# Optimal and ecient approximations of gradients of functions with non-independent variables

Matieyendou Lamboni

# Optimal and efficient approximations of gradients of functions with non-independent variables

Matieyendou Lamboni[1a,b]

[a]*University of Guyane, Department DFR-ST, 97346 Cayenne, French Guiana, France*
[b]*228-UMR Espace-Dev, University of Guyane, University of Réunion, IRD, University of Montpellier, France.*

**Abstract**

Gradients of smooth functions with non-independent variables are relevant for exploring complex models and for the optimization of functions subjected to constraints. In this paper, we investigate new and simple approximations and computations of such gradients by making use of independent, central and symmetric variables. Such approximations are well-suited for applications in which the computations of the gradients are too expansive or impossible. The derived upper-bounds of the biases of our approximations do not suffer from the curse of dimensionality for any 2-smooth function, and theoretically improve the known results. Also, our estimators of such gradients reach the optimal (mean squared error) rates of convergence (i.e., $\mathcal{O}(N^{-1})$) for the same class of functions. Numerical comparisons based on a test case and a high-dimensional PDE model show the efficiency of our approach.

*Keywords:* Dependent variables, Gradients, High-dimensional models,, Optimal estimators, Tensor metric of non-independent variables
**AMS**: 26A24, 60H25, 62Gxx, 49Qxx.

## 1. Introduction

Non-independent variables arise when at least two variables do not vary independently, and such variables are often characterized by their covariance matrices, distribution functions, copulas, weighted distributions (see e.g., [1, 2, 3, 4, 5, 6, 7]). Recently, dependency models provide explicit functions that link these variables together by means of additional independent variables ([8, 9, 10, 11, 12]). Models with non-independent input variables, including functions subjected to constraints, are widely encountered in different scientific fields, such as data analysis, quantitative risk analysis, and uncertainty quantification (see e.g., [13, 14, 15]).

[1]Corresponding author: matieyendou.lamboni[at]gmail.com or [at]univ-guyane.fr; June 16, 2024

Analyzing such functions requires being able to calculate or to compute their dependent gradients, that is, the gradients that account for the dependencies among the inputs. Recall that gradients are involved in i) inverse problems and optimization (see e.g., [16, 17, 18, 19, 20]), ii) exploring complex mathematical models or simulators (see [21, 22, 23, 24, 25, 26, 27, 28] for independent inputs and [9, 15] for non-independent variables); iii) Poincaré inequalities and equalities ([29, 30, 9, 28]), and recently in iv) derivative-based ANOVA (i.e., exact expansions) of functions ([28]). While the first-order derivatives of functions with non-independent variables have been derived in [9] for screening dependent inputs of high-dimensional models, the theoretical expressions of the gradients of such functions (dependent gradients) have been introduced in [15], enhancing the difference between the gradients and the first-order partial derivatives when the input variables are dependent or correlated.

In high-dimensional settings and for time-demanding models, having an efficient approach for computing the dependent gradients provided in [15] using a few model evaluations is worth investigating. So far, the adjoint methods can provide the exact classical gradients for some classes of PDE/ODE-based models ([31, 32, 33, 34, 35, 36]). Additionally, Richardson's extrapolation and its generalization considered in [37] provide accurate estimates of the classical gradients using a number of model runs that strongly depends on the dimensionality. In contrary, the Monte-Carlo approach allows for computing the classical gradients using a number of model runs that can be very less than the dimensionality (i.e., $d \in \mathbb{N}$) ([38, 39, 17]). The Monte-Carlo approach is a consequence of the Stokes theorem, which claims that the expectation of a function evaluated at a random point about $\mathbf{x} \in \mathbb{R}^d$ is the gradient of a certain function. Such a property leads to randomized approximations of the classical gradients in derivative-free optimization or zero-order stochastic optimization (see [16, 18, 19, 20] and references therein). Such approximations are also relevant for applications in which the computations of the gradients are impossible ([20]).
Most of the randomized approximations of the classical gradients, including the Monte-Carlo approach, rely on randomized kernels and/or random vectors that are uniformly distributed on the unit ball. The qualities of such approximations are often assessed by the upper-bounds of the biases and the rates of convergence. The upper-bounds provided in [40, 19, 20] depend on the dimensionality in general.

In this paper, we propose new surrogates of the gradients of smooth functions with non-independent inputs and the associated estimators that

- are simple and applicable to a wide class of functions by making use of model

evaluations at randomized points, which are only based on independent, central and symmetric variables;

- lead to a dimension-free upper-bound of the bias, and improve the best known upper-bounds of the bias for the classical gradients;
- lead to the optimal and parametric (mean squared error) rates of convergence;
- are going to increase the computational efficiency and accuracy of the gradients estimates by means of a set of constraints.

Surrogates of dependent gradients are derived in Section 3 by combining the properties of i) the generalized Richardson extrapolation approach thanks to a set of constraints, and ii) the Monte-Carlo approach based only on independent random variables that are symmetrically distributed about zero. Such expressions are followed by their order of approximations, biases and a comparison with known results for the classical gradients. We also provide the estimators of such surrogates and their associated mean squared errors, including the rates of convergence for a wide class of functions (see Section 3.3). A number of numerical comparisons is considered so as to assess the efficiency of our approach. While Section 4 presents comparisons of our approach to other methods, simulations based on a high-dimensional PDE (spatio-temporal) model with given auto-collaborations among the initial conditions are considered in Section 5 to compare our approach to the adjoint-based methods. We conclude this work in Section 6.

## 2. Preliminaries

For an integer $d > 0$, let $\mathbf{X} := (X_1, \ldots, X_d)$ be a random vector of continuous and non-independent variables having $F$ as the joint cumulative distribution function (CDF) (i.e., $\mathbf{X} \sim F$). For any $j \in \{1, \ldots, d\}$, we use $F_{x_j}$ or $F_j$ for the marginal CDF of $X_j$ and $F_j^{-1}$ for its inverse. Also, we use $(\sim j) := (1, \ldots, j-1, j+1, \ldots, d)$ and $\mathbf{X}_{\sim j} := (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)$. The equality (in distribution) $X \overset{d}{=} Z$ means that $X$ and $Z$ have the same CDF.

As the sample values of $\mathbf{X}$ are dependent, here we use $\frac{\partial f}{\partial x_k}$ for the formal partial derivative of $f$ w.r.t. $x_k$, that is, the partial derivative obtained by considering other inputs as constant or independent of $x_k$. Thus, $\nabla f := \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d}\right]^T$ stands for the formal or classical gradient of $f$.

Given an open set $\Omega \subseteq \mathbb{R}^d$, consider a weak partial differentiable function $f : \Omega \to \mathbb{R}$ ([41, 42]). Given $\vec{\boldsymbol{\imath}} := (i_1, \ldots, i_d) \in \mathbb{N}^d$, denote $\mathcal{D}^{(\vec{\boldsymbol{\imath}})} f := \left(\prod_{k=1}^d \frac{\partial^{i_k}}{\partial x_k}\right) f$;

$(\mathbf{x})^{\vec{\boldsymbol{\imath}}} = \mathbf{x}^{\vec{\boldsymbol{\imath}}} := \prod_{k=1}^{d} x_k^{i_k}$, $\vec{\boldsymbol{\imath}}! = i_1! \dots i_d!$, and consider the Hölder space of $\alpha$-smooth functions given by $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathcal{H}_\alpha := \left\{ f : \mathbb{R}^d \to \mathbb{R}^n \, : \, \left| f(\mathbf{x}) - \sum_{0 \le i_1 + \dots + i_d \le \alpha - 1} \frac{\mathcal{D}^{(\vec{\boldsymbol{\imath}})} f(\mathbf{y})}{\vec{\boldsymbol{\imath}}!} \, (\mathbf{x} - \mathbf{y})^{\vec{\boldsymbol{\imath}}} \right| \le M_\alpha \, ||\mathbf{x} - \mathbf{y}||_2^\alpha \right\},$$

with $\alpha \ge 1$ and $M_\alpha > 0$. We use $||\cdot||_2$ for the Euclidean norm, $||\cdot||_1$ for the $L_1$-norm, $\mathbb{E}(\cdot)$ for the expectation and $\mathbb{V}(\cdot)$ for the variance.

For the stochastic evaluations of functions, consider $L, q \in \mathbb{N} \setminus \{0\}$, $\beta_\ell \in \mathbb{R}$ with $\ell = 1, \dots, L$, $\boldsymbol{h} := (h_1, \dots, h_d) \in \mathbb{R}_+^d$, and denote with $\mathbf{V} := (V_1, \dots, V_d)$ a $d$-dimensional random vectors of independent variables satisfying: $\forall\, j \in \{1, \dots, d\}$,

$$\mathbb{E}\left[V_j\right] = 0; \qquad \mathbb{E}\left[(V_j)^2\right] = \sigma^2; \qquad \mathbb{E}\left[(V_j)^{2q+1}\right] = 0; \qquad \mathbb{E}\left[(V_j)^{2q}\right] < +\infty\,.$$

Random vectors of independent variables that are symmetrically distributed about zero are instances of $\mathbf{V}$, including the standard Gaussian random vector and symmetric uniform distributions about zero.
Also, denote $\boldsymbol{h}\mathbf{V} := (h_1 V_1; \dots, h_d V_d)$; $\boldsymbol{h}^{-1}\mathbf{V} := (V_1/h_1; \dots, V_d/h_d)$ and $\beta_\ell \boldsymbol{h}\mathbf{V} := (\beta_\ell h_1 V_1; \dots, \beta_\ell h_d V_d)$. The reals $\beta_\ell$'s are used for controlling the order of approximations and the order of derivatives (i.e., $||\vec{\boldsymbol{\imath}}||_1 = 1, 2$) we are interested in. Finally, $h_j$'s are used to define a neighborhood of a sample point of $\mathbf{X}$ (i.e., $\mathbf{x}$). Thus, using $\beta_{max} := \max(|\beta_1|, \dots, |\beta_L|)$ and keeping in mind the variance of $\beta_\ell h_j V_j$, we assume that $\forall\, j \in \{1, \dots, d\}$,

Assumption (A1) : $\beta_{max} h_j \sigma \le 1/2$ or equivalently $0 < \beta_{max} h_j |V_j| \le 1$ for bounded $V_j's$.

## 3. Main results

This section aims at providing new expressions of the gradient of a function with non-independent variables, and the associated order of approximations. We are also going to derive the estimators of such a gradient, including the optimal and parametric rates of convergence. Recall that the input variables are said to be non-independent whenever there exists at least two variables $X_j$, $X_k$ such that the joint CDF $F_{j,k}(x_j, x_k) \neq F_j(x_j) F_k(x_k)$.

*3.1. Stochastic expressions of the gradients of functions with dependent variables*

Using the fact $\mathbf{X} \sim F$ with $F(\mathbf{x}) \neq \prod_{j=1}^{d} F_j(x_j)$, we are able to model $\mathbf{X}$ as follows ([8, 10, 9, 14, 11, 12, 43]):

$$\begin{aligned} \mathbf{X}_{\sim j} &\overset{d}{=} r_j\left(X_j, \mathbf{Z}_{\sim j}\right) \\ &= \left[r_{1,j}\left(X_j, \mathbf{Z}_{\sim j}\right), \ldots, r_{j-1,j}\left(X_j, \mathbf{Z}_{\sim j}\right), r_{j+1,j}\left(X_j, \mathbf{Z}_{\sim j}\right) \ldots, r_{d,j}\left(X_j, \mathbf{Z}_{\sim j}\right)\right]^T , \end{aligned} \tag{1}$$

where $r_j : \mathbb{R}^d \to \mathbb{R}^{d-1}$; $X_j$ and $\mathbf{Z}_{\sim j} := (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots Z_d)$ are independent. Moreover, we have $(X_j, \mathbf{X}_{\sim j}) \overset{d}{=} (X_j, r_j(X_j, \mathbf{Z}))$, and it is worth noting that the function $r_j$ is invertible w.r.t. $\mathbf{Z}_{\sim j}$ for continuous variables, that is,

$$\mathbf{Z}_{\sim j} = r_j^{-1}\left(\mathbf{X}_{\sim j} \,|\, X_j\right) .$$

Note that the formal Jacobian matrix of $g : \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{x} \mapsto \mathbf{x}$ is the identity matrix. As $\mathbf{x}$ is a sample value of $\mathbf{X}$, the dependent Jacobean of $g$ based on the above dependency function is clearly not the identity matrix due to the fact that such a matrix accounts for the dependencies among the elements of $\mathbf{x}$. The dependent partial derivatives of $\mathbf{x}$ w.r.t. $x_j$ is then given by ([9, 15])

$$J^{(j)}(\mathbf{x}) := \frac{\partial \mathbf{x}}{\partial x_j} = \left[\frac{\partial r_{1,j}}{\partial x_j} \ \ldots \ \underbrace{1}_{j^{\text{th}} \text{ position}} \ \ldots \ \frac{\partial r_{d,j}}{\partial x_j}\right]^T \left(x_j,\, r_j^{-1}\left(\mathbf{x}_{\sim j} | x_j\right)\right) ,$$

and the dependent Jacobian matrix becomes (see [15] for more details)

$$J^d(\mathbf{x}) := \left[J^{(1)}(\mathbf{x}), \ldots, J^{(d)}(\mathbf{x})\right] .$$

Moreover, the gradient of $f$ with non-independent variables is given by ([15])

$$grad(f)(\mathbf{x}) := \left[J^d(\mathbf{x})^T J^d(\mathbf{x})\right]^{-1} \nabla f(\mathbf{x}) = G^{-1}(\mathbf{x}) \nabla f(\mathbf{x}) , \tag{2}$$

with $G(\mathbf{x}) := J^d(\mathbf{x})^T J^d(\mathbf{x})$ the tensor metric and $G^{-1}(\mathbf{x})$ its generalized inverse. Based on the above framework, Theorem 1 provides the stochastic expression of $grad(f)(\mathbf{x})$. In what follows, denote $\mathbb{1}_\bullet := [1, \ldots, 1]^T \in \mathbb{R}^d$.

**Theorem 1.** *Assume $f \in \mathcal{H}_\alpha$ with $\alpha \geq 2L$, (A1) holds and $\beta_\ell$'s are distinct. Then, there exists $\alpha_1 \in \{1, \ldots, L\}$ and reals coefficients $C_1, \ldots, C_L$ such that*

$$grad(f)(\mathbf{x}) = G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L} C_\ell \, \mathbb{E}\left[f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\right] + \mathcal{O}\left(||\mathbf{h}||_2^{2\alpha_1}\right) \mathbb{1}_\bullet . \tag{3}$$

*Proof.* See Appendix A for the detailed proof. □

Using the Kronecker symbol $\delta_{1,r}$, the setting $L=1, \beta_1=1, C_1=1$ or the constraints $\sum_{\ell=1}^{L=2} C_\ell \beta_\ell^r = \delta_{1,r};\ r=0,1$ lead to the order of approximation $\mathcal{O}\left(||\mathbf{h}||_2^2\right)$, while the constraints $\sum_{\ell=1}^{L} C_\ell \beta_\ell^r = \delta_{1,r};\ r=1,3,5,\ldots,2L-1$ allow for increasing that order up to $\mathcal{O}\left(||\mathbf{h}||_2^{2L}\right)$. For distinct $\beta$'s, the above constraints lead to the existence of the constants $C_1,\ldots,C_L$. Indeed, some constraints rely on the Vandermonde matrix of the form

$$A_L := \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \beta_1 & \beta_2 & \ldots & \beta_L \\ \beta_1^2 & \beta_2^2 & \ldots & \beta_L^2 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

which is invertible for distinct values of $\beta_\ell$'s (i.e., $\beta_{\ell_1} \neq \beta_{\ell_2}$) because the determinant $\det(A_L) = \prod_{1\leq \ell_1 < \ell_2 \leq L} (\beta_{\ell_1} - \beta_{\ell_2})$.

**Remark 1.** *For an even integer $L$ , the following nodes may be considered: $\{\beta_1,\ldots,\beta_L\} = \left\{\pm 2^k,\ k=0,\ldots,\frac{L-2}{2}\right\}$. When $L$ is odd, one may add 0 to the above set. Of course, there are other possibilities provided that $\sum_{\ell=1}^{L} C_\ell \beta_\ell = 1$.*

Beyond the strong assumption made on functions in Theorem 1, and knowing that increasing $L$ will require more evaluations of $f$ at random points, we are going to derive the upper-bounds of the biases of our appropriations under different structural assumptions on the deterministic functions $f$ and $\mathbf{V}$, such as $f \in \mathcal{H}_\alpha$ with $\alpha > 1$. To that end, denote with $\mathbf{R} := (R_1,\ldots,R_d)$ a $d$-dimensional random vector of independent variables that are centered about zero and standardized (i.e., $\mathbb{E}[R_k^2]=1$, $k=1,\ldots,d$), and $\mathcal{R}_c$ the set of such random vectors. Define

$$K_1 := \inf_{\mathbf{R}\in\mathcal{R}_c} \left|\left|\left|G^{-1}(\mathbf{x})\right| \mathbb{E}\left[\mathbf{R}^2 ||\mathbf{R}||_2\right]\right|\right|_1; \qquad K_2 := \inf_{\mathbf{R}\in\mathcal{R}_c} \left|\left|\left|G^{-1}(\mathbf{x})\right| \mathbb{E}\left[\mathbf{R}^2 ||\mathbf{R}||_2\right]\right|\right|_2;$$

with $|G^{-1}|$ the matrix obtained by putting the entries of $G^{-1}$ in the absolute value. When $1<\alpha\leq 2$, only $L=1$ or $L=2$ can be considered for any function that belongs to $\mathcal{H}_\alpha$. To be able to derive the parametric rates of convergence, Corollary 1 starts providing the upper-bounds of the bias when $L=2$.

**Corollary 1.** *Consider $\beta_1=1, \beta_2=-1$; $C_1=1/2$; $C_2=-1/2$. If $f\in\mathcal{H}_2$ and (A1) holds, then there exists $M_2>0$ such that*

$$\left|\left| grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_1 \leq \sigma M_2 K_1 ||\boldsymbol{h}||_2; \quad (4)$$

$$\left|\left|grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \leq \sigma M_2 K_2 \left|\left|\boldsymbol{h}\right|\right|_2 . \quad (5)$$

*Proof.* Detailed proofs are provided in Appendix B. □

For a particular choice of $\mathbf{V}$, we obtain the results below.

**Corollary 2.** *Consider $\beta_1 = 1, \beta_2 = -1$; $C_1 = 1/2$; $C_2 = -1/2$. If $V_k \sim \mathcal{U}(-\xi, \xi)$ with $\xi > 0$, $k = 1, \ldots, d$; $f \in \mathcal{H}_2$ and (A1) holds, then*

$$\left|\left|grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_1 \leq M_2 \xi \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{1}_\bullet \right|\right|_1 \left|\left|\boldsymbol{h}\right|\right|_1 ; \quad (6)$$

$$\left|\left|grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \leq M_2 \xi \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{1}_\bullet \right|\right|_2 \left|\left|\boldsymbol{h}\right|\right|_1 . \quad (7)$$

*Proof.* Since $|V_k| \leq \xi$, we have $\left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_1 \leq \xi \left|\left|\boldsymbol{h}\right|\right|_1$ and the results hold using the upper-bounds $M_2 \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{E}\left[ \frac{\mathbf{V}^2}{\sigma^2} \left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_1 \right] \right|\right|_1$ and $M_2 \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{E}\left[ \frac{\mathbf{V}^2}{\sigma^2} \left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_1 \right] \right|\right|_2$ obtained in Appendix B. □

It is worth noting that, choosing $h_k = h$ and $\xi = 1/d^2$ leads to the dimension-free upper-bound of the bias, that is,

$$\left|\left|grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_1 \leq \frac{M_2 h}{d} \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{1}_\bullet \right|\right|_1 ,$$

because $\left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{1}_\bullet \right|\right|_2$ is a function of $d$ in general.

For the sequel of generality, Corollary 3 provides the bias of our approximations for highly smooth functions. To that end, define

$$K_{2,L} := \inf_{\mathbf{R} \in \mathcal{R}_c} \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{E}\left[ \mathbf{R}^2 \left|\left|\mathbf{R}\right|\right|_2^L \right] \right|\right|_2 ; \qquad K_3 := \sum_{\ell=1}^{L+1} \left| C_\ell \beta_\ell^{1+L} \right| .$$

**Corollary 3.** *For an odd integer $L > 2$, consider $\sum_{\ell=1}^{L+1} C_\ell \beta_\ell^r = \delta_{1,r}$; $r = 0, 1, \ldots, L$. If $f \in \mathcal{H}_{1+L}$ and (A1) holds, then there exists $M_{1+L} > 0$ such that*

$$\left|\left|grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L+1} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \leq \sigma^L M_{1+L} K_{2,L} K_3 \left|\left|\boldsymbol{h}\right|\right|_2^L .$$

*Moreover, if* $V_k \sim \mathcal{U}(-\xi, \xi)$ *with* $\xi > 0$ *and* $k = 1, \ldots, d$, *then*

$$\left|\left| grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L+1} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\boldsymbol{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \leq \xi^L M_{1+L} \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{1}_\bullet \right|\right|_2 ||\boldsymbol{h}||_1^L \, K_3 \, .$$

*Proof.* The proofs are similar to those of Corollary 1 (see Appendix B). □

In view of the results provided in Corollary 3, finding $\beta$'s and $C$'s that minimize the quantity $K_3 = \sum_{\ell=1}^{L+1} \left| C_\ell \beta_\ell^{1+L} \right|$ might be helpful for improving the above upper-bounds.

### *3.2. Links to other works for independent input variables*

Recall that for independent input variables, the matrix $|G^{-1}(\mathbf{x})|$ comes down to the identity matrix, and $grad(f) = \nabla f$. Thus, Equation (7) becomes

$$\left|\left| \nabla f(\mathbf{x}) - \sum_{\ell=1}^{L=2} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\boldsymbol{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \leq M_2 h \, ,$$

when $\xi = \sqrt{d}/d^2$. Taking $\xi = \sqrt{d}/d$ leads to the upper-bound $M_2 h d$.

Other results about the upper-bounds of the bias of the (formal) gradient approximations have been provided in [19, 20] (and the references therein) under the same assumptions made on $f$ and evaluations of $f$. Such results rely on a random vector $\mathbf{S}$ that is uniformly distributed on the unit ball and a kernel $K$. Under such a framework, the upper-bound derived in [19, 20] is

$$\left|\left| \nabla f(\mathbf{x}) - \frac{d}{h} \mathbb{E}\left[ f(\mathbf{x} + Uh\mathbf{S}) \mathbf{S} K(U) \right] \right|\right|_2 \leq 2\sqrt{2} \alpha d M_\alpha h^{\alpha - 1} \, ,$$

where $U \sim \mathcal{U}(-1, 1)$ is independent of $\mathbf{S}$. Therefore, our results improve the upper-bound obtained in [19, 20] when $\alpha = 2$ for instance.

### *3.3. Computation of the gradients of functions with dependent variables*

Consider a sample of $\mathbf{V}$ given by $\{\mathbf{V}_i := (V_{i,1}, \ldots, V_{i,d})\}_{i=1}^N$. Using Equation (3), the estimator of $grad(f)(\mathbf{x})$ is derived as follows:

$$\widehat{grad(f)}(\mathbf{x}) := G^{-1}(\mathbf{x}) \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{L} C_\ell f\left(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}_i\right) \frac{\mathbf{V}_i \boldsymbol{h}^{-1}}{\sigma^2} \, .$$

To assess the quality of such an estimator, it is common to use the mean squared error (MSE), including the rates of convergence. The MSEs are often used in statistics for determining the optimal value of $\boldsymbol{h}$ as well. Theorem 2 and Corollary 4

provide such quantities of interest. To that end, define

$$K_4 := \inf_{\mathbf{R}\in\mathcal{R}_c} \mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\mathbf{R}\mathbf{h}^{-1}\right|\right|_2^2 \left|\left|\mathbf{R}^2\right|\right|_2\right] .$$

**Theorem 2.** *Consider $\beta_1 = 1, \beta_2 = -1$; $C_1 = 1/2$; $C_2 = -1/2$. If $f \in \mathcal{H}_2$ and (A1) holds, then*

$$\mathbb{E}\left[\left|\left|\widehat{grad(f)}(\mathbf{x}) - grad(f)(\mathbf{x})\right|\right|_2^2\right] \leq \sigma^2 M_2^2 K_2^2 \left|\left|\boldsymbol{h}\right|\right|_2^2 + \frac{M_1^2 K_4 \left|\left|\boldsymbol{h}^2\right|\right|_2}{N} . \tag{8}$$

*Moreover, if $V_k \sim \mathcal{U}(-\xi, \xi)$ with $\xi > 0$ and $k = 1, \ldots, d$ and $\mathbf{R_0} := \mathbf{V}/\sigma$, then*

$$\begin{aligned}\mathbb{E}\left[\left|\left|\widehat{grad(f)}(\mathbf{x}) - grad(f)(\mathbf{x})\right|\right|_2^2\right] \leq\ & M_2^2\xi^2 \left|\left|\left|G^{-1}(\mathbf{x})\right| \mathbb{1}_\bullet\right|\right|_2^2 \left|\left|\boldsymbol{h}\right|\right|_1^2 \\ & + \frac{3\sqrt{d}M_1^2 \left|\left|\boldsymbol{h}^2\right|\right|_2}{N} \mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\mathbf{R_0}\mathbf{h}^{-1}\right|\right|_2^2\right] .\end{aligned} \tag{9}$$

*Proof.* See Appendix C. □

Using a uniform bandwidth, that is, $h_k = h$ with $k = 1, \ldots, d$, the upper-bounds of MSEs provided in Theorem 2 have simple expressions. Indeed, the upper-bounds in Equations (8)-(9) become, respectively,

$$\sigma^2 M_2^2 K_2^2 d h^2 + \frac{M_1^2\sqrt{d}}{N} \inf_{\mathbf{R}\in\mathcal{R}_c} \mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\mathbf{R}\right|\right|_2^2 \left|\left|\mathbf{R}^2\right|\right|_2\right] ;$$

$$M_2^2\xi^2 \left|\left|\left|G^{-1}(\mathbf{x})\right| \mathbb{1}_\bullet\right|\right|_2^2 d^2 h^2 + \frac{3dM_1^2}{N} \mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\mathbf{R}_0\right|\right|_2^2\right] .$$

It comes out that the second-terms of the above upper-bounds do not depend on the bandwidth $h$. This key observation leads to the derivation of the optimal and parametric rates of convergence of the proposed estimator.

**Corollary 4.** *Under the assumptions made in Theorem 2, if $\xi = d^{-3/2}$ and $h_k = h \propto N^{-\gamma/2}$ with $\gamma \in ]1, 2[$, then we have*

$$\mathbb{E}\left[\left|\left|\widehat{grad(f)}(\mathbf{x}) - grad(f)(\mathbf{x})\right|\right|_2^2\right] = \mathcal{O}\left(N^{-1}d^2\right) .$$

*Proof.* The proof is straightforward since $h^2 \propto N^{-\gamma}$ and $Nh \to \infty$ when $N \to \infty$. □

It is worth noting that the upper-bound of the squared bias obtained in Corollary 4 does not depend on the dimensionality thanks to the choice $\xi = d^{-3/2}$. But, the derived rate of convergence depends on $d^2$, meaning that our estimator suffers from

the curse of dimensionality. In higher-dimensions, an attempt to improve our results consists in controlling the upper-bound of the second-order moment of the estimator through $\sum_{\ell=1}^{L} |C_\ell \beta_\ell|$. For instance, requiring $\sum_{\ell=1}^{L} |C_\ell \beta_\ell| = 1/d^2$ with $L = 2$ admits a solution in $\mathbb{C}$ and not in $\mathbb{R}$.

**Remark 2.** *For highly smooth functions (i.e., $f \in \mathcal{H}_{1+L}$ with $L > 3$) and under the assumptions made in Corollary 3, we can check that (see Appendix C)*

$$\mathbb{E}\left[\left|\left|\widehat{grad(f)}(\mathbf{x}) - grad(f)(\mathbf{x})\right|\right|_2^2\right] \leq \xi^{2L} M_{1+L}^2 \left|\left|\left|G^{-1}(\mathbf{x})\right| \mathit{\Pi}_\bullet\right|\right|_2^2 ||\boldsymbol{h}||_1^{2L} K_3^2 + \frac{3\sqrt{d} M_1^2 \left|\left|\boldsymbol{h}^2\right|\right|_2}{N} \left(\sum_{\ell=1}^{L+1} |C_\ell \beta_\ell|\right)^2 \mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x}) \mathbf{R}_0 \mathbf{h}^{-1}\right|\right|_2^2\right].$$

## 4. Computations of the formal gradient of Rosenbrock's function

For comparing our approach to i) the finite differences method (FDM) using the R-package numDeriv ([44]) with $h = 10^{-4}$, ii) the Monte Carlo (MC) approach provided in [17] with $h = 10^{-4}$, let us consider the Rosenbrock function given as follows: $\forall \mathbf{x} \in \mathbb{R}^d$,

$$r(\mathbf{x}) := \sum_{k=1}^{d-1} \left[(1 - x_k)^2 + 100 \left(x_{k+1} - x_k^2\right)^2\right].$$

The gradient of that function at $\mathbf{0}$ is $\nabla r(\mathbf{0}) = [-2, \ldots, -2, 0]^T \in \mathbb{R}^{100}$ (see [17]). To assess the numerical accuracy of each approach, the following measure is considered:

$$Err := \frac{\left|\left|\nabla r(\mathbf{0}) - \widehat{\nabla r}(\mathbf{0})\right|\right|_1}{||\nabla r(\mathbf{0})||_1},$$

where $\widehat{\nabla r}(\mathbf{0})$ is the estimated value of the gradient. Table 1 reports the values of $Err$ for the three approaches. To obtain the results using our approach, we have used $h = 1/\sqrt{N}$ with $N$ the sample size and $\xi = 1/d^2 = 10^{-4}$ with $d = 100$. Also, the Sobol sequence is used for generating the values of $V_j$'s, and the Gram-schmidt algorithm is applied to obtain (perfect) orthogonal vectors for a given $N$.

Based on Table 1, our approach provides efficient results compared to other methods. Since the FDM is not possible when $N < 2d = 200$, it comes out that our approach is much flexible thanks to $L$ and the fact that the gradient can be computed for every value of $N$. Increasing $N$ improves our results, as expected.

| | Number of total model evaluations (i.e., $LN$) | | | | | |
|---|---|---|---|---|---|---|
| | 100 | 150 | 200 | 200 | 1000 | 1000 |
| Methods | | | | | | |
| FDM ([44]) | - | - | - | 0.005 | - | - |
| MC ([17]) | 0.042 | - | - | - | - | - |
| | $L=1$ | $L=1$ | $L=1$ | $L=2$ | $L=1$ | $L=2$ |
| This paper | 0.035 | 0.014 | 0.009 | 0.009 | 0.0020 | 0.00199 |

Table 1: Values of $Err$ for three different approximations of the formal gradients.

## 5. Application to a heat PDE model with stochastic initial conditions

### *5.1. Heat diffusion model and its formal gradient*

Consider a time-dependent model $f(x,t)$ defined by the one-dimensional (1-D) diffusion PDE with stochastic initial conditions, that is,

$$\begin{cases} \frac{\partial f}{\partial t} - D\frac{\partial^2 f}{\partial^2 x} = 0, & x \in ]0,\,1[,\, t \in [0,T] \\ f(x,t=0) = Z(x) & x \in [0,\,1] \\ f(x=0,t) = 0, \quad f(x=1,t) = 1, & t \in [0,\,T] \end{cases},$$

where $D \in \mathbb{R}_+$ represents the diffusion coefficient. It is common to consider $J(Z(x)) := \frac{1}{2}\int_0^T \int_0^{10} \left(f(x,t)\right)^2 \, dxdt$ as the quantity of interest (QoI). The spatial discretisation consists in subdividing the spatial domain $[0,\,1]$ in $d$ equally-sized cells, which lead to $d$ initial conditions or inputs given by $Z(x_j)$ with $j = 1,\ldots,d$. Given zero-mean random variables $(R_j,\, j = 1,\ldots,d)$, assume that $X_j := Z(x_j) = \sin(2\pi x_j) + s_j R_j,\ j = 1,\ldots,d$, where $s_j \in \mathbb{R}_+$ represents the inverse precision about our knowledge on the initial conditions. For the dynamic aspect, a time step of 0.025 is considered starting from 0 up to $T = 5$.

Given a direction $z(x)$ and the Gâteaux derivative $\check{f}(x,t) := \frac{\partial f}{\partial z(x)}$, the tangent linear model is derived as follows:

$$\begin{cases} \frac{\partial \check{f}}{\partial t} - D\frac{\partial^2 \check{f}}{\partial^2 x} = 0, & x \in ]0,\,1[,\, t \in [0,T] \\ \check{f}(x,t=0) = z(x), & x \in [0,\,1] \\ \check{f}(x=0,t) = \check{f}(x=1,t) = 0, & t \in [0,\,T] \end{cases},$$

and we can check that the adjoint model (AM) (i.e., $f^a$) is given by

$$\begin{cases} -\frac{\partial f^a}{\partial t} - D\frac{\partial^2 f^a}{\partial^2 x} = f, & x \in ]0,\,1[,\, t \in [0,T] \\ f^a(x=0,t) = f^a(x=1,t) = 0, & t \in [0,\,T] \\ f^a(x,T) = 0, & x \in [0,\,1] \end{cases}.$$

The formal gradient of $J(Z(x))$ w.r.t. the inputs $Z(x)$ is $\nabla_Z J(Z(x)) = f^a(x, 0)$. Remark that the above gradient relies on $f^a(x, 0)$, and only one evaluation of such a function is needed.

*5.2. Spatial auto-correlations of initial conditions and the tensor metric*

Recall that the above gradient is based on the assumption of independent input variables, suggesting that the initial conditions within different cells are uncorrelated. To account for the spatial auto-correlations between different cells, assume that the $d$ input variables follow the Gaussian process with the following auto-correlation function:

$$\rho(X_{j_1}, X_{j_2}) = \left(\frac{1}{2}\right)^{|j_1 - j_2|} \mathbb{1}_{[0,3]}(|j_1 - j_2|); \qquad \forall\, j_1, j_2 \in \{1, \ldots, d\}\,,$$

where $\mathbb{1}_{[0,3]}(|j_1 - j_2|) = 1$ if $|j_1 - j_2| \in [0, 3]$ and zero otherwise. Such spatial auto-correlations lead to the correlation matrix of the form

$$\mathcal{R} := \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 & 0 & 0 & 0 & \ldots & 0 \\ 0.5 & 1 & 0.5 & 0.25 & 0.125 & 0 & 0 & \ldots & 0 \\ 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0.125 & 0 & \ldots & 0 \\ 0.125 & 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0.125 & 0 & \ldots\ 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots\ \ldots \end{bmatrix}.$$

Using the same standard deviation $s_j = s$ leads to the following covariance matrix $\Sigma = s^2\mathcal{R}$, and $\mathbf{X} = (X_1, \ldots, X_d) \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} := (\sin(2\pi c_1), \ldots, \sin(2\pi c_d))$ and $c_1, \ldots, c_d$ the centers of the cells. The associated dependency model is given below.

Consider the diagonal matrix $D_{\sim j} = diag\,(\Sigma_{1,1}, \ldots, \Sigma_{j-1,j-1}, \Sigma_{j+1,j+1}, \ldots, \Sigma_{d,d})$, and the Gaussian random vector $\mathbf{W} \sim \mathcal{N}_{d-1}\left(\boldsymbol{\mu}_{\sim j}, D_{\sim j}\right)$. Denote with $\Sigma^{(j)}$ the matrix obtained by moving the $j^{\text{th}}$ row and column of $\Sigma$ to the the first row and column; $\mathcal{L}^{(j)}$ the Cholesky factor of $\Sigma^{(j)}$, and $\boldsymbol{\mu}^{(j)} := (\mu_j, \mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_d)$. We can see that $(X_j, \mathbf{X}_{\sim j}) \sim \mathcal{N}_d\left(\boldsymbol{\mu}^{(j)}, \Sigma^{(j)}\right)$, and the dependency model is given by ([10])

$$(X_j, \mathbf{X}_{\sim j}) \;=\; \mathcal{L}^{(j)} \begin{bmatrix} \frac{1}{\sqrt{\Sigma_{j,j}}}\left(X_j - \mathbb{E}[X_j]\right) \\ D_{\sim j}^{-1/2}\left(\mathbf{W} - \boldsymbol{\mu}_{\sim j}\right) \end{bmatrix} + \boldsymbol{\mu}^{(j)}; \quad j = 1, \ldots, d\,. \tag{10}$$

Based on Equation (10), we have $\frac{\partial \mathbf{X}_{\sim j}}{\partial x_j} = \frac{\mathcal{L}^{(j)}_{\sim 1,1}}{\sqrt{\Sigma_{j,j}}} = \frac{\Sigma^{(j)}_{\sim 1,1}}{\Sigma_{j,j}} = \frac{\Sigma_{\sim j,j}}{\Sigma_{j,j}}$. Thus, we can deduce that $J^{(j)} = \frac{\Sigma_{\bullet,j}}{\Sigma_{j,j}}$ with $\Sigma_{\bullet,j}$ the $j^{\text{th}}$ column of $\Sigma$, and the dependent Jacobian becomes $J^d = \left[J^{(1)}, \ldots, J^{(d)}\right] = \left[\frac{\Sigma_{\bullet,1}}{\Sigma_{1,1}}, \ldots, \frac{\Sigma_{\bullet,d}}{\Sigma_{d,d}}\right] = \frac{\Sigma}{s^2} = \mathcal{R}$, as $\Sigma_{j,j} = s_j^2 = s^2$ and

$\Sigma = s^2\mathcal{R}$. The tensor metric is given by $G = \mathcal{R}^T\mathcal{R}$.

*5.3. Comparisons between exact gradient and estimated gradients*

For running the above PDE-based model using the R-package deSolve ([45]), we are given $D = 0.0011$ and $s = 1.96$. The exact and formal gradient associated with the mean values of the initial conditions is obtained by running the corresponding adjoint model. For estimating the gradient using the proposed estimators, we consider $L = 2, 3$ and $N = 50, 100, 150, 200$. We also use $h = 1/\sqrt{N}$ and $V_j \sim \mathcal{N}(0, 1)$, $j = 1, \ldots, d = 50$. The Sobol sequence is used for generating the random values of $V_j$'s, and the Gram-schmidt algorithm is applied to obtain perfect orthogonal vectors for a given $N$.

Figure 1 shows the comparisons between the estimated and the exact values of the formal gradient $\nabla f$ (i.e., $\rho(X_{j_1}, Z_{j_2}) = 0$) for $L = 1, 2$. Likewise, Figures 2-3 depict the dependent gradient $grad(f) = \left(\mathcal{R}^T\mathcal{R}\right)^{-1}\nabla f$ and its estimation. The estimates of both gradients are in line with the exact values using only $NL = 50$ (resp. $NL = 100$) model evaluations when $L = 1$ and $N = 50$ (resp. $L = 1$ and $N = 100$ or $L = 2$ and $N = 50$). Increasing the values of $L$ and $N$ gives the same quasi-perfect results for both the formal and dependent gradients (see Figure 3).

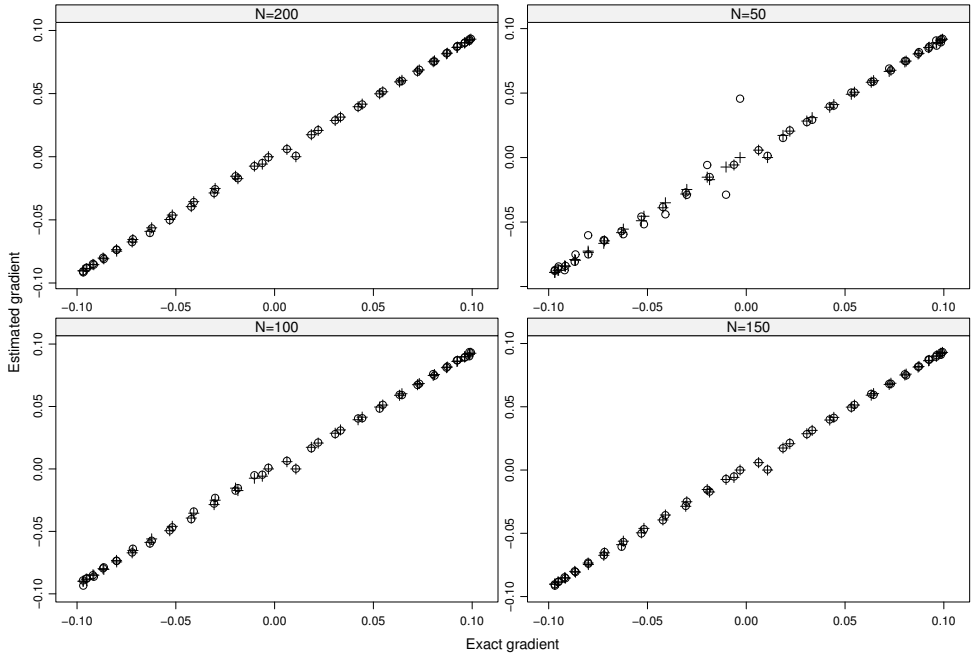


Figure 1: Exact gradient versus estimated gradients using $L = 1$ ($\circ$) and $L = 2$ ($+$) of the QoI by considering the inputs as independent (formal gradients).

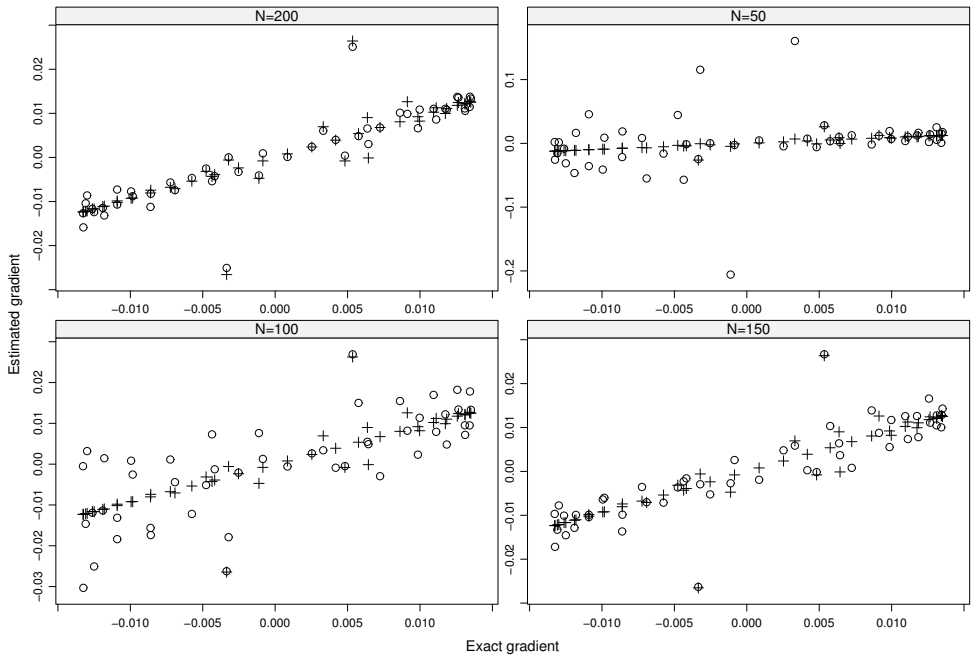


Figure 2: Exact gradient versus estimated gradients using $L = 1$ ($\circ$) and $L = 2$ ($+$) of the QoI by considering the auto-correlations anong the inputs (dependent gradients).

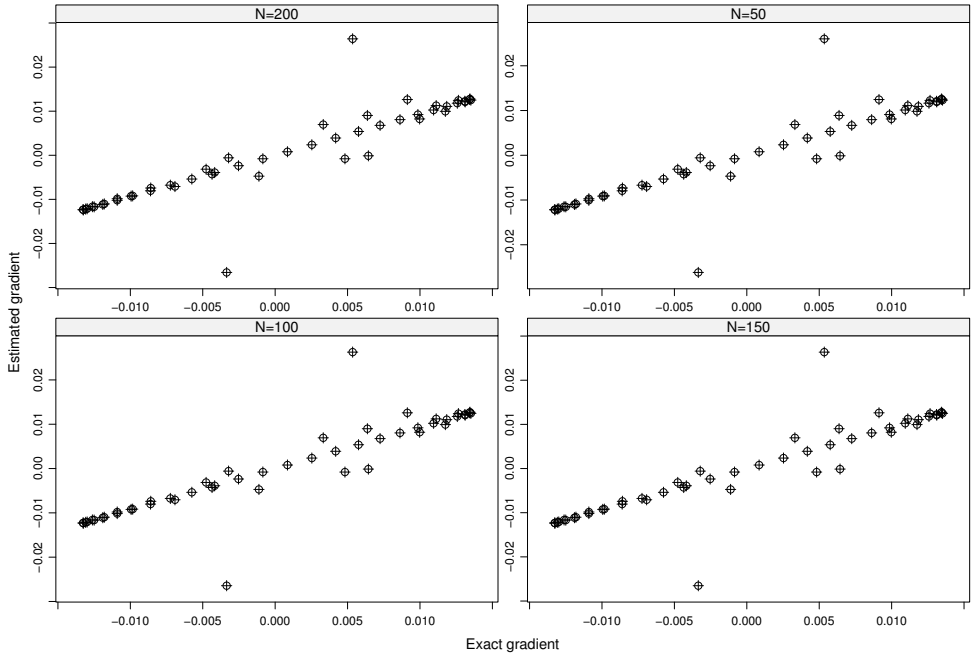


Figure 3: Exact gradient versus estimated gradients using $L = 2$ ($\circ$) and $L = 3$ ($+$) of the QoI by considering the auto-correlations anong the inputs (dependent gradients).

## 6. Conclusion

In this paper, we have proposed new, simple and generic approximations of the gradients of functions with non-independent input variables by means of independent, central and symmetric variables and a set of constraints. It comes out that the biases of our approximations for a wide class of functions, such as 2-smooth functions, do not suffer from the curse of dimensionality by properly choosing the set of independent, central and symmetric variables. For functions including only independent input variables, a theoretical comparison has shown that the upper-bounds of the bias of the formal gradient derived in this paper outperform the best known results.

For computing the dependent gradient of the function of interest, we have provided estimators of such a gradient by making use of evaluations of that function at $LN$ randomized points. Such estimators reach the optimal (mean squared error) rates of convergence (i.e., $\mathcal{O}(N^{-1}d^2)$) for a wide class of functions. Numerical comparisons using a test case and simulations based on a PDE model with given auto-collaborations among the initial conditions have shown the efficiency of our approach, even when $L = 1, 2$ constraints are used. Our approach is then flexible thanks to $L$ and the fact that the gradient can be computed for every value of the sample size $N$ in general.

While the proposed estimators reach the parametric rate of convergence, note that the second-order moments of such estimators depend on $d^2$. An attempt to reach a dimension-free rate of convergence requires working in $\mathbb{C}$ rather than $\mathbb{R}$ when $L = 2$. In next future, it is worth investigating the derivation of the optimal rates of convergence that are dimension-free or (at least) are linear with respect to $d$ by considering $L > 3$ constraints. Also, combining such a promising approach with a transformation of the original space might be helpful for reducing the number of model evaluations in higher dimensions.

## Acknowledgments

We would like to thank the reviewers for their comments that have helped improving our manuscript.

## Appendix A Proof of Theorem 1

As $\vec{\imath} = (\imath_1, \ldots, \imath_d)$, let $\vec{k} = \left(0, \ldots, 0, \underbrace{1}_{\text{k th position}}, 0 \ldots, 0\right) \in \mathbb{R}^d$ and $\vec{\mathbf{q}} = (q_1, \ldots, q_d) \in \mathbb{N}^d$. Multiplying the Taylor expansion of $f(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V})$ about $\mathbf{x}$, that is,

$$f(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}) = \sum_{p=0}^{m} \sum_{||\vec{\boldsymbol{\imath}}||_1 = p} \frac{\mathcal{D}^{(\vec{\boldsymbol{\imath}})} f(\mathbf{x})}{\vec{\boldsymbol{\imath}}!} \beta_\ell^p (\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{\imath}}} + \mathcal{O}\left(||\beta_\ell \boldsymbol{h}\mathbf{V}||_2^{m+1}\right) ,$$

by $\frac{\mathbf{V}\boldsymbol{h}^{-1}}{\sigma^2} \in \mathbb{R}^d$ and the constant $C_\ell$, and taking the sum over $\ell = 1, \ldots, L$, we can see that the expectation $E := \sum_{\ell=1}^{L} C_\ell \mathbb{E}\left[f(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}) \frac{\mathbf{V}\boldsymbol{h}^{-1}}{\sigma^2}\right]$ becomes

$$E = \sum_{p \geq 0} \sum_{||\vec{\boldsymbol{\imath}}||_1 = p} \frac{\mathcal{D}^{(\vec{\boldsymbol{\imath}})} f(\mathbf{x})}{\vec{\boldsymbol{\imath}}!} \left(\sum_{\ell} C_\ell \beta_\ell^p\right) \mathbb{E}\left[\frac{(\mathbf{V})^{\vec{\boldsymbol{\imath}}} (\boldsymbol{h})^{\vec{\boldsymbol{\imath}}} \mathbf{V}\boldsymbol{h}^{-1}}{\sigma^2}\right] .$$

Firstly, for a given $k \in \{1, \ldots, d\}$ and by independence, we can see that

$$\mathbb{E}\left[(\mathbf{V})^{\vec{\boldsymbol{\imath}}} (\boldsymbol{h})^{\vec{\boldsymbol{\imath}}} V_k h_k^{-1}\right] = \mathbb{E}\left[(\mathbf{V})^{\vec{\boldsymbol{\imath}} + \vec{k}} (\boldsymbol{h})^{\vec{\boldsymbol{\imath}} - \vec{k}}\right] \neq 0$$

iff $\imath_k = 2q_k + 1$; $\imath_j = 2q_j$ for any $j \in \{1, \ldots, d\} \setminus \{k\}$ with $q_k, q_j \in \mathbb{N}$, which implies that $\vec{\boldsymbol{\imath}} = \vec{k} + 2\vec{\mathbf{q}}$. Thus, one obtains $\frac{\partial f}{\partial x_k}$ when $||\vec{\boldsymbol{\imath}}||_1 = ||\vec{k} + 2\vec{\mathbf{q}}||_1 = 1$, and the fact that $\mathbb{E}[V_k^2] = \sigma^2$; $\mathbb{E}[V_j] = 0$ and $\sum_\ell C_\ell \beta_\ell = 1$. At this point, by taking $k = 1, \ldots, d$ and setting $L = 1$, $\beta_\ell = 1$ and $C_\ell = 1$ result in the approximation of $\nabla f(\mathbf{x})$ of order $\mathcal{O}(||h||_2^2)$ because when $||\vec{\boldsymbol{\imath}}||_1 = 2$, $\mathbb{E}\left[(\mathbf{V})^{\vec{\boldsymbol{\imath}}} (\boldsymbol{h})^{\vec{\boldsymbol{\imath}}} V_k h_k^{-1}\right] = 0$.
Secondly, for $L > 1$ the constraints $\sum_{\ell=1}^{L} C_\ell \beta_\ell^{r+1} = \delta_{0,r}$ $r = 0, 2, \ldots, 2(L-1)$ allow to eliminate some higher-order terms so as to reach the order $\mathcal{O}\left(||\mathbf{h}||_2^{2L}\right)$. Using other constraints complete the proof, bearing in mind Equation (2).

## Appendix B Proof of Corollary 1

For $\vec{\mathbf{q}} = (q_1, \ldots, q_d) \in \mathbb{N}^d$; $k \in \{1, \ldots, d\}$ and $\vec{k} = \left(0, \ldots, 0, \underbrace{1}_{\text{k th position}}, 0 \ldots, 0\right) \in \mathbb{R}^d$, consider $\boldsymbol{s}_k := \left\{\vec{\mathbf{q}} + \vec{k} \ : \ ||\vec{\mathbf{q}}||_1 = 1\right\}$. As $f \in \mathcal{H}_2$, we can write

$$f(\mathbf{x} + \beta_\ell \boldsymbol{h}\mathbf{V}) = \sum_{||\vec{\boldsymbol{\imath}}||_1 = 0}^{1} \mathcal{D}^{(\vec{\boldsymbol{\imath}})} f(\mathbf{x}) \beta_\ell^{||\vec{\boldsymbol{\imath}}||_1} \frac{(\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{\imath}}}}{\vec{\boldsymbol{\imath}}!} + \sum_{\substack{||\vec{\boldsymbol{\imath}}||_1 = 2 \\ \vec{\boldsymbol{\imath}} \notin \boldsymbol{s}_k}} \mathcal{D}^{(\vec{\boldsymbol{\imath}})} f(\mathbf{x}) \frac{\beta_\ell^2 (\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{\imath}}}}{\vec{\boldsymbol{\imath}}!} + R_k(\boldsymbol{h}, \beta_\ell, \mathbf{V}) ,$$

with the remainder term

$$\begin{aligned} R_k\left(\boldsymbol{h},\beta_\ell,\mathbf{V}\right) &= \sum_{\substack{||\vec{\boldsymbol{i}}||_1=2 \\ \vec{\boldsymbol{i}}\in \boldsymbol{s}_k}} \mathcal{D}^{(\vec{\boldsymbol{i}})} f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}) \frac{\beta_\ell^2 (\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{i}}}}{\vec{\boldsymbol{i}}!} = \sum_{\substack{||\vec{\boldsymbol{q}}||_1=1 \\ \vec{\boldsymbol{i}}\in \boldsymbol{s}_k}} \mathcal{D}^{(\vec{k}+\vec{\boldsymbol{q}})} f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}) \frac{\beta_\ell^2 (\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{q}}+\vec{k}}}{(\vec{k}+\vec{\boldsymbol{q}})!} \\ &= h_k V_k \beta_\ell^2 \sum_{||\vec{\boldsymbol{q}}||_1=1} \mathcal{D}^{(\vec{k}+\vec{\boldsymbol{q}})} f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}) \frac{(\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{q}}}}{(\vec{k}+\vec{\boldsymbol{q}})!}. \end{aligned}$$

Denote $R_k^0 := \sum_{||\vec{\boldsymbol{q}}||_1=1} \mathcal{D}^{(\vec{k}+\vec{\boldsymbol{q}})} f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}) \frac{(\boldsymbol{h}\mathbf{V})^{\vec{\boldsymbol{q}}}}{(\vec{k}+\vec{\boldsymbol{q}})!}$, and remark that $|R_k^0| \leq M_2 \left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_1$.
Using Theorem 1, we can see that the absolute value of the bias, that is,
$B := \left|\left| grad(f)(\mathbf{x}) - G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L} C_\ell \, \mathbb{E}\left[ f\left(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_1$ is given by

$$\begin{aligned} B &= \left|\left| G^{-1}(\mathbf{x}) \mathbb{E}\left[ \nabla f(\mathbf{x}) - \sum_{\ell=1}^{L} C_\ell f\left(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_1 \\ &= \left|\left| G^{-1}(\mathbf{x}) \sum_{\ell=1}^{L} C_\ell \frac{\beta_\ell^2}{\sigma^2} \mathbb{E}\left[ V_1^2 R_1^0, \ldots, V_d^2 R_d^0 \right]^T \right|\right|_1 \\ &\leq \sum_{\ell=1}^{L} |C_\ell| \frac{\beta_\ell^2}{\sigma^2} \left|\left| G^{-1}(\mathbf{x}) \mathbb{E}\left[ V_1^2 R_1^0, \ldots, V_d^2 R_d^0 \right]^T \right|\right|_1 \\ &\leq \sum_{\ell=1}^{L} |C_\ell| \beta_\ell^2 M_2 \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{E}\left[ \frac{\mathbf{V}^2}{\sigma^2} \left|\left| \boldsymbol{h}\mathbf{V} \right|\right|_1 \right] \right|\right|_1, \end{aligned}$$

using the expansion of the product between matrices.
Using the same reasoning and taking the Euclidean norm, we obtain

$$\begin{aligned} B_2 &= \left|\left| G^{-1}(\mathbf{x}) \mathbb{E}\left[ \nabla f(\mathbf{x}) - \sum_{\ell=1}^{L} C_\ell f\left(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V}\right) \frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2} \right] \right|\right|_2 \\ &\leq \sum_{\ell=1}^{L} |C_\ell| \beta_\ell^2 M_2 \left|\left| \left| G^{-1}(\mathbf{x}) \right| \mathbb{E}\left[ \frac{\mathbf{V}^2}{\sigma^2} \left|\left| \boldsymbol{h}\mathbf{V} \right|\right|_1 \right] \right|\right|_2. \end{aligned}$$

The results hold using $\mathbf{R} := \mathbf{V}/\sigma$.

## Appendix C Proof of Theorem 2

Firstly, remark that $MSE := \mathbb{E}\left[ \left|\left| \widehat{grad(f)}(\mathbf{x}) - grad(f)(\mathbf{x}) \right|\right|_2^2 \right]$ is given by

$$MSE = \mathbb{E}\left[ \left|\left| \widehat{grad(f)}(\mathbf{x}) - \mathbb{E}\left[ \widehat{grad(f)}(\mathbf{x}) \right] \right|\right|_2^2 + \left|\left| \mathbb{E}\left[ \widehat{grad(f)}(\mathbf{x}) \right] - grad(f)(\mathbf{x}) \right|\right|_2^2 \right].$$

Since, the bias $\mathbb{E}\left[\left|\left|\mathbb{E}\left[\widehat{grad(f)}(\mathbf{x})\right]-grad(f)(\mathbf{x})\right|\right|_2^2\right]$ has been derived in previous Corollaries, we are going to treat the second-order moment.
Secondly, as $f\in\mathcal{H}_2$ implies that $f\in\mathcal{H}_1$, we have $|f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})-f(\mathbf{x})|\leq M_1\left|\left|\beta_\ell \boldsymbol{h}\mathbf{V}\right|\right|_2$ . Also, as $\sum_{\ell=1}^{L}C_\ell=0$, we then have

$$\sum_{\ell=1}^{L}C_\ell f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})=\sum_{\ell=1}^{L}C_\ell\left[f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})-f(\mathbf{x})\right],$$

which leads to $\left|\sum_{\ell=1}^{L}C_\ell^{(|u|)}f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})\right|\leq\sum_{\ell=1}^{L}|C_\ell\beta_\ell|\,M_1\left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_2$ and

$$Q(\mathbf{x}):=G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\sum_{\ell=1}^{L}C_\ell f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})=G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\sum_{\ell=1}^{L}C_\ell\left[f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})-f(\mathbf{x})\right]. \tag{11}$$

Thirdly, using (3), we can see that $\mathbb{E}\left[Q(\mathbf{x})\right]=\mathbb{E}\left[\widehat{grad(f)}(\mathbf{x})\right]$. Bearing in mind the definition of the Euclidean norm and the variance, the centered second-order moment, that is, $\mathbb{V}_{grad}:=\mathbb{E}\left[\left|\left|\widehat{grad(f)}(\mathbf{x})-\mathbb{E}\left[\widehat{grad(f)}(\mathbf{x})\right]\right|\right|_2^2\right]$ is given by

$$\begin{aligned}
\mathbb{V}_{grad} &\leq \frac{1}{N}\mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\sum_{\ell=1}^{L}C_\ell f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})-\mathbb{E}\left[\widehat{grad(f)}(\mathbf{x})\right]\right|\right|_2^2\right]\\
&\leq \frac{1}{N}\mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\sum_{\ell=1}^{L}C_\ell f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})\right|\right|_2^2\right]\\
&\overset{(11)}{=} \frac{1}{N}\mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\sum_{\ell=1}^{L}C_\ell\left\{f(\mathbf{x}+\beta_\ell \boldsymbol{h}\mathbf{V})-f(\mathbf{x})\right\}\right|\right|_2^2\right]\\
&\leq \frac{1}{N}\mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\right|\right|_2^2\left|\left|\boldsymbol{h}\mathbf{V}\right|\right|_2^2\right]M_1^2\left(\sum_{\ell=1}^{L}|C_\ell\beta_\ell|\right)^2\\
&\leq \frac{1}{N}\mathbb{E}\left[\left|\left|G^{-1}(\mathbf{x})\frac{\mathbf{V}\mathbf{h}^{-1}}{\sigma^2}\right|\right|_2^2\left|\left|\mathbf{V}^2\right|\right|_2\right]\left|\left|\boldsymbol{h}^2\right|\right|_2M_1^2\left(\sum_{\ell=1}^{L}|C_\ell\beta_\ell|\right)^2
\end{aligned}$$

bearing in mind the Hölder inequality. The results hold using $\mathbf{R}:=\mathbf{V}/\sigma$, and the fact that when $V_k\sim\mathcal{U}(-\xi,\xi)$, $\left|\left|\mathbf{V}^2\right|\right|_2\leq\sqrt{d}\xi^2$ and $\sigma^2=\xi^2/3$.

## References


[1] M. Rosenblatt, Remarks on a multivariate transformation, Ann. Math. Statist. 23 (3) (1952) 470–472.

[2] A. Nataf, Détermination des distributions dont les marges sont données, Comptes Rendus de l'Académie des Sciences 225 (1962) 42–43.

[3] H. Joe, Dependence Modeling with Copulas, Chapman & Hall/CRC, London, 2014.

[4] A. J. McNeil, R. Frey, P. Embrechts, Quantitative Risk Management, Princeton University Press, Princeton and Oxford, 2015.

[5] J. Navarro, J. M. Ruiz, Y. D. Aguila, Multivariate weighted distributions: a review and some extensions, Statistics 40 (1) (2006) 51–64.

[6] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, Publications de l'Institut Statistique de l'Université de Paris 8 (1959) 229–231.

[7] F. Durante, C. Ignazzi, P. Jaworski, On the class of truncation invariant bivariate copulas under constraints, Journal of Mathematical Analysis and Applications 509 (1) (2022) 125898.

[8] A. V. Skorohod, On a representation of random variables, Theory Probab. Appl 21 (3) (1976) 645–648.

[9] M. Lamboni, S. Kucherenko, Multivariate sensitivity analysis and derivative-based global sensitivity measures with dependent variables, Reliability Engineering & System Safety 212 (2021) 107519.

[10] M. Lamboni, Efficient dependency models: Simulating dependent random variables, Mathematics and Computers in Simulation (2022). doi:https://doi.org/10.1016/j.matcom.2022.04.018.

[11] M. Lamboni, On exact distribution for multivariate weighted distributions and classification, Methodology and Computing in Applied Probability 25 (2023) 1–41.

[12] M. Lamboni, Measuring inputs-outputs association for time-dependent hazard models under safety objectives using kernels, International Journal for Uncertainty Quantification - (2024) 1–17. doi:10.1615/Int.J.UncertaintyQuantification.2024049119.

[13] S. Kucherenko, O. Klymenko, N. Shah, Sobol' indices for problems defined in non-rectangular domains, Reliability Engineering & System Safety 167 (2017) 218 – 231.

[14] M. Lamboni, On dependency models and dependent generalized sensitivity indices, arXiv preprint arXiv2104.12938 (2021).

[15] M. Lamboni, Derivative formulas and gradient of functions with non-independent variables, Axioms 12 (9) (2023). doi:10.3390/axioms12090845.

[16] A. Nemirovsky, D. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley & Sons, New York, 1983.

[17] E. Patelli, H. Pradlwarter, Monte Carlo gradient estimation in high dimensions, International Journal for Numerical Methods in Engineering 81 (2) (2010) 172–188.

[18] A. Agarwal, O. Dekel, L. Xiao, Optimal algorithms for online convex optimization with multi-point bandit feedback., in: Colt, Citeseer, 2010, pp. 28–40.

[19] F. Bach, V. Perchet, Highly-smooth zero-th order online optimization, in: V. Feldman, A. Rakhlin, O. Shamir (Eds.), 29th Annual Conference on Learning Theory, Vol. 49, 2016, pp. 257–283.

[20] A. Akhavan, M. Pontil, A. B. Tsybakov, Exploiting higher order smoothness in derivative-free optimization and continuous bandits, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[21] I. M. Sobol, S. Kucherenko, Derivative based global sensitivity measures and the link with global sensitivity indices, Mathematics and Computers in Simulation 79 (2009) 3009–3017.

[22] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides, N. Shah, Monte Carlo evaluation of derivative-based global sensitivity measures, Reliability Engineering and System Safety 94 (2009) 1135–1148.

[23] M. Lamboni, B. Iooss, A.-L. Popelin, F. Gamboa, Derivative-based global sensitivity measures: General links with Sobol' indices and numerical tests, Mathematics and Computers in Simulation 87 (0) (2013) 45 – 54.

[24] O. Roustant, J. Fruth, B. Iooss, S. Kuhnt, Crossed-derivative based sensitivity measures for interaction screening, Mathematics and Computers in Simulation 105 (2014) 105 – 118.

[25] J. Fruth, O. Roustant, S. Kuhnt, Total interaction index: A variance-based sensitivity index for second-order interaction screening, Journal of Statistical Planning and Inference 147 (2014) 212 – 223.

[26] M. Lamboni, Derivative-based generalized sensitivity indices and Sobol' indices, Mathematics and Computers in Simulation 170 (2020) 236 – 256.

[27] M. Lamboni, Derivative-based integral equalities and inequality: A proxy-measure for sensitivity analysis, Mathematics and Computers in Simulation 179 (2021) 137 – 161.

[28] M. Lamboni, Weak derivative-based expansion of functions: ANOVA and some inequalities, Mathematics and Computers in Simulation 194 (2022) 691–718.

[29] S. Bobkov, Isoperimetric and analytic inequalities for log-concave probability measures, The Annals of Probability 27 (1999) 1903–1921.

[30] O. Roustant, F. Barthe, B. Iooss, Poincaré inequalities on intervals - application to sensitivity analysis, Electron. J. Statist. 11 (2) (2017) 3081–3119.

[31] F.-X. Le Dimet, O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, Tellus A: Dynamic Meteorology and Oceanography 38 (2) (1986) 97–110.

[32] F. X. Le Dimet, H. E. Ngodock, B. Luong, J. Verron, Sensitivity analysis in variational data assimilation, Journal-Meteorological Society of Japan 75 (1997) 245–255.

[33] D. G. Cacuci, Sensitivity and uncertainty analysis - Theory, Chapman & Hall/CRC, 2005.

[34] M. D. Gunzburger, Perspectives in flow control and optimization, SIAM, Philadelphia, 2003.

[35] A. Borzi, V. Schulz, Computational Optimization of Systems Governed by Partial Differential Equations, SIAM, Philadelphia, 2012.

[36] R. Ghanem, D. Higdon, H. Owhadi, Handbook of Uncertainty Quantification, Springer International Publishing, 2017.

[37] E. Guidotti, calculus: High-dimensional numerical and symbolic calculus in R, Journal of Statistical Software 104 (5) (2022) 1–37.

[38] B. Ancell, G. J. Hakim, Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting, Monthly Weather Review 135 (2007) 4117–4134.

[39] H. Pradlwarter, Relative importance of uncertain structural parameters. part i: algorithm, Computational Mechanics 40 (2007) 627–635.

[40] B. Polyak, A. Tsybakov, Optimal accuracy orders of stochastic approximation algorithms, Probl. Peredachi Inf. (1990) 45–53.

[41] A. Zemanian, Distribution Theory and Transform Analysis: An Introduction to Generalized Functions, with Applications, Dover Books on Advanced Mathematics, Dover Publications, 1987.

[42] R. Strichartz, A Guide to Distribution Theory and Fourier Transforms, Studies in advanced mathematics, CRC Press, Boca, 1994.

[43] M. Lamboni, Kernel-based measures of association between inputs and outputs using anova, Sankhya A - (2024). doi:10.1007/s13171-024-00354-w.

[44] P. Gilbert, R. Varadhan, R-package numDeriv: Accurate Numerical Derivatives, CRAN Repository, 2019.

[45] K. Soetaert et al., R-package deSolve: Solvers for Initial Value Problems of Differential Equations, CRAN Repository, 2022.