Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-021-01994-w

A Mixed Finite Differences Scheme for Gradient Approximation

Marco Boresta[1] · Tommaso Colombo[1] · Alberto De Santis[1] · Stefano Lucidi[1]

Received: 17 November 2020 / Accepted: 26 December 2021

Abstract

In this paper, we focus on the linear functionals defining an approximate version of the gradient of a function. These functionals are often used when dealing with optimization problems where the computation of the gradient of the objective function is costly or the objective function values are affected by some noise. These functionals have been recently considered to estimate the gradient of the objective function by the expected value of the function variations in the space of directions. The expected value is then approximated by a sample average over a proper (random) choice of sample directions in the domain of integration. In this way, the approximation error is characterized by statistical properties of the sample average estimate, typically its variance. Therefore, while useful and attractive bounds for the error variance can be expressed in terms of the number of function evaluations, nothing can be said on the error of a single experiment that could be quite large. This work instead is aimed at deriving an approximation scheme for linear functionals approximating the gradient, whose error of approximation can be characterized by a deterministic point of view in the case of noise-free data. The previously mentioned linear functionals are no longer considered as expected values over the space of directions, but rather as the filtered derivative of the objective function by a Gaussian kernel. By using this new approach, a gradient estimation based on a suitable linear combination of central

Communicated by Gianni Di Pillo.

✉ Alberto De Santis
desantis@diag.uniroma1.it

Marco Boresta
boresta@diag.uniroma1.it

Tommaso Colombo
colombo@diag.uniroma1.it

Stefano Lucidi
lucidi@diag.uniroma1.it

1 Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy

Published online: 18 February 2022

finite differences at different step sizes is proposed and deterministic bounds that do not depend on the particular sample of points considered are computed. In the noisy setting, on the other end, the variance of the estimation error of the proposed method is showed to be strictly lower than the one of the estimation error of the Central Finite Difference scheme. Numerical experiments on a set of test functions are encouraging, showing good performances compared to those of some methods commonly used in the literature, also in the noisy setting.

Keywords Gradient approximation · Filtered derivative · Derivative free optimization

1 Introduction

DFO algorithms have become increasingly important since they provide a proper methodology to tackle most of the optimization problems considered in various fields of application. As reported in [4,8,16], typical applications fall within the simulation-based optimization problems such as policy optimization in reinforcement learning. DFO methods arise when derivative information is either unavailable, or quite costly to obtain, not to mention when only noisy sample of the objective function are available. In the latter case, it is known that most methods based on finite difference are of little use [11,19].

One of the approaches in DFO algorithms is that of computing a proper estimate of the gradient of the objective function. Finite difference approximation schemes were already present in early times [15] and have recently been reconsidered as sample average approximations of functionals defining a "filtered version" of the objective function [2,3,9,13]. These functionals arise when defining a gradient approximation as the average of the function variation along all the directions in the whole space. In the most popular methods, the average is performed by weighting the function variations along directions generated either with a uniform kernel on the unit ball [9], or with a Gaussian kernel [2]. These integrals are considered as ensemble averages over the space of the directions of differentiation, and then are approximated by sample averages over a random sample of directions, with various methods. As a general policy, the approximation error is then characterized by its statistical properties (even in the noise-free setting), the variance is expressed in terms of the number of function calculations, and nice bounds are provided to trade-off precision of the gradient estimation and computational costs. Nevertheless it is plain that the error on a single sample may be quite large, even though its variance is bounded.

In this paper, we focus on a different point of view. The functional defining a filtered version of the objective function is considered as weak derivative of the objective function rather than expected values over the space of the directions [20]. The gradient estimation is therefore obtained by considering a numerical approximation of the functional integral, and the estimation error is evaluated in a deterministic fashion. The estimate is obtained by a suitable linear combination of central finite differences at steps with increasing size. Bounds on the approximation error with the proposed method are derived, and the variance of the error in the case of noisy data is also presented.

The goodness of the approximation is experimentally evaluated by comparing the proposed method with those considered benchmarks by the literature—namely: Forward Finite Differences (FFD), Central Finite Differences (CFD) [15], Gaussian Smoothed Gradient (GSG), Central Gaussian Smoothed Gradient (cGSG) [9,13]—over the benchmark of the Schittkowski functions [17]. Encouraging results are obtained, both in the noise-free and in the noisy setting.

The paper is organized as follows: Sect. 2 formally introduces the gradient estimation problem, highlighting the difference between the approach proposed in this article and the one of several estimates proposed in the literature. In Sect. 3, we present the proposed approximation scheme—NMXFD, with an emphasis on its link with the Finite Difference Method. A theoretical comparison between the variance of the estimation errors of the proposed method and of the CFD scheme is proposed in Sect. 4. Section 5 presents numerical results and conclusions are drawn in Sect. 6.

2 The Gradient Estimate

In this paper, we consider the following unconstrained optimization problem in the derivative free optimization (DFO) setting [6,12]:

$$\min_{x \in R^n} f(x), \tag{1}$$

where $f : R^n \mapsto R$ is a function with continuous derivative, i.e., $f \in \mathcal{C}^1(R^n)$, and we denote the gradient $\nabla f : R^n \mapsto R^n$ such that for any $x \in R^n$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}.$$

In this section, the problem of a numerical approximation of the gradient $\nabla f(x)$ is considered. The most popular approximation scheme is the standard finite difference method [15], but interesting alternative schemes are proposed in papers [2,9]. A general estimate is obtained according to the following formula:

$$G_\sigma(x) := \frac{1}{\sigma} \int_{R^n} f(x + \sigma s)\, s\, \varphi(s)\, \mathrm{d}s, \tag{2}$$

where $\varphi(s) : R^n \mapsto R$ denotes either a standard Gaussian Kernel $\mathcal{N}(0, I_n)$ or a uniform kernel on the unit ball $\mathcal{B}(0, 1)$, $\mathrm{d}s = \mathrm{d}s_1 \cdot \mathrm{d}s_2 \cdot \cdots \cdot \mathrm{d}s_n$ is the volume element in R^n, and $\sigma > 0$ is a scale parameter. The approximation error has different bounds depending on the assumptions on f (see [4]). If the function f is continuously differentiable, and its gradient is L-Lipschitz continuous for all $x \in R^n$, then

$$||G_\sigma(x) - \nabla f(x)|| \le C_\varphi L \sigma, \tag{3}$$

where C_φ is a positive constant whose value depends on the kernel. If the function f is twice continuously differentiable, and its Hessian is H-Lipschitz continuous for all $x \in R^n$, then

$$||G_\sigma(x) - \nabla f(x)|| \leq C_\varphi H \sigma^2. \tag{4}$$

Both bounds (3) and (4) show that

$$\lim_{\sigma \to 0} G_\sigma(x) = \nabla f(x).$$

We will now work out formula (2) considering the (standard) Gaussian kernel

$$\varphi(s) \sim \mathcal{N}(0, I_n) = \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} {s_i}^2\right\} = \prod_{i=1}^{n} \varphi(s_i) \tag{5}$$

but the considerations that follow hold also if a uniform kernel over the unit ball is considered.

Let us consider this further notation: for any $x \in R^n$ denote by $\bar{x}_i \in R^{n-1}$ the following vector $\left[x_1, x_2 \ \ldots \ , x_{i-1}, x_{i+1}, \ \ldots \ , x_n\right]^{\mathrm{T}}$. With some abuse of notation, but for sake of simplicity in the use of formulas, when addressing a given coordinate x_i in a vector x let us write x as $[x_i \ \bar{x}_i]^{\mathrm{T}}$ and denote $f(x)$ as $f(x_i, \bar{x}_i)$ and $\varphi(s) = \varphi(s_i)\varphi(\bar{s}_i)$, with $\varphi(\bar{s}_i) = \prod_{j \neq i}^{n} \varphi(s_j)$; consistently, the volume element becomes $\mathrm{d}s = \mathrm{d}s_i \cdot \mathrm{d}\bar{s}_i$. In case of a vector function $f(z)$, to address explicitly its $i-th$ entry we write it as $[(f(z))_i \ \overline{(f(z))_i}]^{\mathrm{T}}$. Then, estimate (2) is rewritten as follows

$$G_\sigma(x) = \frac{1}{\sigma}\int_{R^n} f(x_1 + \sigma s_1, \ldots, x_n + \sigma s_n) \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix} \prod_{i=1}^{n} \varphi(s_i)\, \mathrm{d}s \tag{6}$$

$$= \begin{bmatrix} \frac{1}{\sigma}\int_{R^n} f(x_1 + \sigma s_1, \bar{x}_1 + \sigma \bar{s}_1)\, s_1\, \varphi(s_1)\varphi(\bar{s}_1)\, \mathrm{d}s_1\, \mathrm{d}\bar{s}_1 \\ \vdots \\ \frac{1}{\sigma}\int_{R^n} f(x_i + \sigma s_i, \bar{x}_i + \sigma \bar{s}_i)\, s_i\, \varphi(s_i)\varphi(\bar{s}_i)\, \mathrm{d}s_i\, \mathrm{d}\bar{s}_i \\ \vdots \\ \frac{1}{\sigma}\int_{R^n} f(x_n + \sigma s_n, \bar{x}_n + \sigma \bar{s}_n)\, s_n\, \varphi(s_n)\varphi(\bar{s}_n)\, \mathrm{d}s_n\, d\bar{s}_n \end{bmatrix}. \tag{7}$$

Let us consider the generic entry of vector (7)

$$(G_\sigma(x))_i = \frac{1}{\sigma}\int_{R^n} f(x_i + \sigma s_i, \bar{x}_i + \sigma \bar{s}_i) s_i \varphi(s_i)\varphi(\bar{s}_i) \mathrm{d}s_i\, \mathrm{d}\bar{s}_i. \tag{8}$$

By the Fubini theorem, we can compute it as follows

$$(G_\sigma(x))_i = \int_{R^{n-1}} \varphi(\bar{s}_i)\left(\frac{1}{\sigma}\int_{-\infty}^{+\infty} f(x_i + \sigma s_i, \bar{x}_i + \sigma \bar{s}_i) s_i \varphi(s_i) \mathrm{d}s_i\right) \mathrm{d}\bar{s}_i. \tag{9}$$

The expression in parentheses is the estimate of the directional derivative of $f(x)$ along the i-th coordinate x_i and computed at the point $(x_i, \bar{x}_i + \sigma\bar{s}_i)$, i.e.,

$$g_\sigma(x_i, \bar{x}_i + \sigma\bar{s}_i) := \frac{1}{\sigma}\int_{-\infty}^{+\infty} f(x_i + \sigma s_i, \bar{x}_i + \sigma\bar{s}_i) s_i \varphi(s_i) \mathrm{d}s_i. \tag{10}$$

Hence, expression (8) becomes

$$(G_\sigma(x))_i = \frac{1}{\sigma}\int_{R^{n-1}} g_\sigma(x_i, \bar{x}_i + \sigma\bar{s}_i)\, \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i. \tag{11}$$

Therefore, the generic entry of the gradient estimate $G_\sigma(x)$ in formula (7) is the average of function (10) weighted by a $(n-1)$-dimensional Gaussian kernel $\varphi(\bar{s}_i) = \mathcal{N}(0, I_{n-1})$ over the subspace R^{n-1} of R^n. As a consequence, the computation of any entry of vector $G_\sigma(x)$ implies an integration over R^n. In papers [2,3], this problem is overcome by considering that (2) is indeed an ensemble average of function $f(x+\sigma s)s$ over all the directions $s \in R^n$ weighted by the Gaussian distribution $\varphi(s) \sim \mathcal{N}(0, I_n)$. Therefore, we can write

$$G_\sigma(x) = \frac{1}{\sigma} E_\varphi[f(x + \sigma s)s]. \tag{12}$$

Now the ensemble average can be well approximated by sampling a set of M independent directions $\{s_i\}$ in R^n according to $\mathcal{N}(0, I_n)$, and considering the *sample average approximation* of $E_\varphi[f(x + \sigma s)s]$

$$G_\sigma(x) \simeq \frac{1}{M}\sum_{i=1}^{M} \frac{(f(x + \sigma s_i) - f(x))s_i}{\sigma}, \tag{13}$$

or its simmetric version

$$G_\sigma(x) \simeq \frac{1}{M}\sum_{i=1}^{M} \frac{(f(x + \sigma s_i) - f(x - \sigma s_i))s_i}{2\sigma}. \tag{14}$$

The same argument holds if a uniform distribution over the unit ball is considered for the ensemble average [9]. Now, only $M+1$ function computations in case of (13) or $2M$ in case of (14) are needed and the convergence properties of the sample estimate to the ensemble average are well established: the sample average is an unbiased estimate and its accuracy increases with increasing M. In [3], suitable expressions of the estimation error variance are found in terms of the number of samples M and the values of some smoothness parameters of function f. Therefore, very useful formulas are given that define the required sample size to obtain a chosen accuracy, with a fixed level of confidence $1-\alpha$. This is a typical statistical characterization of the error, that is robust over the whole ensemble of possible trials, but of course leaves a risk α to have a large error on a single experiment.

In this paper, by exploiting formula (10), the following gradient estimate is proposed

$$\overline{G}_\sigma(x) := \left[g_\sigma(x_1, \bar{x}_1), \ldots, \; g_\sigma(x_i, \bar{x}_i), \ldots, g_\sigma(x_n, \bar{x}_n)\right]^{\mathrm{T}}, \tag{15}$$

where

$$g_\sigma(x_i, \bar{x}_i) = \frac{1}{\sigma} \int_{-\infty}^{+\infty} f(x_i + \sigma s_i, \bar{x}_i)\, s_i\, \varphi(s_i)\, \mathrm{d}s_i \tag{16}$$

is obtained from (10) with $\bar{s}_i = 0$, $i = 1, \ldots, n$. This is a different result from estimate (7) and appears to be more practical since only line integrals are involved in the formula.

The following theorem shows that estimate $\overline{G}_\sigma(x)$ is close to $G_\sigma(x)$ and converges to it as σ tends to zero.

Theorem 2.1 *Let $\nabla f(x)$ be Lipschitz continuous with constant L for all $x \in R^n$. Then we have that*

$$||G_\sigma(x) - \overline{G}_\sigma(x)|| \le L\,\sigma\, \sqrt{n(15 + 7(n-1))}. \tag{17}$$

Proof See Appendix for the proof.

Next theorem shows that $\overline{G}_\sigma(x)$ is indeed a good approximation of the true gradient $\nabla f(x)$ and converges to it as σ tends to zero. □

Theorem 2.2 *Let $f(x)$ be continuously differentiable for all $x \in R^n$. The following holds:*

$$\lim_{\sigma \to 0} \overline{G}_\sigma(x) = \nabla f(x). \tag{18}$$

Proof We prove (18) component-wise. By integration by parts, we have

$$\begin{aligned} g_\sigma(x_i, \bar{x}_i) &= \frac{1}{\sigma} \int_{-\infty}^{+\infty} f(x_i + \sigma s_i, \bar{x}_i)\, s_i\, \varphi(s_i)\, \mathrm{d}s_i \\ &= \frac{1}{\sigma} \int_{-\infty}^{+\infty} \frac{\partial f(z_i, \bar{x}_i)}{\partial z_i} \frac{n z_i}{\mathrm{d}s_i} \varphi(s_i)\, \mathrm{d}s_i \\ &= \int_{-\infty}^{+\infty} \frac{\partial f(z_i, \bar{x}_i)}{\partial z_i} \varphi(s_i)\, \mathrm{d}s_i, \end{aligned} \tag{19}$$

where $z_i = x_i + \sigma s_i$. By changing of variable, $s_i = \frac{z_i - x_i}{\sigma}$ we obtain that

$$g_\sigma(x_i, \bar{x}_i) = \int_{-\infty}^{+\infty} \frac{\partial f(z_i, \bar{x}_i)}{\partial z_i} \frac{1}{\sigma} \varphi\left(\frac{z_i - x_i}{\sigma}\right) \mathrm{d}z_i \tag{20}$$

and therefore, taking into account that a series of Gaussians $\frac{1}{\sigma_n}\varphi(\frac{z_i - x_i}{\sigma_n})$ with $\sigma_n \to 0$ defines a δ-dirac distribution centered in x_i [10], we have that

$$\lim_{\sigma \to 0} g_\sigma(x_i, \bar{x}_i) = \frac{\partial f(x)}{\partial x_i}. \tag{21}$$

□

Any entry of (15) is a weak definition of the derivative of $f(x)$ along x_i [10]. Note that (19) is well defined even though $f(x)$ is not differentiable at $(x_i, \bar{x}_i)$.[1]

3 A New Estimate of the Gradient

We consider the functional $g_\sigma(x_i, \bar{x}_i)$ which is the i_{th} component of the gradient estimate (15) and, for the sake of simplicity, we write in a single formula the result of (19) and (20).

$$\begin{aligned} g_\sigma(x_i, \bar{x}_i) =& \frac{1}{\sigma}\int_{-\infty}^{+\infty} f(x_i + \sigma s_i, \bar{x}_i)\, s_i\, \varphi(s_i)\, \mathrm{d}s_i \\ =& \int_{-\infty}^{+\infty} \frac{\partial f(z_i, \bar{x}_i)}{\partial z_i} \frac{1}{\sigma}\varphi\left(\frac{z_i - x_i}{\sigma}\right) \mathrm{d}z_i. \end{aligned} \tag{22}$$

Note that $\frac{1}{\sigma}\varphi(\frac{z_i - x_i}{\sigma})$ is $\mathcal{N}(x_i, \sigma^2)$. Our goal consists in finding a numerical approximation of the first integral in (22). To do that, we compute the integral in a finite range, namely between -S and S

$$\begin{aligned} \tilde{g}_\sigma(x_i, \bar{x}_i) :=& \frac{1}{\sigma}\int_{-S}^{+S} f(x_i + \sigma s_i, \bar{x}_i)\, s_i\, \varphi(s_i)\, \mathrm{d}s_i \\ =& -\frac{1}{\sigma}\int_{-S}^{+S} f(x_i + \sigma s_i, \bar{x}_i)\, \varphi'(s_i)\, \mathrm{d}s_i. \end{aligned} \tag{23}$$

For S sufficiently big the error between (22) and (23) is negligible due to the fast decreasing of the Gaussian to infinity. The definite integral in (23) can be approximated by a quadrature formula, e.g., Trapezoidal Rule [1]. Dividing the interval $[-S, S]$ in $2m$ sub-intervals, each of size $h = \frac{S}{m}$ we obtain:

$$\begin{aligned} \tilde{g}_\sigma(x_i, \bar{x}_i) = & -\frac{h}{2\sigma}\Bigg[\Bigg(f(x_i - \sigma S, \bar{x}_i)\, \varphi'(-S) + f(x_i + \sigma S, \bar{x}_i)\, \varphi'(S) \\ & +2\sum_{j=1}^{2m-1} f(x_i + \sigma(-S + jh), \bar{x}_i)\, \varphi'(-S + jh)\Bigg)\Bigg] \\ & + \frac{h^2 S}{6\sigma}\frac{\mathrm{d}}{\mathrm{d}\tau^2} f(x_i + \sigma\tau, \bar{x}_i)\, \varphi'(\tau) \qquad \tau \in [-S, S]. \end{aligned} \tag{24}$$

[1] Any L_1 function satisfying (19), in place of $\frac{\partial f(z_i, \bar{x}_i)}{\partial z_i}$, is a *weak derivative* of $f(x)$ along x_i.

It is well known that, under very general conditions, the trapezoidal quadrature formula (24) has an error that is $\mathcal{O}(1/m^2)$ [5]. Indeed, once σ and S are chosen, we can easily check this property in our case. Let

$$\begin{aligned}\epsilon_\sigma(\tau, m) &= \frac{h^2 S}{6\sigma}\frac{\mathrm{d}}{\mathrm{d}\tau^2} f(x_i + \sigma\tau, \bar{x}_i)\,\varphi'(\tau) \qquad \tau \in [-S, S]. \\ &= \frac{S^3}{6\sigma\, m^2}\frac{\mathrm{d}}{\mathrm{d}\tau^2} f(x_i + \sigma\tau, \bar{x}_i)\,\varphi'(\tau) \qquad \tau \in [-S, S].\end{aligned} \tag{25}$$

Note that the derivatives of a guassian kernel $|\varphi^{(k)}(\tau)|$, up to the third order, are all less than 1 in absolute value for any τ, and decrease rapidly as τ increases. Therefore, for f sufficiently smooth in $(x_i \pm \sigma S)$, let

$$K(x_i) = \max\left(|f(x_i + \sigma\tau, \bar{x}_i)|, \left|\frac{\mathrm{d}}{\mathrm{d}\tau} f(x_i + \sigma\tau, \bar{x}_i)\right|, \left|\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} f(x_i + \sigma\tau, \bar{x}_i)\right|\right).$$

We can write:

$$|\epsilon_\sigma(\tau, m)| \le \frac{h^2 S}{6\sigma} K(x_i) = \frac{S^3}{6\sigma\, m^2} K(x_i), \quad \tau \in [-S, +S].$$

Let us rewrite (24) as follows

$$\tilde{g}_\sigma(x_i, \bar{x}_i) = \bar{g}_\sigma(x_i, \bar{x}_i) + \epsilon_\sigma(\tau, m).$$

The larger the number of function evaluation m, the smaller the error term $\epsilon_\sigma(\tau, m)$. On the other hand, $\bar{g}_\sigma(x_i)$ can be interpreted as a combination of finite differences with some coefficients. Keeping in mind that $\varphi'(t) = -\varphi'(-t)$ and that $\varphi'(0) = 0$, after some simple algebra we can write:

$$\begin{aligned}\bar{g}_\sigma(x_i, \bar{x}_i) = -\frac{h}{2\sigma}\Bigg[&|\varphi'(m\,h)|\Big(f(x_i - \sigma\, m\, h, \bar{x}_i) - f(x_i + \sigma\, m\, h, \bar{x}_i)\Big) \\ &+ 2\sum_{j=1}^{m-1} |\varphi'(jh)|\Big(f(x_i - \sigma\, jh, \bar{x}_i) - f(x_i + \sigma\, jh, \bar{x}_i)\Big)\Bigg]\end{aligned}$$

from which

$$\begin{aligned}\bar{g}_\sigma(x_i, \bar{x}_i) = \frac{h}{2\sigma}\Bigg[&|\varphi'(m\,h)|\, 2\sigma\, m\, h\, \frac{f(x_i + \sigma\, m\, h, \bar{x}_i) - f(x_i - \sigma\, m\, h, \bar{x}_i)}{2\sigma\, m\, h} \\ &+ 2\sum_{j=1}^{m-1} |\varphi'(jh)|\, 2\sigma\; j\, h\, \frac{f(x_i + \sigma\, jh, \bar{x}_i) - f(x_i - \sigma\, jh, \bar{x}_i)}{2\sigma\; j\, h}\Bigg].\end{aligned} \tag{26}$$

It is clear that $\bar{g}_\sigma(x_i, \bar{x}_i)$ is a linear combination of finite difference approximations, with different step sizes; for $\sigma h \to 0$, each one converges to the true value of the partial derivative $\partial f(x_i, \bar{x}_i)/\partial x_i$. Therefore, the estimate $\bar{g}_\sigma(x_i, \bar{x}_i)$ converges to the true value only if the sum of its coefficients equals one. For this reason, it is advisable to *normalize* the coefficients of the linear combination in (26) to eliminate the estimate bias for σ finite. To this aim, let C be the sum of all the coefficients:

$$\left.\begin{aligned} &C = \textstyle\sum_{j=1}^m a'_j, \\ &a'_j = 2\,j\,h^2\,|\varphi'(jh)|, \quad j = 1, \ldots, m-1, \\ &a'_m = m\,h^2\,|\varphi'(mh)|, \end{aligned}\right\}. \tag{27}$$

We can then write the normalized version of (26) as:

$$\hat{g}_\sigma(x_i, \bar{x}_i) = \sum_{j=1}^m a_j \, \frac{f(x_i + \sigma\, j\, h, \bar{x}_i) - f(x_i - \sigma\, j\, h, \bar{x}_i)}{2\sigma\, j\, h} \tag{28}$$

where

$$a_j = \frac{a'_j}{C}, \qquad \sum_{j=1}^m a_j = 1. \tag{29}$$

For σ small enough the normalization of the coefficients may not be necessary, the distorsion of the estimate being negligible. Let us now evaluate the error bound corresponding to estimate (28), from here on referred to as NMXFD (Normalized Mixed Finite Difference).

Theorem 3.1 *Let $f(x)$ be twice continuously differentiable and its Hessian be H-Lipschitz for all $x \in R^n$. Consider the gradient approximation obtained by* (28)

$$\widehat{G}_\sigma(x) = \left[\, \hat{g}_\sigma(x_1), \ldots \hat{g}_\sigma(x_n) \,\right]^{\mathrm{T}}. \tag{30}$$

We have that

$$\|\widehat{G}_\sigma(x) - \nabla f(x)\| \le \sqrt{n}\, \frac{H\sigma^2 S^2}{6}. \tag{31}$$

Proof Any single finite difference term in (28) has an error with respect to the true value $\partial f(x_i, \bar{x}_i)/\partial x_i$ whose bound depends on the step size and on the regularity properties of function f. From [4], we have that

$$\left| \frac{f(x_i + \sigma\, j\, h, \bar{x}_i) - f(x_i - \sigma\, j\, h, \bar{x}_i)}{2\sigma\, j\, h} - \frac{\partial f(x_i, \bar{x}_i)}{\partial x_i} \right| \le \frac{H\sigma^2 (jh)^2}{6} \tag{32}$$

for $j = 1, \ldots, m$. Therefore, since $\sum_{j=1}^m a_j = 1$, and $a_j > 0$, $j = 1, \ldots, m$, we can write

$$\left| \hat{g}_\sigma(x_i) - \frac{\partial f(x_i, \bar{x}_i)}{\partial x_i} \right| = \left| \hat{g}_\sigma(x_i) - \sum_{j=1}^m a_j \, \frac{\partial f(x_i, \bar{x}_i)}{\partial x_i} \right|$$

$$\le \sum_{j=1}^{m} a_j \left| \frac{f(x_i + \sigma\, j\, h, \bar{x}_i) - f(x_i - \sigma\, j\, h, \bar{x}_i)}{2\sigma\, j\, h} - \frac{\partial f(x_i, \bar{x}_i)}{\partial x_i} \right|$$

$$\le \frac{H\sigma^2 h^2}{6} \left(\sum_{j=1}^{m} a_j\, j^2 \right) \le \frac{H\sigma^2 h^2 m^2}{6} = \frac{H\sigma^2 S^2}{6},$$

which applied to all entries of $\widehat{G}_\sigma(x) - \nabla f(x)$, proves the theorem. □

Here we used the equality $m\,h = S$ that implies that the error bound does not depend on the number of function evaluations.

4 Estimation Error with Noisy Data

Let us now evaluate how the performance of the gradient estimate $NMXFD$ (30) here referred to as $\hat{G}_\sigma^{\mathrm{MXF}}(x)$ compares with that of the Central Finite Differences (CFD), taking also into account the presence of an additive noise affecting the sampled function values $f(x)$. Let $\{e_i\}$ be the canonical base of R^n, then we can write:

$$\hat{G}_\sigma^{\mathrm{MXF}}(x) = \sum_{i=1}^{n} \hat{g}_\sigma(x_i)\, e_i \tag{33}$$

with the same notation we can easily write the gradient estimate according to the CFD scheme here denoted as $\hat{G}_\sigma^{\mathrm{CFD}}(x)$:

$$\hat{G}_\sigma^{\mathrm{CFD}}(x) = \sum_{i=1}^{n} \frac{f(x_i + \sigma\, h, \bar{x}_i) - f(x_i - \sigma\, h, \bar{x}_i)}{2\sigma\, h}\, e_i = \sum_{i=1}^{n} \delta f_\sigma(x_i)\, e_i. \tag{34}$$

Let $\{\epsilon_i\}$ denote a discrete random field modeling the additive noise on the sampled function values with the following properties: $\epsilon_i \sim N(0, \lambda^2)$ and $E[\epsilon_i\, \epsilon_j] = 0$ for $i \neq j$. We now compute the estimation errors for the two schemes and compare them in terms of accuracy (mean value) and precision (variance). The accuracy evaluates the estimate bias, i.e., the systematic source of the error, like the limited the number N of function evaluations used to build the estimate. The precision is the dispersion of the estimation error around its mean value and evaluates the variability of the statistic source of the error.

The CFD scheme

According to (34), a number $N = 2n$ of function evaluations is considered to obtain

$$\hat{G}_\sigma^{\mathrm{CFD}}(x) = \sum_{i=1}^{n} \delta f_\sigma(x_i)\, e_i + \sum_{i=1}^{n} \frac{\epsilon_i^+ - \epsilon_i^-}{2\,\sigma\, h}\, e_i$$

with $\epsilon_i^{\pm}$ denoting the noise on the function values $f_\sigma(x_i \pm \sigma\, h, \bar{x}_i)$. Let

$$e_{\text{CFD}}(x) = \hat{G}_\sigma^{\text{CFD}}(x) - \nabla f(x)$$

be the estimation error. We can see that

$$E[e_{\text{CFD}}(x)] = \sum_{i=1}^{n} \delta f_\sigma(x_i)\, e_i - \nabla f(x)$$

and

$$var[e_{\text{CFD}}(x)] = n\,\frac{2\lambda^2}{4\,\sigma^2\,h^2} = \frac{n\,\lambda^2}{2\sigma^2\,h^2} \tag{35}$$

where $var[z]$, $z \in R^n$ with $E[z] = 0$, indicates the trace of the covariance matrix $E[z\,z^{\mathrm{T}}]$. Now, for functions f as in theorem (3.1), let us consider the property (32), with $j = 1$, for all the components of $E[e_{\text{CFD}}(x)]$. We obtain that

$$\|E[e_{\text{CFD}}(x)]\| \le \sqrt{n}\,\frac{H\,\sigma^2\,h^2}{6}.$$

Therefore, as the increment $\sigma h \to 0$, the error goes to zero as well *on average*, but its variance increases without bound as $\mathcal{O}\left(1/(\sigma h)^2\right)$.

The NMXFD scheme
In this case, according to (33), a number $N = 2m\,n$ of function evaluations is considered to obtain

$$\hat{G}_\sigma^{\text{MXF}}(x) = \sum_{i=1}^{n} \hat{g}_\sigma(x_i)\, e_i + \sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_j\, \frac{\epsilon_{i,j}^{+} - \epsilon_{i,j}^{-}}{2\sigma\, j\, h} \right) e_i$$

with $\epsilon_{i,j}^{\pm}$ denoting the error terms on the function values $f(x_i \pm \sigma\, jh, \bar{x}_i)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. For the estimation error

$$e_{\text{MXF}}(x) = \hat{G}_\sigma^{\text{MXF}}(x) - \nabla f(x),$$

we readily obtain that

$$E\,[e_{\text{MXF}}(x)] = \sum_{i=1}^{n} \hat{g}_\sigma(x_i)\, e_i - \nabla f(x)$$

$$var\,[e_{\text{MXF}}(x)] = \frac{n\,\lambda^2}{2\sigma^2 h^2} \left(\sum_{j=1}^{m} \frac{a_j^2}{j^2} \right). \tag{36}$$

Under the assumptions of theorem (3.1), and taking into account (31), we obtain

$$\| E\left[e_{\mathrm{MXF}}(x)\right] \| \leq \sqrt{n}\, \frac{H\sigma^2 m^2 h^2}{6}. \tag{37}$$

As for the error variance, two interesting results can be proved.

Proposition 4.1 *For any $m > 1$, the variance of the estimation error of the $NMXFD$ scheme is strictly lower than the variance of the estimation error of the CFD scheme, i.e.,*

$$var\left[e_{MXF}(x)\right] < var\left[e_{CFD}(x)\right] \tag{38}$$

in any $x \in R^n$ and for any σ, h.

Proof The sum of squares $\sum_{j=1}^{m} a_j^2$ is strictly less then 1 since the coefficients a_j, $j = 1, \ldots, m$, are all positive and their sum is 1. Therefore, from (36) we obtain that

$$var\left[e_{\mathrm{MXF}}(x)\right] = \frac{n\lambda^2}{2\sigma^2 h^2} \sum_{j=1}^{m} \frac{a_j^2}{j^2} < \frac{n\lambda^2}{2\sigma^2 h^2} = var\left[e_{\mathrm{CFD}}(x)\right]. \tag{39}$$

□

Now we further show that $var\left[e_{\mathrm{MXF}}(x)\right]$ goes to zero as N increases.

Proposition 4.2 *For any $x \in R^n$, the variance of the estimation error of the $NMXFD$ scheme has the following asymptotic behavior*

$$var\left[e_{MXF}(x)\right] \sim \mathcal{O}\left(\frac{1}{N}\right). \tag{40}$$

Proof By taking into account relations (27), we have that

$$\begin{aligned} C &= m\, h^2 \left(|\varphi'(mh)| + 2 \sum_{j=1}^{m-1} \frac{j}{m} |\varphi'(jh)| \right) \\ &\leq 2m\, h\, \frac{h}{2} \left(|\varphi'(mh)| + 2 \sum_{j=1}^{m-1} |\varphi'(jh)| \right). \end{aligned} \tag{41}$$

Let us denote with $I_{\varphi'}^{(1)}(m)$ the following quantity

$$I_{\varphi'}^{(1)}(m) = \frac{h}{2} \left(|\varphi'(mh)| + 2 \sum_{j=1}^{m-1} |\varphi'(jh)| \right)$$

that is the trapezoidal quadrature formula for the integral

$$\int_0^S |\varphi'(t)|\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\left(1 - e^{-\frac{S^2}{2}}\right).$$

Due to the $\mathcal{O}(1/N^2)$ property of the error of the trapezoidal rule, we have that

$$\left| I_{\varphi'}^{(1)}(m) - \frac{1}{\sqrt{2\pi}}\left(1 - e^{-\frac{S^2}{2}}\right)\right| = \mathcal{O}(1/N^2).$$

Therefore, from (41), we easily obtain that

$$\begin{aligned}\left| C - \frac{2m\,h}{\sqrt{2\pi}}\left(1 - e^{-\frac{S^2}{2}}\right)\right| &\leq 2m\,h \left| I_{\varphi'}^{(1)}(m) - \frac{1}{\sqrt{2\pi}}\left(1 - e^{-\frac{S^2}{2}}\right)\right| \\ &= \mathcal{O}(1/N^2) \end{aligned} \tag{42}$$

so that C is a bounded quantity as $N = 2m\,n$ increases (by increasing m), taking into account that $mh = S$. Now, according to the relations (29) we can write

$$\begin{aligned}\sum_{j=1}^{m} \frac{a_j^2}{j^2} &= \frac{1}{C^2}\left(\frac{m^2\,h^4 |\varphi'(mh)|^2}{m^2} + \sum_{j=1}^{m-1} \frac{4\,j^2\,h^4\,|\varphi'(jh)|^2}{j^2}\right) \\ &= \frac{h^4}{C^2}\left(|\varphi'(mh)|^2 + 2\sum_{j=1}^{m-1} 2|\varphi'(jh)|^2\right) \\ &\leq \frac{2\,h^3}{C^2}\,\frac{h}{2}\left(2|\varphi'(mh)|^2 + 2\sum_{j=1}^{m-1} 2|\varphi'(jh)|^2\right). \end{aligned}$$

Define now $I_{\varphi'}^{(2)}(m)$ as follows

$$I_{\varphi'}^{(2)}(m) = \frac{h}{2}\left(2|\varphi'(mh)|^2 + 2\sum_{j=1}^{m-1} 2|\varphi'(jh)|^2\right).$$

It is the trapezoidal quadrature rule for the integral

$$2\int_0^S |\varphi'(t)|^2\,\mathrm{d}t = \sqrt{\pi}\ \mathrm{erf}(S) - Se^{-S^2} = \Phi(S),$$

where $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, \mathrm{d}t$ is the Gauss error function. Hence, for the usual property of the error, we can write

$$\left| I^{(2)}_{\varphi'}(m) - \Phi(S) \right| = \mathcal{O}(1/N^2).$$

Therefore, we obtain that

$$\begin{aligned} var\,[e_{\mathrm{MXF}}(x)] =& \frac{n\,\lambda^2}{2\sigma^2\,h^2} \left(\sum_{j=1}^{m} \frac{a_j^2}{j^2} \right) \le \frac{n\,\lambda^2}{2\sigma^2\,h^2} \frac{2\,h^3}{C^2} I^{(2)}_{\varphi'}(m) \\ \le& \frac{n\,\lambda^2}{\sigma^2} \frac{h}{C^2} \left(\left| I^{(2)}_{\varphi'}(m) - \Phi(S) \right| + \left| \Phi(S) \right| \right). \end{aligned}$$

Now recalling that $m\,h = S$, and that $N = 2m\,n$, we can write

$$\begin{aligned} var\,[e_{\mathrm{MXF}}(x)] \le& \frac{n\,\lambda^2}{\sigma^2} \frac{S}{m\,C^2} \left(\left| I^{(2)}_{\varphi'}(m) - \Phi(S) \right| + \left| \Phi(S) \right| \right) \\ \le& \frac{2}{N} \frac{n^2\,\lambda^2\,S}{\sigma^2\,C^2} \left(\left| I^{(2)}_{\varphi'}(m) - \Phi(S) \right| + \left| \Phi(S) \right| \right), \end{aligned}$$

which along with (42), proves the proposition. □

5 Numerical Experiments

We tested our method for estimating the gradient by comparing its performance with those of other methods on 69 functions from the Schittkowski test set [17].

For each function, we did the following: we generated a random starting point x^0 and minimized the function using the quasi-Newton method of Broyden, Fletcher, Goldfarb and Shanno (BFGS) [14], finding the optimal point x^* with $\nabla f(x^*) \approx 0$. We then identified the first instance of a point x^k where

$$\frac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|} \le \alpha$$

for each of the following values of α: $10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$. In this way, we generated seven different buckets, one for each α, of 69 different points, one for each function. Bucket i indicates the one associated to $\alpha = 10^{-i}$. Bucket 0 is therefore the one with the points that are farther from the optimal solution and bucket 6 is the one with points closer to the optimal solution.

Then, for each point we computed the gradient approximations obtained with the Normalized MiXed Finite Differences scheme (NMXFD) and with those considered

Table 1 Median log of relative error with $\sigma = 10^{-2}$

Scheme	N	$B0$	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$
FFD	$n+1$	0.08	1.22	2.20	3.43	4.43	5.47	6.52
CFD	$2n$	**−2.26**	**−1.13**	**−0.32**	**0.69**	**1.60**	**2.41**	**3.58**
	$2n+1$	1.84	1.92	2.52	3.57	4.69	5.63	6.92
GSG	$4n+1$	1.70	1.78	2.34	3.41	4.50	5.46	6.82
	$8n+1$	1.54	1.66	2.10	3.31	4.49	5.38	6.67
	$2n$	1.97	1.96	1.96	1.99	2.08	2.44	3.46
cGSG	$4n$	1.86	1.82	1.86	1.90	2.06	2.95	4.01
	$8n$	1.71	1.69	1.74	1.81	2.13	3.14	4.28
	$2n$	−1.66	−0.53	0.28	1.29	2.26	3.06	4.18
NMXFD	$4n$	-1.98	−0.84	−0.03	0.97	1.92	2.71	3.87
	$8n$	*-1.99*	*-0.86*	*-0.05*	*0.96*	*1.90*	*2.68*	*3.85*

benchmarks by the literature, namely: Forward Finite Differences (FFD), Central Finite Differences (CFD), Gaussian Smoothed Gradient (GSG), Central Gaussian Gmoothed Gradient (cGSG) as defined in [4]. Different tables will summarize the results of this comparison.

The tables show, for different values of the number of function evaluations (N) and different buckets (B), the median value of the log of the relative approximation error over all the 69 points in each bucket.

We define relative approximation error as

$$\eta = \frac{\|g(x) - \nabla f(x)\|}{\|\nabla f(x)\|},$$

where $g(x)$ is the generic gradient estimate. The number of function evaluations N is expressed in the following tables as a function of the number of dimensions n. FFD and CFD schemes only allow for a specific value of N ($n+1$ and $2n$, respectively). In GSG and in $cGSG$, N is linked to the number of direction sampled to build the gradient approximation ($N = (M+1)$ in (13) and $N = 2\,M$ in (14)). In the NMFXD scheme, the value of N is linked to the value of m in formula (28). In particular, we have that $N = 2$ mn. In each table, the lowest entry for every bucket is highlighted in bold, and the second lowest is italic.

5.1 Noise-Free Setting

For the noise-free setting, we report three different tables obtained using a different value of σ (shared by all the schemes) to compute the gradient approximation (Tables 1, 2, 3).

It is possible to notice that in a noise-free setting, lower values of σ tend to yield to better results, as one would expect from the theory. The closer the point is to the minimum value of a function, the harder it is to obtain an accurate estimate of its

Table 2 Median log of relative error with $\sigma = 10^{-5}$

Scheme	N	$B0$	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$
FFD	$n+1$	−2.92	−1.78	−0.80	0.43	1.43	2.47	3.51
CFD	$2n$	**−8.17**	**−6.99**	**−6.14**	**−4.87**	**−3.75**	**−3.07**	**−1.73**
	$2n+1$	1.84	1.84	1.85	1.84	2.00	2.64	3.92
GSG	$4n+1$	1.69	1.71	1.71	1.74	1.90	2.48	3.80
	$8n+1$	1.53	1.57	1.56	1.57	1.77	2.41	3.67
	$2n$	1.96	1.96	1.96	1.96	1.96	1.97	1.93
cGSG	$4n$	1.86	1.82	1.85	1.85	1.85	1.85	1.83
	$8n$	1.71	1.68	1.70	1.68	1.71	1.71	1.71
	$2n$	−7.66	−6.47	−5.58	−4.43	−3.24	−2.75	−1.21
NMXFD	$4n$	−7.90	−6.74	−5.85	−4.67	−3.55	−2.79	−1.45
	$8n$	*−7.95*	*−6.76*	*−5.87*	*−4.73*	*−3.57*	*−2.84*	−1.54

Table 3 Median log of relative error with $\sigma = 10^{-8}$

Scheme	N	$B0$	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$
FFD	$n+1$	−5.56	−4.73	−3.74	1.43	−1.50	−0.44	0.67
CFD	$2n$	−6.00	−6.20	−6.23	**−3.75**	−6.20	−6.25	**−6.23**
	$2n+1$	1.84	1.84	1.84	2.00	1.82	1.83	1.91
GSG	$4n+1$	1.69	1.71	1.72	1.90	1.70	1.69	1.79
	$8n+1$	1.53	1.57	1.56	1.77	1.55	1.55	1.66
	$2n$	1.96	1.96	1.96	1.96	1.96	1.96	1.93
cGSG	$4n$	1.86	1.82	1.85	1.85	1.84	1.85	1.82
	$8n$	1.71	1.68	1.70	1.71	1.71	1.71	1.71
	$2n$	−6.48	−6.36	−6.52	−3.24	−6.41	−6.42	−6.09
NMXFD	$4n$	*−6.17*	*−6.29*	*−6.41*	−3.55	*−6.43*	*−6.48*	*6.20*
	$8n$	**−6.42**	**−6.44**	**−6.44**	*−3.57*	**−6.50**	**−6.51**	−6.15

gradient, unless σ is very small. As a matter of fact, for points belonging to lower index buckets—thus far from the minimum of the function, the value $\sigma = 10^{-5}$ yields the better performances, while accurate estimates of the gradient of points closer to the minimum value of a function require using of a lower value of σ. We can also see that the error of the proposed method, NMXFD, is of the same order of magnitude of that of CFD, and almost always better than that of the other methods.

In our experiments, we have also produced gradient estimates using two more methods:

- by removing the normalization of the coefficients in the computation of NMXFD, i.e., implementing the gradient approximation as in (26).
- by computing the estimate as the raw average of central finite differences at different stepsizes, that is (28) with $a_j = \frac{1}{m}$.

Table 4 Median log of relative error with $\sigma = 10^{-2}$, noisy setting

Scheme	N	$B0$	$B1$	$B2$	B3	$B4$	$B5$	$B6$
FFD	$n+1$	0.22	1.45	2.46	3.57	4.86	5.74	6.86
CFD	$2n$	− 1.06	0.10	1.34	2.23	3.50	4.32	5.49
	$4n+1$	1.72	1.79	2.56	3.66	4.82	5.69	6.84
GSG	$8n+1$	1.56	1.66	2.43	3.51	4.67	5.56	6.70
	$12n+1$	1.47	1.56	2.35	3.43	4.59	5.48	6.61
	$4n$	1.85	1.85	1.86	2.39	3.61	4.33	5.65
cGSG	$8n$	1.71	1.71	1.73	2.29	3.55	4.28	5.61
	$12n$	1.62	1.63	1.65	2.24	3.52	4.25	5.58
	$4n$	− 1.22	0.00	1.17	2.23	3.43	4.24	5.52
NMXFD	$8n$	*− 1.31*	*− 0.09*	*1.08*	*2.15*	*3.40*	*4.19*	*5.42*
	$12n$	**− 1.36**	**− 0.15**	**1.05**	**2.11**	**3.39**	**4.15**	**5.38**

Both of these methods performed consistently worse than NMXFD, and they have not been reported in the tables for brevity. Still, the better performances of NMXFD over the raw average of central finite differences seem to confirm that the rationale behind the choice of coefficients used to weight the CFDs in the proposed approach is promising from a computational point of view.

5.2 Noisy Setting

We also show results of the noisy scenario, where the noise term is described in Sect. 4 and has $\lambda = 0.001$. The estimation procedure is slightly different from the one of the noise-free setting. In Table 4, the median log of the relative errors η_i of the 69 different Schittkowski function is reported. Each η_i is computed as the average of 100 relative approximation errors, resulting from 100 independent noise realizations. The rationale behind this choice was to mitigate the dependence of the results from one particular noise realization. Results are shown in Table 4, where the gradient estimates are obtained with $\sigma = 0.01$.

Table 4 shows that NMXFD performs better than the other schemes in presence of noise, although reasonably low relative approximation errors are obtained only for the first three buckets. For the other ones, the error η increases significantly. This is due to the fact that the denominator of η gets smaller as we move to points close to the minimum value of the function, while the variance of the approximation error does not change across different buckets. Just like in the noise-free setting, increasing the number of function evaluations allows to increase the precision of all the schemes, as expected from the theory.

Different values of σ for estimating the gradient (10^{-1} , 10^{-3}, 10^{-4}) have also been used. The associated tables have not been reported for brevity, since they yielded to the same conclusions and since the performances for almost every method and every bucket with those values of σ are significantly worse. This can be inferred from the

Table 5 Variance reduction coefficient on increasing m for NMXFD (1st column) and mCFD (2nd column)

m	$\sum_{j=1}^{m} \frac{a_j^2}{j^2}$	$\frac{1}{m}$
1	1	1
2	0.877023	**0.5**
3	**0.307637**	0.333333
4	**0.128374**	0.25
5	**0.065331**	0.2
6	**0.037682**	0.166667
7	**0.023683**	0.142857
8	**0.015845**	0.125
9	**0.011119**	0.111111
10	**0.008101**	0.1

theory, since the value of σ influences the bias and the variance of the estimate error in opposite directions, as we can see from (36) and (37) in Sect. 4.

The numerical experiments show the good performances of the proposed method when compared with those of the standard methods commonly used in the literature. In particular, the performances of NMXFD are comparable with those of CFD in absence of noise and better with noisy data and are better than those of other schemes in both scenarios.

The results seem to confirm the idea that performing a combination of finite differences in the noisy setting increases the quality of the gradient estimation. In this line, the simplest combination possible is the average of a number m of multiple CFDs ($mCFD$) computed over repeated measures

$$\hat{G}_{\sigma}^{\mathrm{mCFD}}(x) = \frac{1}{m}\sum_{k=1}^{m} \hat{G}_{\sigma,k}^{\mathrm{CFD}}(x) \tag{43}$$

where $\hat{G}_{\sigma,k}^{\mathrm{CFD}}(x)$ is the CFD in (34) computed at the same points, but with a different independent realization k of the noise. This formula, obviously, reduces the error variance of CFD by $1/m$, therefore it becomes interesting to see if

$$var\left[e_{\mathrm{MXF}}(x)\right] = \frac{n\lambda^2}{2\sigma^2 h^2}\sum_{j=1}^{m}\frac{a_j^2}{j^2} < \frac{1}{m}\frac{n\lambda^2}{2\sigma^2 h^2} = var\left[e_{\mathrm{mCFD}}(x)\right]. \tag{44}$$

Because of the complicated structure of the coefficients a_j a formal proof of (44) can be involved. In Table 5, we report a numerical verification of (44) for increasing values of m, with a uniform sampling within the range [-S, S] with $S = mh = 3$ to compute coefficients a_j.

For $m = 1$, the reduction of the variance of the two methods is the same. For all $m > 2$, we can see that the reduction of the error variance of NMXFD is greater than that of mCFD.

Table 6 Median log of relative error with $\sigma = 10^{-2}$, different values of λ

	Scheme	N	B0	B1	B2	B3	B4	B5	B6
$\lambda = 0.001$	mCFD	$4n$	−1.22	−0.05	1.17	2.13	3.35	4.18	5.39
		$8n$	− 1.36	− 0.19	1.02	2.05	3.2	4.07	5.32
		$12n$	**− 1.43**	**− 0.23**	**0.94**	**1.99**	**3.16**	**4.01**	**5.28**
	NMXFD	$4n$	− 1.22	0	1.17	2.23	3.43	4.24	5.52
		$8n$	− 1.31	− 0.09	1.08	2.15	3.4	4.19	5.42
		$12n$	− 1.36	− 0.15	1.05	2.11	3.39	4.15	5.38
$\lambda = 0.01$	mCFD	$4n$	− 0.24	0.93	2.1	3.04	4.33	5.14	6.31
		$8n$	−0.37	0.78	2.02	2.92	4.17	4.99	6.16
		$12n$	− 0.47	0.71	1.93	**2.82**	4.09	4.91	6.08
	NMXFD	$4n$	−0.3	0.89	2.09	3	4.27	5.09	6.25
		$8n$	− 0.39	0.78	2.01	2.92	4.16	4.97	6.16
		$12n$	**− 0.48**	**0.67**	**1.91**	**2.82**	**4.07**	**4.88**	**6.06**
$\lambda = 0.1$	mCFD	$4n$	0.76	1.93	3.05	4.03	5.32	6.03	7.29
		$8n$	0.6	1.78	2.9	3.89	5.16	5.87	7.15
		$12n$	0.52	1.7	2.8	3.8	5.08	5.79	7.06
	NMXFD	$4n$	0.69	1.88	2.97	3.98	5.26	5.97	7.23
		$8n$	0.58	1.77	2.89	3.88	5.15	5.86	7.12
		$12n$	**0.49**	**1.66**	**2.77**	**3.79**	**5.05**	**5.77**	**7.03**

In Table 6, we finally report the comparison of the median log of relative error between $\hat{G}_\sigma^{\mathrm{MXF}}$ and $\hat{G}_\sigma^{\mathrm{mCFD}}$ on increasing noise levels λ, all computed with a value σ of 0.01 and always using the same function evaluation budget. We do not report the performances of other methods for brevity, since they confirm the same conclusions provided by Table 4.

Table 6 shows that the basic combination $\hat{G}_\sigma^{\mathrm{mCFD}}$ is indeed a good gradient approximation due to the effect of the average that reduces the error variance. As the noise level increases, $\hat{G}_\sigma^{\mathrm{MXF}}$ tends to be better than $\hat{G}_\sigma^{\mathrm{mCFD}}$. This supports the idea that a good gradient approximation depends on both the coefficients of the linear combination and the sampling points where the differences are computed. In this respect, the analysis developed in Sect. 3 to define the new gradient estimate, provides a guide to design a more efficient estimate, depending on the following points:

- the parameter S that determines the range of integration in integral (23);
- the integration formula used to approximate integral (23);
- the filter parameter σ;
- the sampling strategy of the function within the integration range $(-S, S)$.

In this early investigation, we heuristically tried several values for the parameters S and σ, without trying different integration formulas or sampling criteria. The choice of σ may be difficult and affects the quality of the approximation. When the noise level is known, there are some strategies to make a proper choice of σ as in [18]. When the noise level is not known, the choice of this parameter becomes harder and represents an open question to be further investigated, along with the other points in the list above, to improve the performances of NMXFD.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

6 Conclusions

In this paper, a novel scheme to estimate the gradient of a function is proposed. It is based on linear functionals defining a filtered version of the objective function. Unlike standard methods where the approximation error is characterized from a statistical point of view and therefore may be quite large on a given experiment, one advantage of the proposed scheme relies on a deterministic characterization of the approximation error in the noise-free setting.

The other advantage lies in its behavior when function evaluations are affected by noise. In fact, the variance of the estimation error of the proposed method is showed to be strictly lower than that of the Central Finite Difference scheme and diminishes as the number of function evaluations increases. The suitable linear combination of finite differences seems to have a filtering role in the case of noisy functions, thus resulting in a more robust estimator.

Numerical experiments on a significant benchmark given by the 69 Schittkowski functions show the good performances of the proposed method when compared with those of the standard methods commonly used in the literature. In particular, the performances of NMXFD are comparable with those of CFD in absence of noise and better with noisy data and seem to be better than those of other schemes in both scenarios. Moreover, we also show the comparison with NMXFD and the average of repeated CFD, thus using the same budget of function evaluations. As the noise level increases, NMXFD tends to perform better than all the other schemes.

This supports the idea that the theory developed to propose this new scheme can be a suitable framework to design gradient estimates with noisy data. The gradient estimate proposed in this paper can be seen as a first design attempt. A future study could be dedicated to the investigation of the best gradient estimates in this framework, along with the analysis of the impact of the obtained gradient approximation when used in optimization algorithms.

Appendix

Proof of Theorem (2.1) we have that

$$||G_\sigma(x) - \overline{G}_\sigma(x)||^2 = \sum_{i=1}^{n} \big((G_\sigma(x))_i - (\overline{G}_\sigma(x))_i\big)^2$$

where $(G_\sigma(x))_i$ is given by (11)

$$(G_\sigma(x))_i = \int_{R^{n-1}} g_\sigma(x_i,\ \bar{x}_i + \sigma\bar{s}_i)\,\varphi(\bar{s}_i)\,\mathrm{d}\bar{s}_i$$

and $(\overline{G}_\sigma(x))_i = g_\sigma(x_i,\ \bar{x}_i)$, by (16). We can write

$$\begin{aligned}\big((G_\sigma(x))_i - (\overline{G}_\sigma(x))_i\big)^2 &= \left(\int_{R^{n-1}} g_\sigma(x_i,\ \bar{x}_i + \sigma\bar{s}_i)\,\varphi(\bar{s}_i)\,\mathrm{d}\bar{s}_i - g_\sigma(x_i,\ \bar{x}_i)\right)^2\\ &= \left(\int_{R^{n-1}} (g_\sigma(x_i,\ \bar{x}_i + \sigma\bar{s}_i) - g_\sigma(x_i,\ \bar{x}_i))\,\varphi(\bar{s}_i)\,\mathrm{d}\bar{s}_i\right)^2\end{aligned} \tag{45}$$

where the last equality holds since $\int_{R^{n-1}} \varphi(\bar{s}_i)\,\mathrm{d}\bar{s}_i = 1$. Now, the integrand in (45) has the following expression

$$\begin{aligned}&g_\sigma(x_i,\ \bar{x}_i + \sigma\bar{s}_i) - g_\sigma(x_i,\ \bar{x}_i)\\ &\quad= \frac{1}{\sigma}\int_{-\infty}^{\infty} (f(x_i + \sigma s_i,\ \bar{x}_i + \sigma\bar{s}_i) - f(x_i + \sigma s_i,\ \bar{x}_i))\ s_i\,\varphi(s_i)\,ds_i,\end{aligned} \tag{46}$$

and for the argument of the integral we can write

$$\begin{aligned}&f(x_i + \sigma s_i,\ \bar{x}_i + \sigma\bar{s}_i) - f(x_i + \sigma s_i,\ \bar{x}_i)\\ &\quad= (f(x_i + \sigma s_i,\ \bar{x}_i + \sigma\bar{s}_i) - f(x_i,\ \bar{x}_i)) - (f(x_i + \sigma s_i,\ \bar{x}_i) - f(x_i,\ \bar{x}_i))\\ &\quad= \nabla f(x')^{\mathrm{T}}\sigma s - (\nabla f(x_i'', \bar{x}_i))_i\,\sigma\, s_i\\ &\quad= (\nabla f(x'))_i\,\sigma\, s_i + \overline{(\nabla f(x'))}_i^{\mathrm{T}}\,\sigma\,\bar{s}_i - (\nabla f(x_i'', \bar{x}_i))_i\,\sigma\, s_i\end{aligned} \tag{47}$$

with $x' \in (x, x + \sigma s)$ and $x_i'' \in (x_i, x_i + \sigma s_i)$.

We further have that

$$\overline{(\nabla f(x'))}_i = \overline{(\nabla f(x'))}_i - \overline{(\nabla f(x))}_i + \overline{(\nabla f(x))}_i \tag{48}$$

Now substituting (47) and (48) into (46), we obtain that

$$g_\sigma(x_i,\ \bar{x}_i + \sigma\bar{s}_i) - g_\sigma(x_i,\ \bar{x}_i)$$

$$= \int_{-\infty}^{\infty} \left[(\nabla f(x'))_i - (\nabla f(x_i'', \bar{x}_i))_i\right] s_i^2\, \varphi(s_i)\, \mathrm{d}s_i$$
$$+ \int_{-\infty}^{\infty} \left(\overline{(\nabla f(x'))_i} - \overline{(\nabla f(x))_i} + \overline{(\nabla f(x))_i}\right)^{\mathrm{T}} \bar{s}_i\, s_i\, \varphi(s_i)\, \mathrm{d}s_i. \tag{49}$$

By the Lipschitz property of the gradient, and recalling that

$$\int_{-\infty}^{\infty} \overline{(\nabla f(x))_i}^{\mathrm{T}}\, \bar{s}_i\, s_i\, \varphi(s_i)\, \mathrm{d}s_i = 0$$

we have:

$$\begin{aligned}
&(g_\sigma(x_i,\, \bar{x}_i + \sigma \bar{s}_i) - g_\sigma(x_i,\, \bar{x}_i))^2 \\
&\quad \le \int_{-\infty}^{\infty} L^2\sigma^2\, \|s\|^2\, s_i^4\, \varphi(s_i)\, \mathrm{d}s_i \\
&\qquad + \int_{-\infty}^{\infty} L^2\sigma^2\, \|s\|^2\, \|\bar{s}_i\|^2\, s_i^2\, \varphi(s_i)\, \mathrm{d}s_i
\end{aligned} \tag{50}$$

We can finally substitute (50) into (45) obtaining:

$$\begin{aligned}
&\left((G_\sigma(x))_i - (\overline{G}_\sigma(x))_i\right)^2 \\
&\quad \le L^2\sigma^2 \int_{R^{n-1}} \int_{-\infty}^{\infty} (s_i^2 + \|\bar{s}_i\|^2)\, s_i^4\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i + \\
&\qquad + L^2\sigma^2 \int_{R^{n-1}} \int_{-\infty}^{\infty} (s_i^2 + \|\bar{s}_i\|^2)\, \bar{s}_i^{\,2}\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i.
\end{aligned} \tag{51}$$

For the first term in (51), we obtain that

$$\begin{aligned}
&\int_{R^{n-1}} \int_{-\infty}^{\infty} (s_i^2 + \|\bar{s}_i\|^2)\, s_i^4\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i \\
&\quad = \int_{R^{n-1}} \int_{-\infty}^{\infty} s_i^6\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i \\
&\qquad + \int_{R^{n-1}} \int_{-\infty}^{\infty} \|\bar{s}_i\|^2\, s_i^4\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i \\
&\quad = 15 + 3(n-1).
\end{aligned} \tag{52}$$

By similar computations, the second term in (51) becomes

$$\begin{aligned}
&\int_{R^{n-1}} \int_{-\infty}^{\infty} (s_i^2 + \|\bar{s}_i\|^2)\, \|\bar{s}_i\|^2\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i \\
&\quad = \int_{R^{n-1}} \int_{-\infty}^{\infty} s_i^2\, \|\bar{s}_i\|^2\, \varphi(s_i)\, \mathrm{d}s_i\;\; \varphi(\bar{s}_i)\, \mathrm{d}\bar{s}_i
\end{aligned}$$

$$+\int_{R^{n-1}}\int_{-\infty}^{\infty}\|\bar{s}_i\|^4\,\varphi(s_i)\,\mathrm{d}s_i\;\;\varphi(\bar{s}_i)\,\mathrm{d}\bar{s}_i$$
$$=(n-1)+3(n-1)=4(n-1). \tag{53}$$

In (52) and (53), we used the property (p. 208 in [7]) that for a zero mean Gaussian z with variance σ^2:

$$E[z^d]=\begin{cases}(d-1)!!\,\sigma^2, & \text{for } d \text{ even},\\ 0, & \text{for } d \text{ odd},\end{cases}$$

where $(d-1)\,!!=(d-1)(d-3)\cdots 3\cdot 1$ and that for any $z\sim\mathcal{N}(0,I_{n-1})$

$$\int_{R^{n-1}}\|z\|^2\varphi(z)\,\mathrm{d}z=\int_{R^{n-1}}\sum_{i=1}^{n-1}z_i^2\varphi(z)\,\mathrm{d}z=n-1.$$

By substituting (52) and (53) in (51), we finally obtain that

$$\left((G_\sigma(x))_i-(\overline{G}_\sigma(x))_i\right)^2\le L^2\sigma^2\,(15+3(n-1)+4(n-1)),$$

which, applied to all the entries, proves the theorem.

References

1. Atkinson, K.E.: An Introduction to Numerical Analysis. Wiley, New York (2008)
2. Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. Foundations of Computational Mathematics pp. 1–42 (2021)
3. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: Linear interpolation gives better gradients than Gaussian smoothing in derivative-free optimization. arXiv preprint arXiv:1905.13043 (2019)
4. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization. Foundations of Computational Mathematics, pp. 1–54 (2021)
5. Boyd, J.P.: Chebyshev and Fourier Spectral Methods. Springer, Berlin (2001)
6. Conn, A.R., Scheinberg, K., Vicente, L.N.: Geometry of interpolation sets in derivative free optimization. Math. Program. **111**(1–2), 141–172 (2008)
7. Cramér, H.: Mathematical Methods of Statistics, vol. 43. Princeton University Press, Princeton (1999)
8. Fazel, M., Ge, R., Kakade, S., Mesbahi, M.: Global convergence of policy gradient methods for the linear quadratic regulator. In: International Conference on Machine Learning, pp. 1467–1476. PMLR (2018)
9. Flaxman, A.D., Kalai, A.T., McMahan, H.B.: Online convex optimization in the bandit setting: gradient descent without a gradient. arXiv:0408.007 (2004)
10. Gel'fand, I.M., Shilov, G.E.: Generalized Functions, Volume 2: Spaces of Fundamental and Generalized Functions, vol. 261. American Mathematical Soc. (2016)
11. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. **45**(3), 385–482 (2003)
12. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer. **28**, 287–404 (2019)
13. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)

14. Nocedal, J., Wright, S.J.: Sequential quadratic programming. Numer. Optim. pp. 529–562 (2006)
15. Polyak, B.T.: Introduction to Optimization, vol. 1. Inc., Publications Division, New York (1987)
16. Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864 (2017)
17. Schittkowski, K.: More Test Examples for Nonlinear Programming Codes, vol. 282. Springer, Berlin (2012)
18. Shi, H.J.M., Xuan, M.Q., Oztoprak, F., Nocedal, J.: On the numerical performance of derivative-free optimization methods based on finite-difference approximations. arXiv preprint arXiv:2102.09762 (2021)
19. Wild, S.M., Regis, R.G., Shoemaker, C.A.: Orbit: optimization by radial basis function interpolation in trust-regions. SIAM J. Sci. Comput. **30**(6), 3197–3219 (2008)
20. Ziemer, W.P.: Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation, vol. 120. Springer, Berlin (2012)