x)-\nabla F(x))\right]\right\} \\
&\quad\leq \max\left\{\frac{1}{N}\left(\frac{3n}{n+2}\|\nabla\phi(x)\|^2 + \frac{nM^2\sigma^4}{36} + \frac{n\epsilon_f^2}{\sigma^2} + \frac{nM\sigma\epsilon_f}{3}\right), \frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right)\right\} \\
&\quad= \frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right).
\end{aligned}$$

By Bernstein's inequality, we have

$$\begin{aligned}
&\mathbb{P}(\|g(x)-\nabla F(x)\|\geq r) \\
&\qquad\leq (n+1)\exp\left(\frac{-r^2/2}{v(g(x)-\nabla F(x)) + \frac{nr}{3N}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)}\right) \\
&\qquad\leq (n+1)\exp\left(\frac{-r^2/2}{\frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{nr}{3N}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)}\right).
\end{aligned}$$

In order to ensure that $\mathbb{P}(\|g(x)-\nabla F(x)\|\geq r)\leq\delta$, for some $\delta\in(0,1)$, we require that

$$(n+1)\exp\left(\frac{-r^2/2}{\frac{n^2}{N}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{nr}{3N}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)}\right)\leq\delta,$$

from which we conclude that

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta}.$$

□

Now, with bounds for both terms in (2.35), we can bound $\|g(x) - \nabla\phi(x)\|$, in probability.

Theorem 2.11. *Suppose that Assumption 1.2 holds and $g(x)$ is calculated via (2.33). If*

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta},$$

then, for all $x \in \mathbb{R}^n$ and $r > 0$,

$$\|g(x) - \nabla\phi(x)\| \leq L\sigma + \frac{n\epsilon_f}{\sigma} + r. \tag{2.41}$$

with probability at least $1 - \delta$.

Alternatively, suppose that Assumption 1.3 holds and $g(x)$ is calculated via (2.34). If

$$N \geq \left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta},$$

then, for all $x \in \mathbb{R}^n$ and $r > 0$,

$$\|g(x) - \nabla\phi(x)\| \leq M\sigma^2 + \frac{n\epsilon_f}{\sigma} + r. \tag{2.42}$$

with probability at least $1 - \delta$.

Proof. The proof for the first part (2.41) is a straightforward combination of the bound in (2.36) and the result of the first part of Lemma 2.10. The proof for the second part (2.42) is a straightforward combination of the bound in (2.37) and the result of the second part of Lemma 2.10. □

In Theorem 2.11 one should notice the improved dependence of the size of the sample set N on the probability δ as compared to Theorem 2.6. While Bernstein's inequality does not apply in the case of the Gaussian smoothed gradient, there may be other ways to establish a better dependence on δ. However, the dependence on n in all cases is linear, which as we have shown for the GSG case is a necessary dependence. A similar lower bound result for BSG can be derived analogously.

Using the results of Theorem 2.11, as before, we derive bounds on σ and N that ensure that (1.3) holds with probability $1 - \delta$. To ensure (1.3), with probability $1 - \delta$, using Theorem 2.11 we want the following to hold

$$L\sigma + \frac{n\epsilon_f}{\sigma} \leq \lambda\theta\|\nabla\phi(x)\|, \tag{2.43}$$

$$r \leq (1-\lambda)\theta\|\nabla\phi(x)\|, \tag{2.44}$$

for some $\lambda \in (0, 1)$.

Let us first consider $g(x)$ calculated via (2.33). As before, to ensure that (2.43) holds, we impose the following conditions,

$$\sqrt{\frac{n\epsilon_f}{L}} \leq \sigma \leq \frac{\lambda\theta\|\nabla\phi(x)\|}{2L}, \quad \|\nabla\phi(x)\| \geq \frac{2\sqrt{nL\epsilon_f}}{\lambda\theta}.$$

Now using these bounds and substituting $r = (1-\lambda)\theta\|\nabla\phi(x)\|$ into the first bound on N in Theorem 2.11 we have

$$\begin{aligned}&\left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{L^2\sigma^2}{4} + \frac{4\epsilon_f^2}{\sigma^2} + 2L\epsilon_f\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + L\sigma + \frac{4\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta} \\ &\quad\leq \left[\frac{2n}{\theta^2}\frac{1}{(1-\lambda)^2} + \left(\frac{n^2}{8} + 2 + n\right)\frac{\lambda^2}{(1-\lambda)^2} + \frac{4n}{3\theta}\frac{1}{(1-\lambda)} + \left(\frac{n}{3} + \frac{4}{3}\right)\frac{\lambda}{(1-\lambda)}\right]\log\frac{n+1}{\delta}.\end{aligned} \tag{2.45}$$

As before, we are interested in making the lower bound on N to scale at most linearly with n. Thus, to achieve this and to simplify the expression we choose $\lambda = \frac{1}{2\sqrt{n}}$ so that $\frac{\lambda^2}{(1-\lambda)^2} \leq \frac{1}{n}$, for all n. Then, using that $\frac{1}{(1-\lambda)^2} \leq \frac{n}{(\sqrt{n}-1)^2}$ and $\frac{1}{n} \leq 1$ the above expression is bounded by

$$\left[\frac{2n}{\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{n}{8} + 3 + \frac{4n}{3\theta}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{\sqrt{n}}{3} + \frac{4}{3}\right]\log\frac{n+1}{\delta}. \tag{2.46}$$

This implies that by choosing N at least as large as the value of (2.46) we ensure that (2.44) holds.

We now summarize the result for the gradient approximation computed via (2.33), using the fact that $\lambda = \frac{1}{2\sqrt{n}}$.

Corollary 2.12. *Suppose that Assumption 1.2 holds, $n > 1$ and $g(x_k)$ is computed via (2.33) with N and σ satisfying*

$$N \geq \left[\frac{2n}{\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{n}{8} + \frac{4n}{3\theta}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{\sqrt{n}}{3} + \frac{13}{3}\right]\log\frac{n+1}{\delta},$$
$$\sqrt{\frac{n\epsilon_f}{L}} \leq \sigma \leq \frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}L}, \quad \|\nabla\phi(x)\| \geq \frac{4n\sqrt{L\epsilon_f}}{\theta},$$

then (1.3) holds with probability $1-\delta$.

We now derive the analogous bounds on N and σ for the case when $g(x)$ is calculated via (2.34). To ensure (1.3), with probability $1-\delta$, using Theorem 2.11 we want the following to hold

$$M\sigma^2 + \frac{n\epsilon_f}{\sigma} \leq \lambda\theta\|\nabla\phi(x)\|, \tag{2.47}$$
$$r \leq (1-\lambda)\theta\|\nabla\phi(x)\|, \tag{2.48}$$

for some $\lambda \in (0,1)$. In order to ensure that (2.47) holds, we use the same logic as was done for Central Finite Differences in Section 2.1. Namely, we require the following,

$$\sqrt[3]{\frac{n\epsilon_f}{M}} \leq \sigma \leq \sqrt{\frac{\lambda\theta\|\nabla\phi(x)\|}{2M}}, \quad \|\nabla\phi(x)\| \geq \frac{2\sqrt[3]{n^2M\epsilon_f^2}}{\lambda\theta}.$$

Now using these bounds and setting $r = (1-\lambda)\theta\|\nabla\phi(x)\|$ into the second bound on N in Theorem 2.6 we have

$$\left[\frac{2n^2}{r^2}\left(\frac{\|\nabla\phi(x)\|^2}{n} + \frac{M^2\sigma^4}{36} + \frac{\epsilon_f^2}{\sigma^2} + \frac{M\sigma\epsilon_f}{3}\right) + \frac{2n}{3r}\left(2\|\nabla\phi(x)\| + \frac{M\sigma^2}{3} + \frac{2\epsilon_f}{\sigma}\right)\right]\log\frac{n+1}{\delta}$$
$$\leq \left[\frac{2n}{\theta^2}\frac{1}{(1-\lambda)^2} + \left(\frac{n^2}{72} + \frac{1}{2} + \frac{n}{6}\right)\frac{\lambda^2}{(1-\lambda)^2} + \frac{4n}{3\theta}\frac{1}{(1-\lambda)} + \left(\frac{n}{9}+2\right)\frac{\lambda}{(1-\lambda)}\right]\log\frac{n+1}{\delta}.$$

As before, we are interested in making the lower bound on N to scale at most linearly with n. Thus, to achieve this and to simplify the expression we choose λ such that $\frac{\lambda^2}{(1-\lambda)^2} \leq \frac{1}{n}$, which, implies that $\lambda \leq \frac{1}{\sqrt{n}}$. Then, using again the fact that $\frac{1}{(1-\lambda)^2} \leq \frac{n}{(\sqrt{n}-1)^2}$ and $\frac{1}{n} \leq 1$ the above expression is bounded by

$$\left[\frac{2n}{\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{n}{72} + \frac{2}{3} + \frac{4n}{3\theta}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{\sqrt{n}}{9} + 2\right]\log\frac{n+1}{\delta}.$$

We now summarize the result for the gradient approximation computed via (2.34), using the fact that $\lambda = \frac{1}{2\sqrt{n}}$.

Corollary 2.13. *Suppose that Assumption 1.3 holds, $n > 1$ and $g(x_k)$ is computed via (2.34) with N and σ satisfying*

$$N \geq \left[\frac{2n}{\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{n}{72} + \frac{8}{3} + \frac{4n}{3\theta}\frac{\sqrt{n}}{(\sqrt{n}-1)} + \frac{\sqrt{n}}{9}\right]\log\frac{n+1}{\delta},$$

$$\sqrt[3]{\frac{n\epsilon_f}{M}} \leq \sigma \leq \sqrt{\frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}M}}, \quad \|\nabla\phi(x)\| \geq \frac{4\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta},$$

then (1.3) holds with probability $1-\delta$.

2.5 Smoothing vs. Interpolation gradients

We now want to give some quick intuition explaining why GSG and BSG method do not provide as high accuracy as linear interpolation. Let us consider the two method of estimating gradients based on the same sample set. In particular, to compare GSG with linear interpolation, we choose the sample set $\mathcal{X} = \{x + \sigma u_1, x + \sigma u_2, \ldots, x + \sigma u_n\}$ for some $\sigma > 0$ with u obeying the standard Gaussian distribution. Recall the definition of the matrix $Q_{\mathcal{X}}$ and the vector $F_{\mathcal{X}}$ (see Section 2.2), and the fact that the gradient estimate computed by linear interpolation satisfies

$$Q_{\mathcal{X}} g_{LI} = F_{\mathcal{X}}/\sigma.$$

The GSG estimate, on the other hand is written as,

$$g_{GSG} = \frac{1}{n}Q_{\mathcal{X}}^T F_{\mathcal{X}}/\sigma = \frac{1}{n}Q_{\mathcal{X}}^T Q_{\mathcal{X}} g_{LI}.$$

Hence, we obtain

$$\|g_{LI} - g_{GSG}\| = \left\|\left(I - \frac{1}{n}Q_{\mathcal{X}}^T Q_{\mathcal{X}}\right) g_{LI}\right\|.$$

We know that, when $\epsilon(x) = 0$ for all $x \in \mathbb{R}^n$, the difference $\|g_{LI} - \nabla\phi(x)\|$ goes to zero as $\sigma \to 0$. However, $\|(I - \frac{1}{n}Q_{\mathcal{X}}^T Q_{\mathcal{X}})g_{LI}\|$ does not, as it does not depend on σ. While we have $\mathbb{E}[\frac{1}{n}Q_{\mathcal{X}}^T Q_{\mathcal{X}}] = I$, nevertheless, with non-negligible probability, the matrix $\|(I - \frac{1}{n}Q_{\mathcal{X}}^T Q_{\mathcal{X}})g_{LI}\| \geq \nu\|g_{LI}\|$ for some fixed non-negligible value of λ, for example, $\nu > 1/2$.

The intuition for the BSG can be derived in the same manner.

2.6 Summary of Results

In this section, we summarize the results for all methods. Specifically, Table 2 summarizes the conditions on N, σ and ρ for each method that we consider in this paper to guarantee $\|g(x) - \nabla\phi(x)\| \leq \rho$, and Table 3 summarizes the conditions on N, σ and $\nabla\phi(x)$ for each method that we consider in this paper to guarantee condition (1.3). Note that for the smoothing methods the bounds hold with probability $1-\delta$ and the number of samples depends on δ. From the tables, it is clear that for large n ($\frac{n}{(\sqrt{n}-1)^2}$ goes to 1 as $n \to \infty$), all methods have the same dependence (order of magnitude) on the dimension n; however, for the smoothing methods the constants in the bound can be significantly larger than those for deterministic methods, such as finite differences. This suggests that deterministic methods may be more efficient, at least in the setting considered in this paper, when accurate gradient estimates are desired. The bounds on the sampling radius are comparable for the smoothing and deterministic methods

Table 2: Bounds on N and σ that ensure $\|g(x) - \nabla\phi(x)\| \leq \rho$ (* denotes result is with probability $1 - \delta$).

Gradient Approximation	N	σ	ρ
Forward Finite Differences	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\rho}{\sqrt{n}L}\right]$	$2\sqrt{nL\epsilon_f}$
Central Finite Differences	n	$\left[\sqrt[3]{\frac{6\epsilon_f}{M}}, \sqrt{\frac{3\rho}{\sqrt{n}M}}\right]$	$\frac{2\sqrt{n}\sqrt[3]{M\epsilon_f^2}}{\sqrt[3]{6}}$
Linear Interpolation	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\rho}{\|Q_{\mathcal{X}}^{-1}\|\sqrt{n}L}\right]$	$2\|Q_{\mathcal{X}}^{-1}\|\sqrt{nL\epsilon_f}$
Gaussian Smoothed Gradients*	$\frac{3n\|\nabla\phi(x)\|^2}{\delta\rho^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+20)}{16\delta}$	$\left[\sqrt{\frac{\epsilon_f}{L}}, \frac{\rho}{6nL}\right]$	$6n\sqrt{L\epsilon_f}$
Centered Gaussian Smoothed Gradients*	$\frac{3n\|\nabla\phi(x)\|^2}{\delta\rho^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+30)}{144\delta}$	$\left[\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}}, \sqrt{\frac{\rho}{12n^{3/2}M}}\right]$	$12\sqrt[3]{n^{7/2}M\epsilon_f^2}$
Sphere Smoothed Gradients*	$\left[\left(\frac{2n\|\nabla\phi(x)\|^2}{\rho^2} + \frac{4\sqrt{n}\|\nabla\phi(x)\|}{3\rho}\right)\frac{n}{(\sqrt{n}-1)^2} + \frac{(3n+8\sqrt{n}+104)}{24}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt{\frac{n\epsilon_f}{L}}, \frac{\rho}{4\sqrt{n}L}\right]$	$4n\sqrt{L\epsilon_f}$
Centered Sphere Smoothed Gradients*	$\left[\left(\frac{2n\|\nabla\phi(x)\|^2}{\rho^2} + \frac{4\sqrt{n}\|\nabla\phi(x)\|}{3\rho}\right)\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+8\sqrt{n}+192)}{27}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt[3]{\frac{n\epsilon_f}{M}}, \sqrt{\frac{\rho}{4\sqrt{n}M}}\right]$	$4\sqrt[3]{n^{7/2}M\epsilon_f^2}$

Table 3: Bounds on N, σ and $\nabla\phi(x)$ that ensure $\|g(x) - \nabla\phi(x)\| \leq \theta\|\nabla\phi(x)\|$ (* denotes result is with probability $1 - \delta$).

Gradient Approximation	N	σ	$\|\nabla\phi(x)\|$
Forward Finite Differences	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{\sqrt{n}L}\right]$	$\frac{2\sqrt{nL\epsilon_f}}{\theta}$
Central Finite Differences	n	$\left[\sqrt[3]{\frac{6\epsilon_f}{M}}, \sqrt{\frac{3\theta\|\nabla\phi(x)\|}{\sqrt{n}M}}\right]$	$\frac{2\sqrt{n}\sqrt[3]{M\epsilon_f^2}}{\sqrt[3]{6}\theta}$
Linear Interpolation	n	$\left[2\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{\|Q_{\mathcal{X}}^{-1}\|\sqrt{n}L}\right]$	$\frac{2\|Q_{\mathcal{X}}^{-1}\|\sqrt{nL\epsilon_f}}{\theta}$
Gaussian Smoothed Gradients*	$\frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+20)}{16\delta}$	$\left[\sqrt{\frac{\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{6nL}\right]$	$\frac{6n\sqrt{L\epsilon_f}}{\theta}$
Centered Gaussian Smoothed Gradients*	$\frac{3n}{\delta\theta^2}\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+30)}{144\delta}$	$\left[\sqrt[3]{\frac{\epsilon_f}{\sqrt{n}M}}, \sqrt{\frac{\theta\|\nabla\phi(x)\|}{12n^{3/2}M}}\right]$	$\frac{12\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta}$
Sphere Smoothed Gradients*	$\left[\left(\frac{2n}{\theta^2} + \frac{4\sqrt{n}}{3\theta}\right)\frac{n}{(\sqrt{n}-1)^2} + \frac{(3n+8\sqrt{n}+104)}{24}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt{\frac{n\epsilon_f}{L}}, \frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}L}\right]$	$\frac{4n\sqrt{L\epsilon_f}}{\theta}$
Centered Sphere Smoothed Gradients*	$\left[\left(\frac{2n}{\theta^2} + \frac{4\sqrt{n}}{3\theta}\right)\frac{n}{(\sqrt{n}-1)^2} + \frac{(n+8\sqrt{n}+192)}{27}\right]\log\frac{n+1}{\delta}$	$\left[\sqrt[3]{\frac{n\epsilon_f}{M}}, \sqrt{\frac{\theta\|\nabla\phi(x)\|}{4\sqrt{n}M}}\right]$	$\frac{4\sqrt[3]{n^{7/2}M\epsilon_f^2}}{\theta}$

3 Numerical Results

In this section, we test our theoretical conclusions via numerical experiments. First, we present numerical results evaluating the quality of gradient approximations constructed via finite differences, linear interpolation, Gaussian smoothing and smoothing on a unit sphere (Section 3.1). We then illustrate the performance of a line search derivative-free optimization algorithm that employs the aforementioned gradient approximations on standard DFO benchmarking problems as well as on Reinforcement Learning tasks (Section 3.2).

3.1 Gradient Approximation Accuracy

We compare the numerical accuracy of the gradient approximations obtained by the methods discussed in Section 2. We compare the resulting θ, which is the relative error,

$$\frac{\|g(x) - \nabla\phi(x)\|}{\|\nabla\phi(x)\|}, \tag{3.1}$$

and report the average log of the relative error, i.e., $\log_{10}\theta$. Theory dictates that an optimization algorithm will converge if $\log_{10}\theta < \log_{10} 1/2 \approx -0.301$, namely $\theta < 1/2$, with sufficiently high probability; see [3].

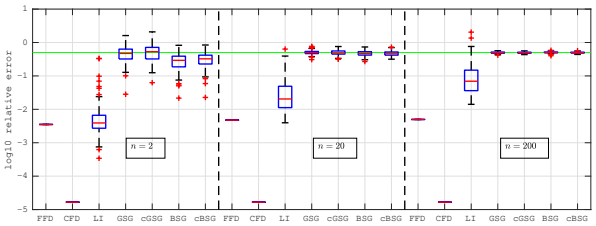
Gradient estimation on a synthetic function We first conduct tests on a synthetic function,

$$\phi(x) = \left(\sum_{i=1}^{n/2} M\sin(x_{2i-1}) + \cos(x_{2i})\right) + \frac{L-M}{2n}x^{\mathsf{T}}1_{n\times n}x, \tag{3.2}$$

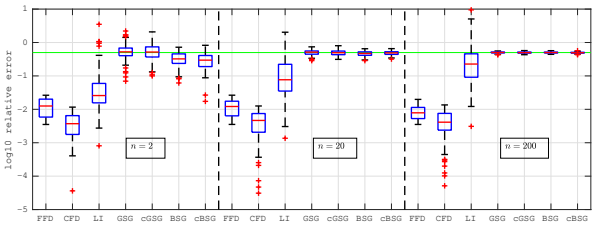
where n is an even number denoting the input dimension, $1_{n\times n}$ denotes an n by n matrix of all ones, and $L > M > 0$. We approximate the gradient of ϕ at the origin, for which $\|\nabla\phi(0)\| = \sqrt{\frac{n}{2}}M$. The Lipschitz constants for the first and second derivatives are L and $\max\{M, 1\}$, respectively. The function given in (3.2) allows us to vary all the moving components in the gradient approximations, namely, the dimension n, the Lipschitz constants L and M of the gradients and Hessians, respectively, the sampling radius σ and the size of the sample set N, in order to evaluate the different gradient approximation methods. We show results for two regimes: (1) the noise-free regime where $f(x) = \phi(x)$ (Figure 2, left column); and, (2) the noisy regime where $f(x) = \phi(x) + \epsilon(x)$ and $\epsilon(x) \sim U([-\epsilon_f, \epsilon_f])$ (Figure 2, right column).

We illustrate the performance of the gradient approximations using two sets (noise-free and noisy) of 5 box plots (Figure 2). The default values of the parameters are: $n = 20$, $M = 1$, $L = 2$, $\sigma = 0.01$, and $N = 4n$ (for the smoothing methods). For each box plot, we vary one of the parameters. Note, when comparing the relative errors for different values of M, the constant L is was set to $M+1$. For all randomized methods, including linear interpolation, $\nabla\phi(0)$ is estimated 100 times, i.e., we compute 100 realizations of $g(0)$. For linear interpolation, the directions $\{u_i\}_{i=1}^n$ are chosen as $u_i \sim \mathcal{N}(0, I)$ for all $i = 1, 2, \ldots, n$, and then normalized so that they lie in a unit ball $u_i \leftarrow u_i / \max_{j\in\{1,\ldots,n\}} \|u_j\|$. Moreover, all experiments in the noisy regime were conducted 100 times.

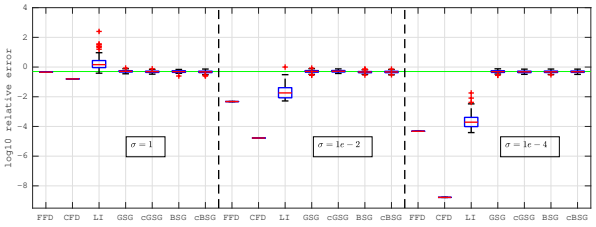
In accordance with our theory, we see in Figure 2a that the performance of most algorithms is not affected by the dimension n as long as the sampling radius and the number of sample points is chosen appropriately. The only method that is affected is interpolation; this is because as the dimension increases the matrix $Q_{\mathcal{X}}$ formed by the sampling directions (chosen randomly) may become more ill-conditioned. The effect of the dimension n becomes more apparent in the noisy regime; see Figure 2b. In Figure 2c, we observe that the size of σ, the sampling radius, has a significant effect on the deterministic methods (FFD and CFD) and LI. As predicted by the theory, in the noise-free setting, the gradient approximations improve as the sampling radius is reduced. For the randomized methods, GSG, cGSG, BSG and cBSG, in the noise-free setting, it appears that the sampling radius has no effect on the approximation quality. This is not surprising as our theory indicates that one of the terms in the error bound does not diminish with σ; see 2c. We should note that the randomized approximations are significantly worse than the approximations constructed by the deterministic methods in the noise-free regime. In the noisy regime, diminishing the sampling radius does not necessarily improve the approximations; see Figure 2d. This is predicted by the theory, as the error bounds have two terms, one that is diminishing with σ and one that is increasing with σ. In Figures 2e and 2f, we see that having more samples improves the accuracy achieved by GSG, cGSG, BSG and cBSG, in both the noise-free and noisy regimes. Finally, in Figures 2g, 2h, 2i and 2j, we see how the approximations are affected by changes in the Lipschitz constants. For example, the FFD, GSG and BSG approximations are affected by changes in L, whereas, the CFD cGSG and cBSG approximations are immune to these changes, but are affected by changes in M. All these effects are predicted by the theory. Note, in our experiments the FFD, GSG and BSG approximations are sensitive to changes in M, this is due to the fact that the constant L is linked to M ($L = M + 1$).



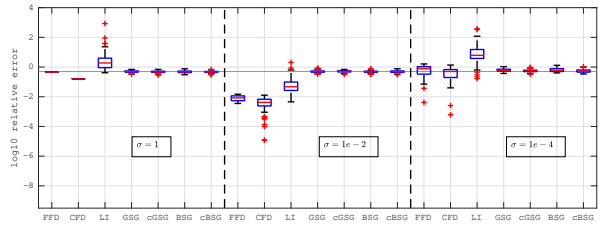
(a) Different n ($n \in \{2, 20, 200\}$).



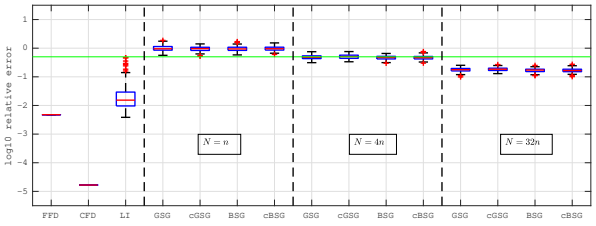
(b) Different n ($n \in \{2, 20, 200\}$).



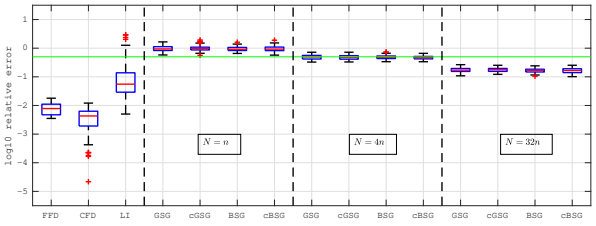
(c) Different σ ($\sigma \in \{10^0, 10^{-3}, 10^{-6}\}$).



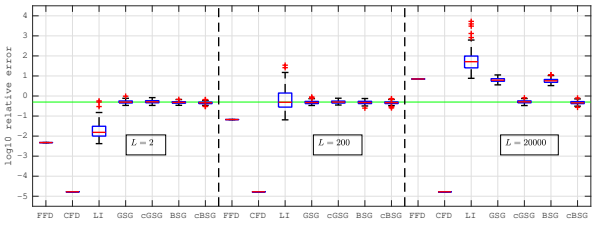
(d) Different σ ($\sigma \in \{10^0, 10^{-2}, 10^{-4}\}$).



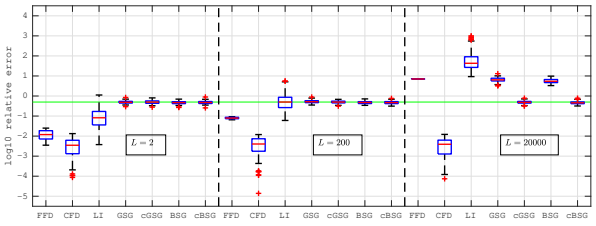
(e) Different N ($N \in \{n, 4n, 32n\}$, smoothing methods).



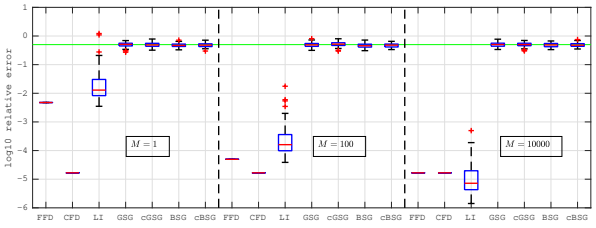
(f) Different N ($N \in \{n, 4n, 32n\}$, smoothing methods).



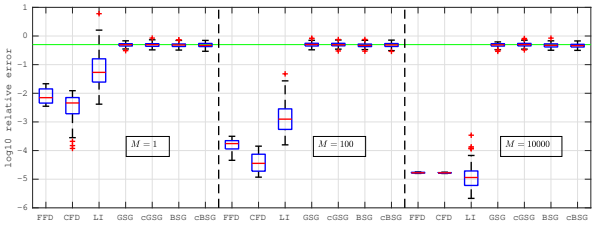
(g) Different L ($L \in \{2, 200, 20000\}$).



(h) Different L ($L \in \{2, 200, 20000\}$).



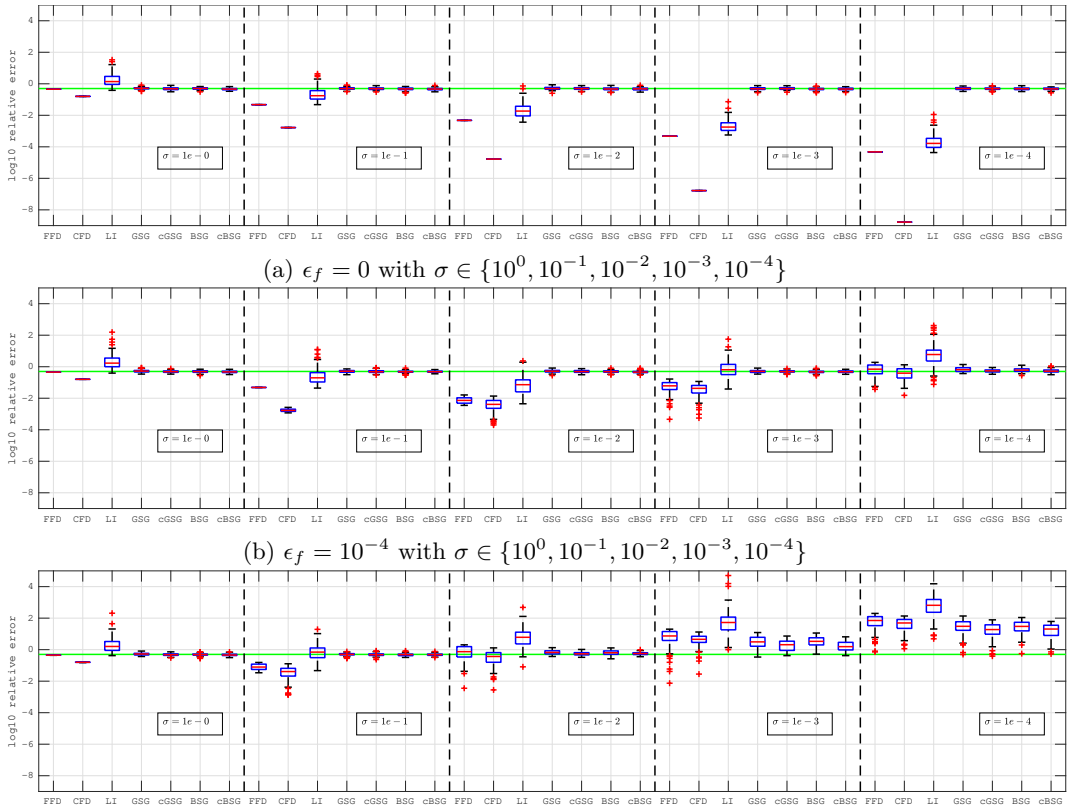
(i) Different M ($M \in \{1, 100, 10000\}$). Note, $L = M + 1$.



(j) different M ($M \in \{1, 100, 10000\}$). Note, $L = M + 1$.

Figure 2: Log of relative error (3.1) of gradient approximations (FFD, CFD, LI, GSG, cGSG, BSG, cBSG) with different n, σ, N, L and M. Left column: noise-free ($\epsilon_f = 0$); Right column: noisy (iid noise $U(-\epsilon_f, \epsilon_f)$ for each point).

In order to further illustrate the effects of noise ϵ_f and sampling radius σ, we ran experiments on the function given in (3.2) and varied these two parameters; see Figure 3. Each row illustrates results for a different noise level $\epsilon_f \in \{0, 10^{-4}, 10^{-2}\}$ for different sampling radii $\sigma \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. In the absence of noise (Figure 3a), as the sampling radius is reduced the approximations get better. As



(a) $\epsilon_f = 0$ with $\sigma \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

(b) $\epsilon_f = 10^{-4}$ with $\sigma \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

(c) $\epsilon_f = 10^{-2}$ with $\sigma \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$

Figure 3: Log of relative error (3.1) of gradient approximations (FFD, CFD, LI, GSG, cGSG, BSG, cBSG) with different σ. Top row: $\epsilon_f = 0$; Middle row: $\epsilon_f = 10^{-4}$; Bottom row: $\epsilon_f = 10^{-2}$.

predicted by the theory, this is not the case in the presence of noise (Figures 3b and 3c).

Gradient estimation on Schittkowski functions [31] Next, we test the different gradient approximations on the 69 functions from the Schittkowski test set [31]. The methods we compare are the same as in the case of the synthetic function. We computed the gradient approximations for a variety of points with diverse values for $\nabla\phi(x_k)$ and local Lipschitz constants L. For each problem we generated points by running gradient descent with a fixed step size for either 100 iterations or until the norm of the true gradient reached a value of 10^{-2}. Since for several problems the algorithm terminated in less than 100 iterations, the actual number of points we obtained was 5330.

Tables 4 and 5 summarize the results of these experiments for the noise-free and noisy ($\epsilon_f = 10^{-4}$) regimes, respectively. We show the average of the log of the relative error (3.1) for the 5330 points and the percentage of gradient estimates achieving $\theta < 1/2$ for different choices of σ, and, where appropriate, different choices of N. The values in bold indicate cases where the average of $\log_{10}\theta < \log_{10} 1/2$ or the percentage of gradient estimates achieving $\theta < 1/2$ is greater than 50%, repectively.

Table 4 illustrates the results in the noise-free regime. For these experiments, the sampling radius was chosen as $\sigma \in \{10^{-2}, 10^{-5}, 10^{-8}\}$. As predicted by the theory, in the noise-free case as the sampling radius decreases the quality of the approximations increases. This is true for all methods. We observe that for the smoothing methods more than $4n$ samples are needed to reliably obtain $\log_{10}\theta < \log_{10} 1/2 \approx -0.301$ (or $\theta < 1/2$). Moreover, this experiment indicates that the relative errors θ for FFD, CFD and LI methods are significantly smaller than those obtained by the smoothing methods.

Table 4: Average (Log) Relative Error of Gradient Approximations for 5330 Problems ($\epsilon_f = 0$).

Method	N	$\sigma = 10^{-2}$	$\sigma = 10^{-5}$	$\sigma = 10^{-8}$
FFD	n	-0.1651/42.68%	**-3.0124** / **95.10%**	**-5.7176** / **98.57%**
CFD	n	**-4.0112** / **93.41%**	**-8.4448** / **98.76%**	**-7.3651** / **98.57%**
LI	n	0.3808 / 27.64%	**-2.4616** / **91.44%**	**-5.0777** / **98.22%**
GSG	n	0.4067 / 4.05%	-0.0060 / 6.19%	-0.0425 / 7.11%
	$2n$	0.3108 / 8.01%	-0.1252 / 14.50%	-0.1754 / 15.91%
	$4n$	0.1790 / 24.39%	-0.2669 / 49.74%	**-0.3188** / **51.73%**
	$8n$	0.0477 / 45.82%	**-0.4117** / **84.00%**	**-0.4625** / **86.85%**
cGSG	n	0.0215 / 6.19%	-0.0435 / 6.90%	-0.0430 / 6.42%
	$2n$	-0.0983 / 14.80%	-0.1822 / 17.58%	-0.1723 / 15.89%
	$4n$	-0.2307 / 48.05%	**-0.3195** / **52.12%**	**-0.3163** / **51.16%**
	$8n$	**-0.3568** / **81.84%**	**-0.4665** / **87.28%**	**-0.4634** / **86.40%**
BSG	n	0.3478 / 6.21%	-0.0823 / 12.38%	-0.1192 / 12.23%
	$2n$	0.2033 / 15.59%	-0.2202 / 28.29%	-0.2609 / 29.55%
	$4n$	0.0544 / 38.46%	**-0.3649** / **67.37%**	**-0.4097** / **70.58%**
	$8n$	-0.0956 / **60.11%**	**-0.5163** / **93.62%**	**-0.5593** / **96.81%**
cBSG	n	-0.0503 / 10.38%	-0.1242 / 11.95%	-0.1258 / 12.36%
	$2n$	-0.1861 / 26.70%	-0.2677 / 30.19%	-0.2639 / 29.64%
	$4n$	**-0.3247** / **66.40%**	**-0.4109** / **70.00%**	**-0.4125** / **71.52%**
	$8n$	**-0.4625** / **91.52%**	**-0.5593** / **97.13%**	**-0.5677** / **96.94%**

Table 5 illustrates the performance of the gradient approximation in the presence of noise ($\epsilon_f = 10^{-4}$). Here the sampling radius was chosen as $\sigma \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. As in the noise-free regime, it appears that overall the gradient approximations computed via FFD, CFD and LI have smaller relative errors than those obtained by the smoothing methods. Moreover, as predicted by the theory in the noisy regime one needs to carefully select the sampling radius in order to achieve the smallest relative error.

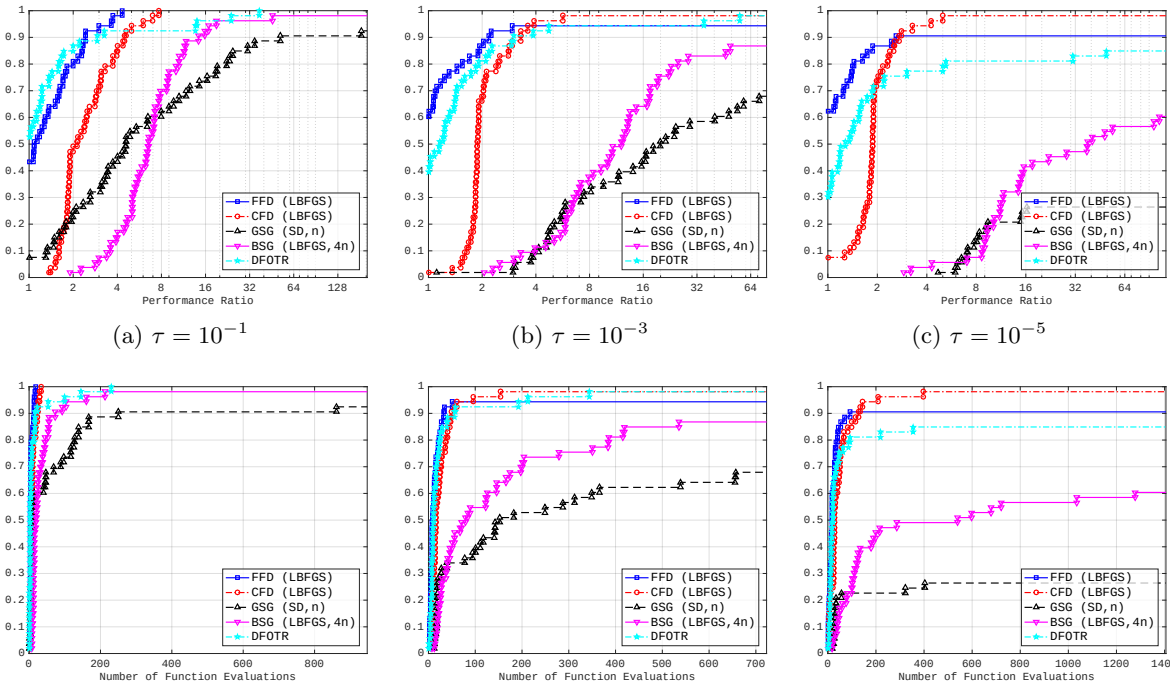
Table 5: Average (Log) Relative Error of Gradient Approximations for 5330 Problems ($\epsilon_f = 10^{-4}$).

Method	N	$\sigma = 10^{-1}$	$\sigma = 10^{-2}$	$\sigma = 10^{-3}$	$\sigma = 10^{-4}$
FFD	n	0.8593 / 12.03%	-0.0827 / 41.71%	**-0.5450** / **58.99%**	0.0724 / 31.26%
CFD	n	**-0.7297** / **62.61%**	**-1.7849** / **91.48%**	**-1.2902** / **80.56%**	**-0.3664** / 45.52%
LI	n	1.4604 / 8.37%	0.4718 / 24.86%	0.0841 / 38.07%	0.7335 / 21.33%
GSG	n	1.1284 / 1.67%	0.4105 / 4.73%	0.2262 / 4.97%	0.4954 / 3.08%
	$2n$	1.0574 / 2.57%	0.3085 / 8.39%	0.1052 / 11.09%	0.3728 / 7.94%
	$4n$	0.9970 / 7.49%	0.1888 / 22.70%	-0.0344 / 32.68%	0.2366 / 20.24%
	$8n$	0.8835 / 14.02%	0.0503 / 45.55%	-0.1686 / **62.55%**	0.0960 / 36.79%
cGSG	n	0.3144 / 4.37%	0.0178 / 6.62%	0.0178 / 6.42%	0.2783 / 4.26%
	$2n$	0.2472 / 10.19%	-0.0988 / 14.95%	-0.1151 / 14.33%	0.1446 / 11.07%
	$4n$	0.2049 / 28.99%	-0.2256 / 46.62%	-0.2499 / 42.27%	0.0054 / 26.21%
	$8n$	0.1441 / **52.51%**	**-0.3594** / **81.97%**	**-0.3891** / **76.68%**	-0.1341 / 47.35%
BSG	n	1.0705 / 1.93%	0.3460 / 6.42%	0.1848 / 7.90%	0.4919 / 4.62%
	$2n$	0.9383 / 5.07%	0.2016 / 16.12%	0.0381 / 20.36%	0.3473 / 11.52%
	$4n$	0.8119 / 12.53%	0.0541 / 38.03%	-0.1149 / 45.57%	0.1957 / 26.68%
	$8n$	0.6725 / 20.49%	-0.0954 / **59.81%**	-0.2603 /**71.31%**	0.0492 / 42.23%
cBSG	n	0.2210 / 7.95%	-0.0510 / 10.94%	-0.0422 / 9.91%	0.2565 / 7.67%
	$2n$	0.1311 / 16.85%	-0.1775 / 25.91%	-0.1833 / 24.50%	0.1093 / 17.02%
	$4n$	0.0369 / 42.20%	**-0.3149** / **64.20%**	**-0.3303** / **56.85%**	-0.0418 / 37.41%
	$8n$	-0.0582 / **63.60%**	**-0.4636** / **90.71%**	**-0.4754** / **86.30%**	-0.1877 / **54.18%**

3.2 Performance of Line Search DFO Algorithm with Different Gradient Approximations

The ability to approximate the gradient sufficiently accurately is a crucial ingredient of model based, and in particular line search, DFO algorithms. The numerical results presented in Section 3.1 illustrated the merits and limitations of the different gradient approximations. In this section, we investigate how these methods perform in conjunction with a line search DFO algorithm [3, Algorithm 1].

Moré & Wild Problems [20] Several algorithms could be considered in this section. We focus on line search DFO algorithms that either compute steepest descent search directions ($d_k = -g(x_k)$) or L-BFGS [23] search directions ($d_k = -H_k g(x_k)$). Moreover, we considered both adaptive line search variants as well as variants that used a constant, tuned step size parameter. Overall, we investigated the performance of 17 different algorithms. We considered algorithms that approximate the gradient using FFD, CFD and the four smoothing methods with steepest descent or L-BFGS search directions and an adaptive line search strategy. We also considered methods that approximate the gradient using the smoothing methods with steepest descent search directions and a constant step size parameter. Finally, as a benchmark, we compared the performance of the aforementioned methods against the popular DFOTR algorithm [1].



(a) $\tau = 10^{-1}$ (b) $\tau = 10^{-3}$ (c) $\tau = 10^{-5}$

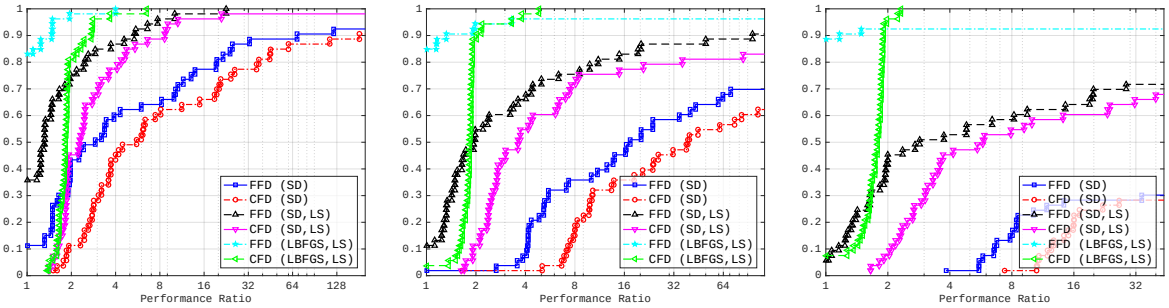
(d) $\tau = 10^{-1}$ (e) $\tau = 10^{-3}$ (f) $\tau = 10^{-5}$

Figure 4: Performance and data profiles for best variant of each method. Top row: performance profiles; Bottom row: data profiles.

We tested the algorithms on the problems described in [20] (53 problems), and illustrate the performance of the methods using performance and data profiles [14, 20]. We first compare the performance of the *best variant* of each gradient approximation for different accuracy levels (τ) (Figure 4). We selected only the best performers among different possible variants by first comparing the variants among themselves. For example, for FFD and CFD the LBFGS variant outperformed the steepest descent variant. With regards to the smoothing methods, GSG with $N = n$ samples per iteration and steepest descent search directions was the best performer out of all GSG methods, and BSG with $N = 4n$ and LBFGS performed best among all

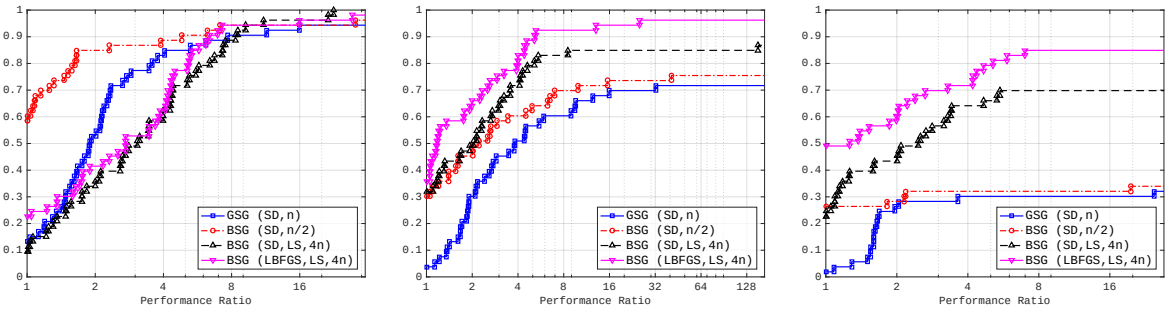
BSG variants. For all the types of gradient approximations, the variants that performed the best used an adaptive step length procedure. We omit illustrations of these comparison for brevity.

Next, we compare the adaptive step size methods against the constant step size ones (Figures 5 and 6).



(a) $\tau = 10^{-1}$ (b) $\tau = 10^{-3}$ (c) $\tau = 10^{-5}$

Figure 5: Performance profiles for Finite Difference variants with steepest descent (SD) and LBFGS search directions, and with and without a line search (LS).



(a) $\tau = 10^{-1}$ (b) $\tau = 10^{-3}$ (c) $\tau = 10^{-5}$

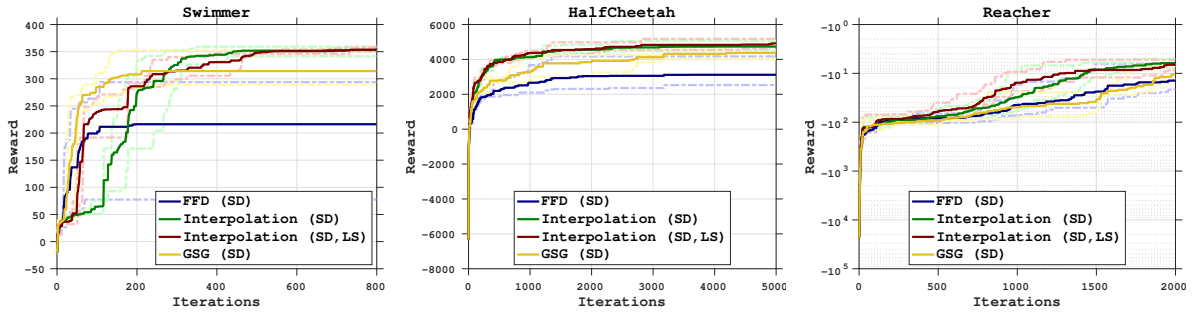
Figure 6: Performance profiles for best smoothed variants with steepest descent (SD) and LBFGS search directions, and with and without a line search (LS).

Reinforcement Learning Tasks [4] In this section, we investigate the performance of the methods on noisy optimization problems. Specifically, we present numerical results for reinforcement learning tasks from OpenAI Gym library [4]. We compare gradient based methods, where the gradients are approximated as follows:

1. Forward Finite Differences (`FFF (SD)`),
2. Linear Interpolation (`Interpolation (SD)`) and (`Interpolation (SD, LS)`),
3. Gaussian Smoothed Gradients (`GSG (SD)`).

For the methods that use interpolation, we implemented two different step length strategies: (1) fixed step length $\alpha_k = \alpha$, and (2) step length chosen via a line search.

In Figure 7, we show the average (solid lines) and max/min (dashed lines) over a number of runs. While our theory is the same for FD and Interpolation, our experiments show that for these tasks, choosing u_i to be orthonormal but random helps the algorithm to avoid getting stuck in local maxima. We observe that the Interpolation method is superior to the GSG and that line-search provides some improvements over a manually tuned choice of α_k. More details are given in the Appendix B.



(a) Swimmer (b) HalfCheetah (c) Reacher

Figure 7: Performance of Methods on Reinforcement Learning Tasks.

4 Final Remarks

We have shown that several derivative-free techniques for approximating gradients provide comparable estimates under reasonable assumptions. More specifically, we analyzed the gradient approximations constructed via finite differences, linear interpolation, Gaussian smoothing and smoothing on a unit sphere using functions values with bounded noise. For each method, we derived bounds on the number of samples and the sampling radius which guarantee favorable convergence properties for a line search or fixed step size descent method. These approximations can be used effectively in conjunction with a line search algorithm, possibly with L-BFGS search directions, provided they are sufficiently accurate. Our theoretical results, and related numerical experiments, show that finite difference and interpolation methods are much more efficient than smoothing methods in providing good gradient approximations.

References

[1] Afonso Bandeira, Katya Scheinberg, and Luis N Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Mathematical Programming, Series B*, 134:223–257, 2012.

[2] Albert S Berahas, Richard H Byrd, and Jorge Nocedal. Derivative-free optimization of noisy functions via quasi-newton methods. *SIAM Journal on Optimization*, 29(2):965–993, 2019.

[3] Albert S Berahas, Liyuan Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *arXiv preprint arXiv:1910.04055*, 2019.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

[5] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

[6] Richard G Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28(1):251–265, 1991.

[7] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, pages 1–39, 2018.

[8] Krzysztof Choromanski, Atil Iscen, Vikas Sindhwani, Jie Tan, and Erwin Coumans. Optimizing simulations with noise-tolerant structured exploration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2970–2977. IEEE, 2018.

[9] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard E Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. *arXiv preprint arXiv:1804.02395*, 2018.

[10] Andrew R Conn, Katya Scheinberg, and Philippe L Toint. On the convergence of derivative-free methods for unconstrained optimization. In A. Iserles and M. Buhmann, editors, *Approximation Theory and Optimization: Tributes to M. J. D. Powell*, pages 83–108, Cambridge, England, 1997. Cambridge University Press.

[11] Andrew R Conn, Katya Scheinberg, and Philippe L Toint. A derivative free optimization algorithm in practice. Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, Missouri, September 2-4, 1998.

[12] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. Geometry of interpolation sets in derivative free optimization. *Mathematical programming*, 111(1-2):141–172, 2008.

[13] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to Derivative-free Optimization*. MPS-SIAM Optimization series. SIAM, Philadelphia, USA, 2008.

[14] Elizabeth D Dolan and Jorge J Moré. Benchmarking Optimization Software with Performance Profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[15] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

[16] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[17] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[18] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3727–3737, 2018.

[19] Alvaro Maggiar, Andreas Wächter, Irina S Dolinskaya, and Jeremy 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(2):1478–1507, 2018.

[20] Jorge J Moré and Stefan M Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.

[21] Jorge J Moré and Stefan M Wild. Estimating computational noise. *SIAM Journal on Scientific Computing*, 33(3):1292–1314, 2011.

[22] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[23] Jorge Nocedal and Stephen J Wright. *Numerical Optimization, Second Edition*. Springer, 2006.

[24] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[25] Raghu Pasupathy, Peter Glynn, Soumyadip Ghosh, and Fatemeh S Hashemi. On sampling rates in simulation-based recursions. *SIAM Journal on Optimization*, 28(1):45–73, 2018.

[26] Valentin V Petrov. On lower bounds for tail probabilities. *Journal of statistical planning and inference*, 137(8):2703–2705, 2007.

[27] Michael J D Powell. Unconstrained minimization algorithms without computation of derivatives. *Bollettino delle Unione Matematica Italiana*, 9:60–69, 1974.

[28] Michael J D Powell. The NEWUOA software for unconstrained optimization without derivatives. In *Large-Scale Nonlinear Optimization*, volume 83, pages 255–297. Springer, US, 2006.

[29] Mark Rowland, Krzysztof Choromanski, François Chalus, Aldo Pacchiano, Tamas Sarlós, Turner Richard E, and Adrian Weller. Geometrically coupled monte carlo sampling. In *Advances in Neural Information Processing Systems*, pages 195–205, 2018.

[30] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. Technical Report arXiv:1703.03864, 2016.

[31] Klaus Schittkowski. *More test examples for nonlinear programming codes*, volume 282. Springer Science & Business Media, 2012.

[32] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[33] Sara Shashaani, Fatemeh S Hashemi, and Raghu Pasupathy. Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

[34] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in neural information processing systems*, pages 2899–2908, 2018.

[35] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

[36] Stefan M Wild, Rommel G Regis, and Christine A Shoemaker. ORBIT: optimization by radial basis function interpolation in trust-regions. *SIAM Journal on Scientific Computing*, 30(6):3197–3219, 2008.

[37] Xun Zhu and James C Spall. A modified second-order spsa optimization algorithm for finite samples. *International Journal of Adaptive Control and Signal Processing*, 16(5):397–409, 2002.

A Derivations

A.1 Derivation of (2.11)

$$
\begin{aligned}
\|\nabla F(x) - \nabla\phi(x))\| &= \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{1}{\sigma}f(x+\sigma u)u\right] - \nabla\phi(x)\right\| \\
&= \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)}{\sigma}u\right] - \nabla\phi(x)\right\| \\
&= \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\nabla\phi(x+\sigma u) - \nabla\phi(x)\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon(x+\sigma u)}{\sigma}u\right]\right\| \\
&\le \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\nabla\phi(x+\sigma u) - \nabla\phi(x)\right]\right\| + \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon(x+\sigma u)}{\sigma}u\right]\right\| \\
&\le \mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|\nabla\phi(x+\sigma u) - \nabla\phi(x)\|] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left\|\frac{\epsilon(x+\sigma u)}{\sigma}u\right\|\right] \\
&\le L\sigma\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|] + \frac{\epsilon_f}{\sigma}\mathbb{E}_{u\sim\mathcal{N}(0,I)}[\|u\|] \\
&= \left(L\sigma + \frac{\epsilon_f}{\sigma}\right)\sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \le \sqrt{n}L\sigma + \frac{\sqrt{n}\epsilon_f}{\sigma}.
\end{aligned}
$$

A.2 Derivation of (2.12)

$$
\begin{aligned}
&\|\nabla F(x) - \nabla\phi(x))\| \\
&= \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x+\sigma u)-\epsilon(x+\sigma u)}{2\sigma}u\right] - \nabla\phi(x)\right\| \\
&= \left\|\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{1}{2}\nabla\phi(x+\sigma u) + \frac{1}{2}\nabla\phi(x-\sigma u) - \nabla\phi(x)\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon(x+\sigma u)-\epsilon(x+\sigma u)}{2\sigma}u\right]\right\| \\
&\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\left(\nabla\phi(x+\sigma u) - \nabla\phi(x)\right) - \left(\nabla\phi(x) - \phi(x-\sigma u)\right)\|\right] \\
&\quad + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left\|\frac{\epsilon(x+\sigma u)-\epsilon(x+\sigma u)}{2\sigma}u\right\|\right] \\
&\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\left(\nabla\phi(x+\sigma u) - \nabla\phi(x)\right) - \left(\nabla\phi(x) - \phi(x-\sigma u)\right)\|\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon_f}{\sigma}\|u\|\right] \\
&= \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\left(\nabla^2\phi(x+\xi_1 u) - \nabla^2\phi(x-\xi_2 u)\right)\sigma u\|\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon_f}{\sigma}\|u\|\right],
\end{aligned}
$$

for some $0 \le \xi_1 \le \sigma$ and $0 \le \xi_2 \le \sigma$ by the intermediate value theorem. Then

$$
\begin{aligned}
\|\nabla F(x) - \nabla\phi(x))\| &\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|\nabla^2\phi(x+\xi_1 u) - \nabla^2\phi(x-\xi_2 u)\|\|\sigma u\|\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon_f}{\sigma}\|u\|\right] \\
&\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[M\|\xi_1 u + \xi_2 u\|\cdot\sigma\|u\|\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon_f}{\sigma}\|u\|\right] \\
&= \frac{1}{2}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[|\xi_1+\xi_2|\cdot\|u\|^2 M\sigma\right] + \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\frac{\epsilon_f}{\sigma}\|u\|\right] \\
&\le nM\sigma^2 + \frac{\sqrt{n}\epsilon_f}{\sigma}.
\end{aligned}
$$

A.3 Derivation of (2.19)

For the first equality, let $A = \mathrm{E}_{u\sim\mathcal{N}(0,I)}(a^\intercal u)^2 uu^\intercal$. Then for any $(i,j) \in \{1,2,\ldots,n\}^2$ with $i \neq j$, we have

$$\begin{aligned}
A_{ij} &= \mathrm{E}\left\{(a^\intercal u)^2 u_i u_j\right\} \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{E}\left\{a_k u_k a_l u_l u_i u_j\right\} \\
&= \sum_{k=i}\sum_{l=i} \mathrm{E}\left\{a_k u_k a_l u_l u_i u_j\right\} + \sum_{k\neq i}\sum_{l=i} \mathrm{E}\left\{a_k u_k a_l u_l u_i u_j\right\} + \sum_{k=i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l u_i u_j\right\} + \sum_{k\neq i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l u_i u_j\right\} \\
&= \mathrm{E}\left\{a_i^2 u_i^3 u_j\right\} + \sum_{k\neq i} \mathrm{E}\left\{a_k a_i u_k u_i^2 u_j\right\} + \sum_{l\neq i} \mathrm{E}\left\{a_i a_l u_l u_i^2 u_j\right\} + \mathrm{E}\left\{u_i\right\} \sum_{k\neq i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l u_j\right\} \\
&= 0 + \sum_{k\neq i} \mathrm{E}\left\{a_k a_i u_k u_i^2 u_j\right\} + \sum_{l\neq i} \mathrm{E}\left\{a_i a_l u_l u_i^2 u_j\right\} + 0 \\
&= \mathrm{E}\left\{a_i a_j u_i^2 u_j^2\right\} + \mathrm{E}\left\{a_i a_j u_i^2 u_j^2\right\} \\
&= 2a_i a_j.
\end{aligned}$$

For any $i \in \{1,2,\ldots,n\}$,

$$\begin{aligned}
A_{ii} &= \mathrm{E}\left\{(a^\intercal u)^2 u_i^2\right\} \\
&= \sum_{k=i}\sum_{l=i} \mathrm{E}\left\{a_k u_k a_l u_l u_i^2\right\} + \sum_{k\neq i}\sum_{l=i} \mathrm{E}\left\{a_k u_k a_l u_l u_i^2\right\} + \sum_{k=i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l u_i^2\right\} + \sum_{k\neq i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l u_i^2\right\} \\
&= \mathrm{E}\left\{a_i^2 u_i^4\right\} + \sum_{k\neq i} \mathrm{E}\left\{a_k a_i u_k u_i^3\right\} + \sum_{l\neq i} \mathrm{E}\left\{a_i a_l u_l u_i^3\right\} + \mathrm{E}\left\{u_i^2\right\} \sum_{k\neq i}\sum_{l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l\right\} \\
&= 3a_i^2 + 0 + 0 + 1 \times \sum_{k=l\neq i} \mathrm{E}\left\{a_k u_k a_l u_l\right\} = 3a_i^2 + \sum_{k\neq i} \mathrm{E}\left\{a_k^2 u_k^2\right\} \\
&= 3a_i^2 + \sum_{k\neq i} a_k^2 = 2a_i^2 + \sum_{k=1}^{n} a_k^2.
\end{aligned}$$

Then by writting the result in matrix format, we get $\mathrm{E}_{u\sim\mathcal{N}(0,I)}\left[(a^\intercal u)^2 uu^\intercal\right] = a^\intercal a I + 2aa^\intercal$.

For the second equality, since the possibility density function of $\mathcal{N}(0,I)$ is even while $a^\intercal u \cdot \|u\|^k \cdot uu^\intercal$ is an odd function, the expectation $\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[a^T u \cdot \|u\|^k \cdot uu^T\right]$ is zero.

Because $\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k u^\intercal u\right] = \mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^{k+2}\right]$ is the $(k+2)$nd moment of a Chi distributed variable for all $k \in \mathbb{N}$, we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k u^\intercal u\right] = \frac{2^{1+k/2}\Gamma((n+k+2)/2)}{\Gamma(n/2)}.$$

This value is also the trace of the matrix $\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k uu^T\right]$. Considering all n elements on the diagonal of this matrix are the same, we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k uu^T\right] = \frac{2^{1+k/2}\Gamma((n+k+2)/2)}{n\Gamma(n/2)} I \text{ for } k = 0,1,2,\ldots.$$

For even k, this quantity is equal to $\prod_{i=1}^{k/2}(n+2i)$. For odd k, this quantity is equal to $\left[\sqrt{2}\Gamma\left(\frac{n+1}{2}\right)/\Gamma\left(\frac{n}{2}\right)\right]\frac{1}{n}\prod_{i=1}^{(k+1)/2}(n+2i-1)$. Use the inequality $\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)/\Gamma\left(\frac{n}{2}\right) \leq \sqrt{n}$ for all $n \in \mathbb{N}$, we have

$$\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\|u\|^k uu^T\right] \preceq (n+1)(n+3)\cdots(n+k)\cdot n^{-0.5} I \text{ for } k = 1,3,5,\ldots.$$

A.4 Derivation of (2.29)

$$\begin{aligned}
\mathbb{E}\left[\|g(x)-\nabla F(x)\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{f(x+\sigma u_i)-f(x)}{\sigma}u_i-\nabla F(x)\right\|^2\right] \\
&= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2 u^\intercal u\right]-\frac{1}{N}\nabla F(x)^\intercal\nabla F(x) \\
&= \frac{1}{N}\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[(a^\intercal u)^2\,u^\intercal u\right]-\frac{1}{N}a^\intercal a \\
&= \frac{1}{N}(n+1)a^\intercal a.
\end{aligned}$$

A.5 Derivation of (2.30)

The expression for $E\left[\|g(x)-\nabla F(x)\|^4\right]$ is a sum of N^4 terms with each term being the product of four vectors:

$$\mathbb{E}\left[\|g(x)-\nabla F(x)\|^4\right] = \frac{1}{N^4}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\prod_{w\in\{i,j,k,l\}}\left(\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right)\right],$$

where $\prod$ denotes the operation which is a product of the inner products of the two pairs of vectors. Specifically, given four vectors $a_1,a_2,a_3,a_4\in\mathbb{R}^n$, $\prod_{i\in\{1,2,3,4\}}a_i=(a_1^\intercal a_2)\cdot(a_3^\intercal a_4)$ and $\prod_{i\in\{1,1,2,2\}}a_i=(a_1^\intercal a_1)\cdot(a_2^\intercal a_2)$.

We first observe that $\prod_{w\in\{i,j,k,l\}}\left(\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right)=0$ whenever one of the indices (i,j,k,l) is different from all of the other ones. This is because all u_w, for $w\in\{i,j,k,l\}$ are independent of each other if their indices are different and

$$\mathbb{E}\left[\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right]=0.$$

Thus we need only to consider the terms having one of the following conditions:

1. $i=j=k=l$;
2. $i=j\neq k=l$;
3. $i=k\neq j=l$;
4. $i=l\neq j=k$.

First we consider the case: $i=j\neq k=l$, which occurs when $N>1$.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\prod_{w\in\{i,i,k,k\}}\left(\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right)\right] \\
&\quad=\sum_{i=1}^{N}\mathbb{E}\left[\left\|\frac{f(x+\sigma u_i)-f(x)}{\sigma}u_i-\nabla F(x)\right\|^2\right]\cdot\sum_{k=1,k\neq i}^{N}\mathbb{E}\left[\left\|\frac{f(x+\sigma u_k)-f(x)}{\sigma}u_k-\nabla F(x)\right\|^2\right] \\
&\quad=N(N-1)\left[(n+1)a^\intercal a\right]^2.
\end{aligned}$$

We now consider two other cases: $i = k \neq j = l$ and $i = l \neq j = k$ that are essentially the same. We have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\prod_{w\in\{i,k,i,k\}}\left(\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right)\right] \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\mathbb{E}\left\{\left[\left(\frac{f(x+\sigma u_i)-f(x)}{\sigma}u_i-\nabla F(x)\right)^{\intercal}\left(\frac{f(x+\sigma u_k)-f(x)}{\sigma}u_k-\nabla F(x)\right)\right]^2\right\} \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\mathbb{E}\left(\left\{\left[(a^{\intercal}u_i)u_i-a\right]^{\intercal}\left[(a^{\intercal}u_k)u_k-a\right]\right\}^2\right) \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\mathbb{E}\left(\left[(a^{\intercal}u_i)(a^{\intercal}u_k)(u_i^{\intercal}u_k)-(a^{\intercal}u_i)^2-(a^{\intercal}u_k)^2+a^{\intercal}a\right]^2\right) \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\mathbb{E}\left[\begin{array}{ll} & (a^{\intercal}u_i)^2(a^{\intercal}u_k)^2(u_i^{\intercal}u_k)^2+(a^{\intercal}u_i)^4+(a^{\intercal}u_k)^4+(a^{\intercal}a)^2 \\ + & 2(a^{\intercal}a)(a^{\intercal}u_i)(a^{\intercal}u_k)(u_i^{\intercal}u_k)-2(a^{\intercal}a)(a^{\intercal}u_i)^2-2(a^{\intercal}a)(a^{\intercal}u_k)^2 \\ - & 2(a^{\intercal}u_i)^3(a^{\intercal}u_k)(u_i^{\intercal}u_k)-2(a^{\intercal}u_i)(a^{\intercal}u_k)^3(u_i^{\intercal}u_k)+2(a^{\intercal}u_i)^2(a^{\intercal}u_k)^2\end{array}\right] \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}\left[\begin{array}{ll} & (n+8)(a^{\intercal}a)^2+3(a^{\intercal}a)^2+3(a^{\intercal}a)^2+(a^{\intercal}a)^2 \\ + & 2(a^{\intercal}a)^2-2(a^{\intercal}a)^2-2(a^{\intercal}a)^2 \\ - & 6(a^{\intercal}a)^2-6(a^{\intercal}a)^2+2(a^{\intercal}a)^2\end{array}\right] \\
&\quad=\sum_{i=1}^{N}\sum_{k=1,k\neq i}^{N}(n+3)(a^{\intercal}a)^2=N(N-1)(n+3)(a^{\intercal}a)^2
\end{aligned}
$$

Finally, we have the $i = j = k = l$ case:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{N}\prod_{w\in\{i,i,i,i\}}\left(\frac{f(x+\sigma u_w)-f(x)}{\sigma}u_w-\nabla F(x)\right)\right] \\
&\quad=N\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\left\|\frac{f(x+\sigma u)-f(x)}{\sigma}u-\nabla F(x)\right\|^4\right] \\
&\quad=N\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left\{\left[\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2u^{\intercal}u-2\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)u^{\intercal}\nabla F(x)+\nabla F(x)^{\intercal}\nabla F(x)\right]^2\right\} \\
&\quad=N\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\begin{array}{ll} & \left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^4(u^{\intercal}u)^2+4\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2(u^{\intercal}\nabla F(x))^2 \\ + & (\nabla F(x)^{\intercal}\nabla F(x))^2-4\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^3(u^{\intercal}u)\,(u^{\intercal}\nabla F(x)) \\ - & 4\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)(u^{\intercal}\nabla F(x))(\nabla F(x)^{\intercal}\nabla F(x)) \\ + & 2\left(\frac{f(x+\sigma u)-f(x)}{\sigma}\right)^2(u^{\intercal}u)(\nabla F(x)^{\intercal}\nabla F(x))\end{array}\right] \\
&\quad=N\mathbb{E}_{u\sim\mathcal{N}(0,I)}\left[\begin{array}{ll} & (a^{\intercal}u)^4\,(u^{\intercal}u)^2+4\,(a^{\intercal}u)^2\,(u^{\intercal}a)^2+(a^{\intercal}a)^2-4\,(a^{\intercal}u)^3\,(u^{\intercal}u)\,(u^{\intercal}a) \\ - & 4\,(a^{\intercal}u)\,(u^{\intercal}a)(a^{\intercal}a)+2\,(a^{\intercal}u)^2\,(u^{\intercal}u)(a^{\intercal}a)\end{array}\right] \\
&\quad=N\left[\begin{array}{ll} & 3(n+4)(n+6)(a^{\intercal}a)^2+12(a^{\intercal}a)^2+(a^{\intercal}a)^2-12(n+4)(a^{\intercal}a)^2 \\ - & 4(a^{\intercal}a)^2+2(n+2)(a^{\intercal}a)^2\end{array}\right] \\
&\quad=N(3n^2+20n+37)(a^{\intercal}a)^2
\end{aligned}
$$

In summary, we have

$$\begin{aligned}
&N^4\mathbb{E}\left[\|g(x)-\nabla F(x)\|^4\right]\\
&\quad= N(N-1)(n+1)^2(a^\intercal a)^2+2N(N-1)(n+3)(a^\intercal a)^2+N(3n^2+20n+37)(a^\intercal a)^2\\
&\quad= N(N-1)(n^2+4n+7)(a^\intercal a)^2+N(3n^2+20n+37)(a^\intercal a)^2.
\end{aligned}$$

A.6 Derivation of (2.36)

$$\begin{aligned}
\|\nabla F(x)-\nabla\phi(x))\| &= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{\sigma}f(x+\sigma u)u\right]-\nabla\phi(x)\right\|\\
&= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{\sigma}(\phi(x+\sigma u)+\epsilon(x+\sigma u))u\right]-\nabla\phi(x)\right\|\\
&= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\nabla\phi(x+\sigma u)-\nabla\phi(x)\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon(x+\sigma u)}{\sigma}u\right]\right\|\\
&\le \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\nabla\phi(x+\sigma u)-\nabla\phi(x)\right]\right\|+\left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon(x+\sigma u)}{\sigma}u\right]\right\|\\
&\le \mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}[\|\nabla\phi(x+\sigma u)-\nabla\phi(x)\|]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left\|\frac{n\epsilon(x+\sigma u)}{\sigma}u\right\|\right]\\
&\le L\sigma\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}[\|u\|]+\frac{n\epsilon_f}{\sigma}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}[\|u\|]\\
&= L\sigma\frac{n}{n+1}+\frac{n\epsilon_f}{\sigma}\le L\sigma+\frac{n\epsilon_f}{\sigma}.
\end{aligned}$$

A.7 Derivation of (2.37)

$$\begin{aligned}
&\|\nabla F(x)-\nabla\phi(x))\|\\
&\quad= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{2\sigma}(\phi(x+\sigma u)+\epsilon(x+\sigma u)-\phi(x+\sigma u)-\epsilon(x+\sigma u))u\right]-\nabla\phi(x)\right\|\\
&\quad= \left\|\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\frac{1}{2}\nabla\phi(x+\sigma u)+\frac{1}{2}\nabla\phi(x-\sigma u)-\nabla\phi(x)\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n}{2\sigma}(\epsilon(x+\sigma u)-\epsilon(x+\sigma u))u\right]\right\|\\
&\quad\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\|\left(\nabla\phi(x+\sigma u)-\nabla\phi(x)\right)-\left(\nabla\phi(x)-\phi(x-\sigma u)\right)\|\right]\\
&\qquad+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\left\|\frac{n}{2\sigma}(\epsilon(x+\sigma u)-\epsilon(x+\sigma u))u\right\|\right]\\
&\quad\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\|\left(\nabla\phi(x+\sigma u)-\nabla\phi(x)\right)-\left(\nabla\phi(x)-\phi(x-\sigma u)\right)\|\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon_f}{\sigma}\|u\|\right]\\
&\quad= \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\|\left(\nabla^2\phi(x+\xi_1u)-\nabla^2\phi(x-\xi_2u)\right)\sigma u\|\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon_f}{\sigma}\|u\|\right],
\end{aligned}$$

for some $0\le\xi_1\le\sigma$ and $0\le\xi_2\le\sigma$ by the intermediate value theorem. Then

$$\begin{aligned}
\|\nabla F(x)-\nabla\phi(x))\| &\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[\|\nabla^2\phi(x+\xi_1u)-\nabla^2\phi(x-\xi_2u)\|\|\sigma u\|\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon_f}{\sigma}\|u\|\right]\\
&\le \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[M\|\xi_1u+\xi_2u\|\cdot\sigma\|u\|\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon_f}{\sigma}\|u\|\right]\\
&= \frac{1}{2}\mathbb{E}_{u\sim\mathcal{U}(\mathcal{B}(0,1))}\left[|\xi_1+\xi_2|\cdot\|u\|^2M\sigma\right]+\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\frac{n\epsilon_f}{\sigma}\|u\|\right]\\
&\le M\sigma^2+\frac{n\epsilon_f}{\sigma}.
\end{aligned}$$

A.8 Derivation of (2.40)

The first and third equalities of (A.8) comes from the first and third equalities of (2.19). Considering any vector of iid Gaussian v, dividing by its own norm, can be expressed as $v = \|v\|u$. Moreover, $\|v\|$ and u are independent. Thus any homogeneous polynomial p in the entries of u of degree k has the property that

$$\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}[p(u)] = \frac{\mathbb{E}_{v\sim\mathcal{N}(0,I)}[p(v)]}{\mathbb{E}_{v\sim\mathcal{N}(0,I)}\|v\|^k}.$$

Then

$$\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[(a^Tu)^2uu^T\right] = \frac{\mathbb{E}_{u\sim\mathcal{N}(0,I))}\left[(a^Tu)^2uu^T\right]}{\mathbb{E}_{u\sim\mathcal{N}(0,I))}\|u\|^4} = \frac{a^TaI + 2aa^T}{n(n+2)}$$

$$\mathbb{E}_{u\sim\mathcal{U}(\mathcal{S}(0,1))}\left[\|u\|^kuu^T\right] = \frac{\mathbb{E}_{u\sim\mathcal{N}(0,I))}\left[\|u\|^kuu^T\right]}{\mathbb{E}_{u\sim\mathcal{N}(0,I))}\|u\|^{k+2}} = \frac{1}{n}I.$$

The second equality of (2.40) being 0 follows the same argument as that for the second equality of (2.19).

B Additional Details: RL Experiments

In all RL experiments the blackbox function f takes as input the parameters of the policy $\pi_\theta : \mathcal{S} \rightarrow \mathcal{A}$ which maps states to proposed actions. The output of f is the total reward obtained by an agent applying that particular policy π_θ in the given environment.

To encode policies π_θ, we used fully-connected feedforward neural networks with two hidden layers, each of $h = 41$ neurons and with tanh nonlinearities. The matrices of connections were encoded by low-displacement rank neural networks (see [9]), as in several recent papers on applying orthogonal directions in gradient estimation for ES methods in reinforcement learning. We did not apply any additional techniques such as state/reward renormalization, ranking or filtering, in order to solely focus on the evaluation of the presented proposals.

All experiments were run with hyperparameter $\sigma = 0.1$. Experiments that did not apply line search were run with the use of Adam optimizer and $\alpha = 0.01$. For line search experiments, we were using adaptive α that was updated via Armijo condition with Armijo parameter $c_1 = 0.2$ and backtracking factor $\tau = 0.3$.

Finally, in order to construct orthogonal samples, at each iteration we were conducting orthogonalization of random Gaussian matrices with entries taken independently at random from $\mathcal{N}(0, 1)$ via Gram-Schmidt procedure (see [9]). Instead of the orthogonalization of Gaussian matrices, we could take advantage of constructions, where orthogonality is embedded into the structure (such as random Hadamard matrices from [9]), introducing extra bias but proven to work well in practice. However in all conducted experiments that was not necessary.

For each environment and each method we run $k = 3$ experiments corresponding to different random seeds.