

**WORKING PAPERS**

N° 1751

June 2026

## “On-line Pick-Freeze Mirror algorithm for Sensitivity Analysis”

Manon Costa, Sébastien Gadat, Xavier Gendre and Thierry Klein

# On-line Pick-Freeze Mirror algorithm for Sensitivity Analysis

Manon Costa<sup>1</sup>, Sébastien Gadat<sup>2</sup>, Xavier Gendre<sup>3,4</sup>, and Thierry Klein<sup>3,5</sup>

<sup>1</sup>Univ Toulouse, INUC, UT2J, INSA Toulouse, TSE, CNRS, IMT, Toulouse, France.

<sup>2</sup>Toulouse School of Economics, CNRS UMR 5314, Université Toulouse 1 Capitole, Toulouse.  
Institut Universitaire de France.

<sup>3</sup>Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse

<sup>4</sup>Pathway, Paris, France

<sup>5</sup>ENAC - Ecole Nationale de l'Aviation Civile, Université de Toulouse, France

December 9, 2025

## Abstract

The main objective of this paper is to propose a new approach for estimating the entire collection of Sobol' indices simultaneously. Our approach exploits the fact that Sobol' indices can be rewritten as solutions to an optimization problem over the simplex of  $\mathbb{R}^d$ , to construct an online sequence of estimators using a stochastic mirror descent algorithm. We prove that our estimation procedure is consistent and provide a non-asymptotic upper bound for its rate of convergence. Furthermore, we demonstrate the numerical accuracy of our method and compare it with other classical estimation procedures. <sup>1</sup>

**Keywords:** global sensitivity analysis; Sobol' indices; stochastic Mirror descent algorithm .

**AMS subject classification:** 62G05, 62L12, 62G20.

## 1 Introduction

### 1.1 Motivation

The study of how a numerical code's output depends on its input variables has become critically important in many fields, including physics, engineering, applied mathematics, and signal and image processing, among others. Sensitivity analysis is particularly important in fields where meta-models

---

<sup>1</sup>**Acknowledgment** The authors thank Paul Rochet for recognizing the Möbius function as well as Sébastien Da Veiga for giving them access to the results of the numerical experiments performed in [10] and running his R code on the test function of Section 4.2. This work benefits from the financial support of ANR project GATSBII (ANR-24-CE23-6645). S. G. also wishes to thank the Lagrange Center for its support during the development of this research project.

are used, such as modeling coastal flooding [3], optimizing aircraft geometry in aeronautics [43], and in various other areas of engineering. Further examples can be found in [4, 37, 25]. The importance of sensitivity analysis is growing with the advent of artificial intelligence techniques, as the use of deep neural networks with highly sophisticated architectures is gradually replacing costly, traditional, physics-based codes for computing outputs through a regression paradigm, for example [17, 16]. However, the complexity of these AI-generated models is often perceived as mysterious and difficult to interpret, particularly with regard to how the output is influenced by the input variables. It is nevertheless crucial to understand the effects of input variables on a code’s output, in order to improve the generation of certain input variables and to enhance public acceptance of how the code operates. Therefore, it is important to quantify the influence of variables on the system with a reasonably limited number of evaluations and computations.

When these inputs are regarded as random elements, this problem is generally referred to as global sensitivity analysis. Global sensitivity analysis considers the input vector as random and provides a measure of the influence, in terms of output fluctuations, of each subset of its components. We refer to the seminal book [48] for an overview, to [8] for a synthesis of recent trends in this field, and to [12, 48, 50, 52] for a discussion of the practical aspects of global sensitivity analysis.

Of the various measures used in global sensitivity analysis, variance-based measures, derived from Hoeffding’s decomposition of variance, are probably the most commonly used. When considering an output  $Y$  of a computer code modeled as  $Y = f(X^1, \dots, X^p)$ , one obtains one of the most common measures of the sensitivity of  $Y$  with respect to a subset of inputs  $X^i, i \in u$  for  $u \subset 1, \dots, p$ : the closed Sobol’ index  $\Sigma_u$ , defined by

$$\Sigma_u = \frac{\text{Var}(\mathbb{E}[Y|X^i, i \in u])}{\text{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X^i, i \in u]^2] - \mathbb{E}[Y]^2}{\text{Var}(Y)}.$$

In recent years, a myriad of different estimators have been proposed (see [8, Chapter 4] for a full review). A first class gathers spectral methods that aim at constructing estimators based on the spectral analysis of the input/output functional relationship and Parseval’s formula. We refer to [55] for a basic description of the method and to [8, Chapter 4] for more recent references. It should be noted that the asymptotic properties of these methods have been little studied, as they are based on the theory of non-linear (quadratic) functional estimation. A second class of methods requires that the covariates  $(X^1, \dots, X^p)$  be independent in order to obtain the asymptotic properties of the estimators. Among them, the so-called Pick–Freeze design of experiments aims at computing a Monte Carlo estimate of  $\Sigma_u$  from a sequence of evaluations of  $f$  on a sample of input values, fixing those of the subset  $u$ . The main advantage is that only minimal assumptions on  $f$  are required to derive consistency and central limit theorems. In particular, assumptions of integrability, but not regularity, on  $f$  are necessary (see, for example, [26, 32]). Moreover, to estimate at rate  $\sqrt{n}$  a single Sobol’ index, one needs a design of experiments of size  $2n$ . This implies that estimating all the  $p$  first-order indices involves a sample of size  $(p + 1)n$ , and a  $2^p n$  sample is required to estimate all Sobol’ indices. This method has two drawbacks: the size of the DOE required to estimate all indices grows exponentially with the number of input variables, and it relies on the very specific Pick–Freeze experimental design.

In order to tackle these limitations, methods based on local averaging have been developed. Among them are kernel estimators, which have been thoroughly studied for the case  $d = 1$  [11, 7, 45, 54, 30], and provide central limit theorems and asymptotic efficiency as soon as the function  $f$  satisfies regularity assumptions. Closely related are nearest- approaches, which have been studied by several authors (see, e.g., [15, 34, 35, 13, 22, 29, 14]). For instance, in [14], the authors propose a plug-in estimator with statistical consistency for any  $d$ , and a central limit theorem with rate  $\sqrt{n}$  for  $d \leq 3$ , provided that regularity assumptions hold. In parallel, [4] consider a variant that is consistent for any  $d$ , but no rate of convergence is provided. When  $d = 1$ , a central limit theorem for estimators based on ranks (i.e., nearest neighbors on the right) is also proved in [24]. Recently, in [9], the authors used high-order kernels to build asymptotically efficient estimators of Sobol' indices of any dimension from a single i.i.d.  $n$  input/output sample. Their result requires strong smoothness assumptions on  $f$  and on the support of the input variables which may be restrictive in some irregular situations.

In order to build an estimator of Sobol' indices using any of the methods described above, the practitioner needs to have access to the complete dataset. Unfortunately, in some applications, the data are generated online and cannot be stored. The goal of this work is to introduce and to study a new numerical scheme based on online data for simultaneously estimating the  $2^p$  Sobol' indices taking into account their geometrical constraints.

## 1.2 Main contributions

In this paragraph, we provide an overview of our contributions, acknowledging that some concepts may not be fully introduced at this stage. Our aim here is to offer a concise and accessible summary that enables the reader to grasp the core contributions of the paper within a short time. Readers seeking a more detailed and systematic exposition will find a thorough presentation in Sections 2 through 5.

**Framework** In this work, we propose a sequential estimation method for all Sobol' indices associated with a numerical model  $Y = f(X)$ , where  $X = (X^1, \dots, X^p)$  is a random vector with an unknown distribution and *independent* components in  $\mathbb{R}^p$ , and  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  is a deterministic and unknown function such that  $\mathbb{E}[f(X)^4] < \infty$ . Our method generates a sequence of estimators  $(\hat{\mathbf{S}}_n)_{n \geq 1}$ , where, at each iteration  $n$ ,  $\hat{\mathbf{S}}_n$  denotes the current estimate of the collection of Sobol' indices, based on  $2n$  evaluations of the numerical model  $f$ . The true collection of Sobol' indices is denoted by  $\mathbf{S}^* = (S_u^*)_{u \in \Upsilon}$ , where  $\Upsilon$  represents the set of all possible subsets of  $1, \dots, p$ .

Our scheme combines two ingredients that were neglected in previous studies: first, the Hoeffding decomposition of the variance implies that the collection of Sobol' indices lies on the unit simplex; and second, Sobol' indices can be interpreted as the solution to an optimization problem, as pointed out by [32] in Equation (2.4). Indeed, the Hoeffding decomposition of the variance provides two key insights: the indices take values between zero and one, and they sum to one. This information is generally not taken into account in classical estimation procedures. Our online estimator is constructed via a stochastic mirror descent algorithm that solves a constrained optimization problem associated with the collection of Sobol' indices. The main advantage of the mirror descent algorithm is that it yields an explicit solution of the one-step optimization on the simplex without any projection, and

a single evaluation of the function  $f$  allows the simultaneous update of  $2^p$  indices, resulting in a significant reduction in computational cost. Our analysis requires access to two online sequences of i.i.d. realizations of the random variable  $X$ , denoted by  $(X_n)_{n \geq 1}$  and  $(X'_n)_{n \geq 1}$ , as well as the ability to perform sequential evaluations of the numerical model  $f$ .

**Variational description of Sobol' indices** Our first contribution, although theoretically straightforward, has significant practical value for implementing our sequential estimation method for  $\mathbf{S}^*$ . Specifically, it provides a variational characterization of  $\mathbf{S}^*$  through a constrained convex optimization problem, which generalizes the result stated in Equation (2.4) of [32]. For this purpose, we introduce  $\Delta_q$  as the set of probability distributions over  $\Upsilon$ , while  $\Delta_q^*$  denotes the same set restricted to vectors with non-zero coefficients.

**Theorem.** *Assume that  $a \in \Delta_q^*$  is a discrete positive probability distribution over  $\Upsilon$ . Then, a strongly convex function  $\Phi^a$  exists (defined in Equation (13)) such that  $\mathbf{S}^*$  is the unique solution to the convex program:*

$$\mathbf{S}^* = \arg \min_{s \in \Delta_q} \Phi^a(s). \quad (1)$$

The details of this statement are provided in Section 2, and in particular in Section 2.3, which presents a detailed version of the previous result stated in Theorem 2.2.

**Estimation of Sobol' indices with a Stochastic Pick-Freeze mirror descent** Our second, and more substantial, contribution is a practical algorithm that solves the aforementioned convex program via a sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$ , constructed from online arrivals of i.i.d. observations  $(X_n, X'_n)_{n \geq 1}$ . We employ a stochastic gradient descent procedure, tailored to our setting, as detailed in Algorithm 1 and developed in Section 3. The sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$  depends on a step-size sequence  $(\eta_n)_{n \geq 1}$ , for which we establish the following result.

**Theorem.** *The sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$  built in Section 3 satisfies*

(a) Asymptotic almost sure convergence result.

*Assume that  $\sum \eta_n = +\infty$  and  $\sum \eta_n^2 < +\infty$ , then:*

$$\lim_{n \rightarrow +\infty} \hat{\mathbf{S}}_n = \mathbf{S}^* \quad a.s. \quad (2)$$

(b) Non-asymptotic upper bound *Consider the Cesaro averaged sequence*

$$\bar{\mathbf{S}}_n^\eta = \frac{\left( \sum_{k=1}^n \eta_{k+1} \hat{\mathbf{S}}_k \right)}{\left( \sum_{k=1}^n \eta_{k+1} \right)}, \quad \forall n \geq 1,$$

*then if  $\eta_0 > 0$  and  $(\eta_n)_{n \geq 1} = (\eta_0 n^{-1/2})_{n \geq 1}$ , we have:*

$$\mathbb{E} \left[ \Phi^a(\bar{\mathbf{S}}_n^\eta) \right] - \min \Phi^a \leq C_d \frac{\log(n)}{\sqrt{n}}, \quad (3)$$

*where  $C_d$  is a dimension-dependent constant that will be made explicit later on.*

The two parts of the previous theorem are detailed in Theorems 3.3 and 3.4, where alternative step-size sequences are also considered.

### 1.3 Organization of the paper

The article is organized as follows. Section 2 introduces a variational characterization of the collection of all Sobol' indices as the minimizer of a strongly convex function (Corollary 2.1). This section also summarizes well-known key concepts from global sensitivity analysis. Section 3 presents the Pick–Freeze Mirror algorithm, proposed to solve the constrained optimization problem and to construct a sequence  $\mathbf{S}_n^*$  of estimators for the Sobol' indices. Section 4 then presents numerical experiments that illustrate the performance and practical behavior of the proposed algorithm on benchmark problems. Finally, Section 5 is devoted to the proofs of our theoretical results.

## 2 Constrained variational characterization of Sobol' indices

We consider  $X = (X^1, \dots, X^p)$ , a real-valued  $p$ -dimensional random vector representing the input of a numerical code, and  $Y$ , the output random variable associated with a black-box function  $f : \mathbb{R}^p \rightarrow \mathbb{R}$ . The output  $Y$  is linked to the input  $X$  through the simple relation

$$Y = f(X) \quad \text{where} \quad X = (X^1, \dots, X^p). \quad (4)$$

We assume that  $X$  follows an unknown distribution  $\mathbb{Q}$ , and that its components  $(X^j)_{j \leq p}$  are *independent*. For numerical purposes, we further assume that the variables  $(X^j)_{j \leq p}$  can be easily simulated. Our theoretical results require only that

$$\mathbb{E}[Y^4] < \infty. \quad (5)$$

**Subset notations** Throughout the paper, we frequently consider subsets of variables, denoted by  $u \subset \{1, \dots, p\} \setminus \emptyset$ . For convenience, we denote by  $\Upsilon$  the set of *all* subsets of  $\{1, \dots, p\}$ . When  $u \in \Upsilon$ ,  $|u|$  denotes the cardinality of the subset  $u$ , with the convention that  $|\emptyset| = 0$ .

### 2.1 Sobol' indices

A now-standard approach for global sensitivity analysis relies on the computation of Sobol' indices, a technique introduced in [42] and later popularized by [53]. These indices are particularly suitable when the numerical code returns a real-valued output, denoted by  $Y$ . In this section, we recall some key elements of global sensitivity analysis, which may be skipped by readers already familiar with the topic.

Using the ANOVA-Hoeffding decomposition of the variance (see *e.g.* [31, 21]), the objective of Sobol' indices is to compare the conditional variance of  $Y$ , when fixing the values of a subset of variables  $(X^i, i \in u)$ , to the total variance of  $Y$

$$\text{Var}(Y) = \sum_{u \in \Upsilon} V_u \quad \text{with} \quad V_u = \sum_{v \subset u} (-1)^{|u|-|v|} \text{Var}(\mathbb{E}[Y|X^i, i \in v]) \geq 0. \quad (6)$$

We shall observe that  $V_\emptyset = \text{Var}(\mathbb{E}[Y]) = 0$ , regardless the distribution of the inputs  $X$ , so that  $V_\emptyset$  is always zero and does not need any estimation procedure.

The Sobol'  $\mathbf{S}^* = (S_u^*)_{u \in \Upsilon}$  and closed Sobol'  $\Sigma^* = (\Sigma_u^*)_{u \in \Upsilon}$  indices with respect to  $(X^i, i \in u)$  are defined by:

$$S_u^* := \frac{V_u}{\text{Var}(Y)} \quad \text{and} \quad \Sigma_u^* := \frac{\text{Var}(\mathbb{E}[Y|X^i, i \in u])}{\text{Var}(Y)}. \quad (7)$$

Dividing both sides of (6) by  $\text{Var}(Y)$ , we observe that  $\mathbf{S}^*$  belongs to the simplex of discrete probability distributions over subsets defined by:

$$\Delta_q := \left\{ s = (s_u)_{u \in \Upsilon}, \forall u \in \Upsilon : s_u \geq 0 \quad \text{and} \quad s_\emptyset = 0 \quad \text{and} \quad \sum_{u \in \Upsilon} s_u = 1. \right\}.$$

Moreover, we also deduce from Equation (6) that Sobol' and closed Sobol' indices are linked through a linear transformation. Using an indexation with subsets of  $\Upsilon$  (see Appendix B for more details), a square matrix  $\mathbf{M}$ , triangular with diagonal terms equal to 1 exists such that:

$$\mathbf{S}^* = \mathbf{M}\Sigma^* \quad \text{with} \quad \forall (u, v) \in \Upsilon \times \Upsilon \quad \mathbf{M}_{u,v} = (-1)^{|u|-|v|} \mathbf{1}_{v \subset u}. \quad (8)$$

Finally, the matrix  $\mathbf{M}$  can be inverted using the Rota–Möbius inversion formula [47], thus

$$\forall (u, v) \in \Upsilon \times \Upsilon \quad \mathbf{M}_{u,v}^{-1} = \mathbf{1}_{v \subset u} \quad \text{and} \quad \Sigma^* = \mathbf{M}^{-1}\mathbf{S}^*. \quad (9)$$

## 2.2 Pick-Freeze trick

In this section, we briefly recall the classical "Pick-Freeze" strategy for estimating Sobol' indices. We consider  $X' = (X'^1, \dots, X'^p)$  an independent copy of  $X = (X^1, \dots, X^p)$  so that  $(X, X') \sim \mathbb{Q} \otimes \mathbb{Q}$ .

**Proposition 2.1** (Pick-Freeze trick). *For any  $u \in \Upsilon$ , we have:*

(i) *If  $Y = f(X)$ ,  $Y^{(u)} = f(X^{(u)})$  where  $X^{(u)}$  is built from  $X$  and  $X'$  as*

$$\forall i \in \{1, \dots, p\} \quad X^{(u),i} = X^i \mathbf{1}_{i \in u} + X'^i \mathbf{1}_{i \notin u},$$

*then  $\text{Var}(\mathbb{E}[Y|X^i, i \in u]) = \text{Cov}(Y, Y^{(u)})$  and*

$$\Sigma_u^* = \frac{\text{Cov}(Y, Y^{(u)})}{\text{Var}(Y)}. \quad (10)$$

*In what follows,  $(Y, Y^{(u)})$  will be referred to as a Pick-Freeze sample.*

(ii) *Furthermore*

$$\Sigma_u^* = \arg \min_{\theta \in [0,1]} \psi_u(\theta) \quad \text{with} \quad \psi_u(\theta) := \mathbb{E} \left[ \left( (Y - \mathbb{E}[Y])\theta - (Y^{(u)} - \mathbb{E}[Y]) \right)^2 \right],$$

*where the expectation has to be considered with respect to the pair of random variables  $(Y, Y^{(u)})$  defined above.*

For a fixed subset  $u \in \Upsilon$ , the “Pick–Freeze” trick leads to the so-called “Pick–Freeze” estimator, which corresponds to a Monte Carlo version of definition (10). These estimators have been extensively studied in the literature [53, 49, 38] and are known to be consistent and asymptotically normal [32]. In addition to their favorable asymptotic properties, it is worth noting that the number of evaluations of  $f$  required to estimate  $\Sigma_u^*$  is  $2n$ , yielding a convergence rate of  $\sqrt{n}$ . To estimate  $K$  Sobol’ indices, this number increases to  $(K + 1)n$ , so that the total computational cost reaches  $2^p n$ , which grows exponentially with  $p$ .

### 2.3 From the Pick-Freeze trick to a variational characterization of $\Sigma^*$ and $\mathbf{S}^*$

In this work, we propose a method for the simultaneous estimation of *all* Sobol’ indices, assuming access to a sequence of data compatible with the Pick–Freeze trick. The variational characterization of  $\Sigma^*$ , stated in Proposition 2.1, implies that a standard gradient descent method can be naturally applied to obtain an accurate approximation of  $\Sigma_u^*$ , since the functional  $\theta \mapsto \psi_u(\theta)$  is strongly convex.

We perform a convex aggregation of all the functions  $\psi_u$  using a collection of weights  $(a_u)_{u \in \Upsilon} \in \Delta_q$ , for now chosen arbitrarily, and define the function  $\Psi^a$  as follows:

$$\forall x = (x_u)_{u \in \Upsilon} \quad \Psi^a(x) = \frac{1}{2} \sum_{u \in \Upsilon} a_u \psi_u(x_u). \quad (11)$$

The probability distribution  $(a_u)_{u \in \Upsilon}$  which will be denoted by  $\mathcal{L}_\Upsilon^a$  is for now chosen arbitrarily such that:

$$\min_{u \in \Upsilon} a_u > 0. \quad (12)$$

For any such  $a$ , a straightforward consequence of Proposition 2.1 (ii) is that  $\Psi^a$  is minimized for  $x = \Sigma^*$ .

Finally, recall that  $\mathbf{S}^* = \mathbf{M}\Sigma^*$  and  $\mathbf{S}^* \in \Delta_q$ , the simplex of discrete probability distributions over the set  $\Upsilon$  and define

$$\forall s \in \Delta_q : \quad \Phi^a(s) := \Psi^a(\mathbf{M}^{-1}s). \quad (13)$$

This leads to the following characterization result for the set of *all* Sobol’ indices, which will serve as the cornerstone of our work.

**Corollary 2.2** (Variational characterization of  $\mathbf{S}^*$ ). *For any discrete positive probability distribution  $a = (a_u)_{u \in \Upsilon}$ , satisfying (12), the collection of Sobol’ indices  $\mathbf{S}^*$  is uniquely defined as:*

$$\mathbf{S}^* = \arg \min_{s \in \Delta_q} \Phi^a(s), \quad (14)$$

where  $\Phi^a$  can be written as an strongly convex function represented as an expectation:

$$\Phi^a(s) = \mathbb{E} \left[ \left( (Y - \mathbb{E}[Y])[ \mathbf{M}^{-1}s ]_U - (Y^{(U)} - \mathbb{E}[Y]) \right)^2 \right], \quad (15)$$

where  $U \sim \mathcal{L}_\Upsilon^a$  and conditionally to  $U$ ,  $(Y, Y^U)$  is a Pick–Freeze sample (see Proposition 2.1).

Corollary 2.2 shows that  $\mathbf{S}^*$  minimizes the function  $\Phi^a$ , which is strongly convex and differentiable, over the space  $\Delta_q$ . Moreover,  $\Phi^a$  is the expectation of a strongly convex function that can be readily simulated using Pick-Freeze realizations. Consequently, the problem of estimating all Sobol' indices can be addressed via a constrained stochastic optimization algorithm. This is precisely the approach we adopt in the following, employing a Stochastic Mirror Descent algorithm detailed in the next section.

### 3 On-line estimation of Sobol' indices with Pick-Freeze S.M.D.

Our estimation strategy is based on a mirror descent introduced in the pioneering work of [40]; it provides a versatile approach that naturally handles constrained optimization problems through an appropriate mirror map. Mirror descent can be seen as a natural generalization of the gradient descent method, seen as an iterative Maximization-Minimization procedure. It is particularly suited for the smooth minimization of a strongly convex function  $\Phi^a$  over the constrained set  $\Delta_q$  (see, *e.g.*, [33], [5]). The mirror descent algorithm defines a smooth trajectory that remains within the constrained set without requiring an additional projection step and effectively "pushes" the boundaries of the simplex to an infinite distance from any point strictly inside  $\Delta_q$ . Moreover, it can be easily extended to a *stochastic* setting using iterative noisy realizations of the gradients.

#### 3.1 Bregman divergence on $\Delta_q$ and mirror descent

We first introduce the strongly convex negative entropy function over  $\Delta_q$ :

$$\forall v \in \Delta_q : \quad h(v) := \sum_{i=1}^q v_i \log v_i. \quad (16)$$

If  $\langle \cdot, \cdot \rangle$  denotes the standard Euclidean inner product, the  $h$ -Bregman divergence  $\mathcal{D}_h$  between two probability distributions is defined as

$$\forall (w, v) \in \Delta_q^2, \quad \mathcal{D}_h(w, v) := h(w) - h(v) - \langle \nabla h(v), w - v \rangle.$$

**Remark 3.1.** *A Bregman divergence induces a natural metric associated to  $u \in \Delta_q$  while we emphasize that the strong convexity of  $h$  induces the following lower bound:*

$$\mathcal{D}_h(w, v) \geq \frac{\|w - v\|^2}{2}. \quad (17)$$

A (deterministic) mirror descent with a step-size sequence  $(\gamma_n)_{n \geq 1}$  consists of minimizing, from  $n$  to  $n + 1$ , the first-order Taylor approximation of  $\Phi^a$  penalized by the Bregman divergence. The iterative step corresponds to a sequence  $(\theta_n)_{n \geq 1}$ :

$$\theta_{n+1} = \arg \min_{\theta \in \Delta_q} \left\{ \Phi^a(\theta_n) + \langle \nabla \Phi^a(\theta_n), \theta - \theta_n \rangle + \frac{1}{\gamma_{n+1}} \mathcal{D}_h(\theta, \theta_n) \right\}. \quad (18)$$

Equation (18) defines a general “proximal” descent method built using the Bregman divergence  $\mathcal{D}_h$  on the function  $\Phi^a$ . When  $\mathcal{D}_h$  is replaced by the standard Euclidean distance, this reduces to the standard gradient descent with step-size sequence  $(\gamma_n)_{n \geq 1}$ . A remarkable feature, see *e.g.* [33, 6], is that the minimization in (18) can be made explicit when  $\mathcal{D}_h$  is derived from Equation (16). Using a Lagrangian formulation, one can indeed verify that:

$$\theta_{n+1} = \frac{\theta_n e^{-\gamma_{n+1} \nabla \Phi^a(\theta_n)}}{\|\theta_n e^{-\gamma_{n+1} \nabla \Phi^a(\theta_n)}\|_1}, \quad (19)$$

where this last equation has to be understood within a  $q$ -dimensional vector structure, coordinate per coordinate. We refer to [33] for further details.

### 3.2 The Pick-Freeze Mirror algorithm

An important feature of the definition of  $\Phi^a$  is that this function can be expressed as an expectation over a triplet of random variables  $(U, Y, Y^U)$ , which are assumed to be easily sampled in our context. To design the Pick-Freeze Mirror algorithm, we begin with the computation of  $\nabla \Phi^a$ , which is given in the following statement as an expectation.

**Proposition 3.2** (Computation of  $\nabla \Phi^a$ ). *Assume that  $U$  is a random subset sampled according to  $\mathcal{L}_\Upsilon^a$  and conditionally on  $U$ ,  $(Y, Y^U)$  is a Pick-Freeze sample, then for any  $s \in \Delta_q$ , one has:*

$$\nabla \Phi^a(s) = \nabla_s \mathbb{E}[\phi(s, Y, Y^U, U)] = \mathbb{E} \left[ \nabla_s \phi(s, Y, Y^U, U) \right]$$

where  $\phi(s, Y, Y^U, U)$  is the random variable defined by:

$$\nabla \phi(s, Y, Y^U, U) := (Y - \mathbb{E}[Y]) M_{U,:}^{-1} \left( (Y - \mathbb{E}[Y])[M^{-1}s]_U - (Y^U - \mathbb{E}[Y]) \right). \quad (20)$$

In our setting, an additional challenge arises from the fact that  $\mathbb{E}[Y]$ , which is involved in the computation of  $\nabla \Phi^a$  via Equation (20), is unknown and must be estimated online. Consequently, we introduce a biased Stochastic Mirror Descent algorithm, which jointly estimates both the mean of  $Y$  and all Sobol’ indices. This biased framework is similar to that used in [6].

**Description of the algorithm with unknown mean** We assume that we have at our disposal a sequence of i.i.d. random variables  $(X_n, X'_n, U_n)_{n \geq 1}$ , where  $(X_n, X'_n, U_n) \sim \mathbb{Q} \otimes \mathbb{Q} \otimes \mathcal{L}^a \Upsilon$ , and  $\mathcal{L}^a \Upsilon$  denotes the discrete distribution over  $\Upsilon$  such that

$$\forall n \geq 1 \quad \forall u \in \Upsilon \quad \mathbb{P}(U_n = u) = a_u.$$

We can therefore compute, for each  $n \geq 1$ , a Pick-Freeze sample  $(Y_n, Y_n^{(U)}) = (f(X_n), f(X_n^{U_n}))$ , as stated in Corollary 2.2. In particular, Proposition 3.2 shows that  $\nabla \phi(s, Y_n, Y_n^{(U)}, U_n)$  is an unbiased estimator of the gradient  $\nabla \Phi^a$  for any  $s \in \Delta_q$ . We estimate  $\mathbb{E}[Y]$  using the empirical mean of the samples  $(Y_n)$ . Let  $\hat{m}_n$  be defined by

$$\widehat{m}_n = \frac{1}{n} \sum_{k=1}^n Y_k.$$

Then, based on Proposition 3.2 and, specifically, Equation (20), a recursive estimator can be derived by substituting the unknown mean with its empirical counterpart  $\widehat{m}_n$ :

$$\widehat{\nabla}\phi_{n+1}(s) = (Y_{n+1} - \widehat{m}_n)M^{-1}U_{n+1}, : \left[ (Y_{n+1} - \widehat{m}_n)[M^{-1}s]_{U_{n+1}} - (Y_{n+1}^{(U)} - \widehat{m}_n) \right]. \quad (21)$$

We emphasize that in Equation (21) above, we use  $\widehat{m}_n$  and not  $\widehat{m}_{n+1}$  to compute  $\widehat{\nabla}\phi_{n+1}$ , for the sake of the simplicity from a mathematical point of view.

Finally we can define the recursive sequence  $(\widehat{\mathbf{S}}_n)_{n \geq 0}$  as:

$$\widehat{\mathbf{S}}_{n+1} = \arg \min_{s \in \Delta_q} \left\{ \langle \widehat{\nabla}\phi_{n+1}(\widehat{\mathbf{S}}_n), s - \widehat{\mathbf{S}}_n \rangle + \frac{1}{\eta_{n+1}} \mathcal{D}_h(s, \widehat{\mathbf{S}}_n) \right\} = \frac{\widehat{\mathbf{S}}_n e^{-\eta_{n+1} \widehat{\nabla}\phi_{n+1}(\widehat{\mathbf{S}}_n)}}{\|\widehat{\mathbf{S}}_n e^{-\eta_{n+1} \widehat{\nabla}\phi_{n+1}(\widehat{\mathbf{S}}_n)}\|_1}. \quad (22)$$

The Pick-Freeze Mirror algorithm then derives from a sequence of iterations that respectively computes  $\widehat{m}_n$ , samples  $(Y_n, Y_n^{(U)}, U_n)$  and uses (22). This new method is summarized in Algorithm 1.

### 3.3 Theoretical results

We state below our theoretical results.

**Almost sure convergence** Theorem 3.3 is a purely asymptotic convergence result. It establishes the convergence of our algorithm  $(\widehat{\mathbf{S}}_n)_{n \geq 1}$  to the set of Sobol' indices  $\mathbf{S}^*$  under weak assumptions on the sampling process (finite fourth moment of  $Y$ ) and on the step-size sequence. In particular, our result does not require any smoothness assumptions on the numerical code  $f$ , unlike recent contributions [11, 7, 45, 54, 30], which use a kernel-based approach that inherently imposes regularity conditions on the numerical code.

**Theorem 3.3** (Almost sure convergence of  $(\widehat{\mathbf{S}}_n)_{n \geq 0}$ ). *Assume that the sequence  $(\eta_n)_{n \geq 1}$  satisfies:*

$$\sum \eta_{n+1} = \infty, \quad \text{and} \quad \sum \eta_{n+1}^2 < \infty. \quad (23)$$

*Assume furthermore that  $Y$  has a 4-th order moment and that the discrete probability distribution  $a$  is lower bounded:*

$$\min_{u \in \Upsilon} a_u > 0,$$

*then  $(\widehat{\mathbf{S}}_n)_{n \geq 1}$  almost surely converges towards  $\mathbf{S}^*$ .*

These results hold for a standard setup in stochastic algorithms with a decreasing learning rate  $(\eta_n)_{n \geq 1}$ . The key condition is the convergence of the series  $\sum_{n \geq 1} \eta_n^2$  together with the divergence of  $\sum_{n \geq 1} \eta_n$ . A typical application of Theorem 3.3 corresponds to  $\eta_n = \eta_0 n^{-\alpha}$  with  $\alpha \in (1/2, 1]$ . Finally,

we note that the previous result is not quantitative, in the sense that the "rate" of convergence of the sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$  towards the target  $\mathbf{S}^*$  is not explicit. In particular, it would be highly valuable to complement our result with an almost sure convergence rate, see [28, 51, 36], which is a stronger statement. Another possible extension would be to establish a central limit theorem for the properly rescaled algorithm  $(\hat{\mathbf{S}}_n - \mathbf{S}^*)_{n \geq 1}$ , to obtain a precise characterization of the limiting variance. See, *e.g.*, [20, 23, 1] for examples of CLTs derived in various contexts using martingale representations or the Markov process perspective. Such extensions could be the topic of further research.

**Non-asymptotic upper bound** Using the perspective introduced in [39, 27] to assess the computational cost of convex stochastic optimization, it is possible to derive a more quantitative result on the sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$ . This result is expressed in terms of the expected value of  $\Phi^a$  throughout the algorithm. For this purpose, we introduce below a quantity related to the discrete probability distribution used for sampling the subsets  $U$  according to  $a$ :

$$|a|_{\text{exp}, \ell} := \mathbb{E} \left[ 2^{\ell|U|} \right] = \sum_{u \in \Upsilon} a_u 2^{\ell|u|}. \quad (24)$$

**Theorem 3.4** (Non-asymptotic risk bound). *Consider any finite horizon  $n > 0$  and the sequence  $(\bar{\mathbf{S}}_k^\eta)_{k \geq 1}$ , the Cesaro averaged sequence built from the sequence  $(\hat{\mathbf{S}}_k)_{k \geq 1}$ :*

$$\bar{\mathbf{S}}_n^\eta = \frac{\sum_{k=1}^n \eta_{k+1} \hat{\mathbf{S}}_k}{\sum_{k=1}^n \eta_{k+1}}.$$

*i) Assume that  $(\eta_{k+1})_{k \geq 0}$  is piecewise constant defined by:*

$$\eta_{k+1} = \frac{\mathbf{1}_{k \leq n-1}}{\sqrt{(1 + \sqrt{\|a\|_{\text{exp}, 2}})n}},$$

*then*

$$\mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a \leq \sqrt{1 + \sqrt{\|a\|_{\text{exp}, 2}}} \left( \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1) + C\mathbb{E}[Y^4] \right) \frac{1}{\sqrt{n}} + 5\|a\|_{\text{exp}} \text{Var}(Y) \frac{\log(n)}{n}.$$

*ii) If  $(\eta_{k+1})_{k \geq 0}$  is horizon free (that does not depend on the stopping iteration  $n$ ) defined by:*

$$\eta_{k+1} = \eta_0(k+1)^{-1/2},$$

*then  $\forall n \in \mathbb{N}$*

$$\mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a \leq C \left( \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1) + \|a\|_{\text{exp}, 2} + \mathbb{E}[Y^4] \right) \frac{\log n}{\sqrt{n}},$$

We emphasize that item *i)* of Theorem 3.4 does not constitute a true convergence result, which is indeed impossible to obtain for a constant step-size stochastic algorithm. Nevertheless, it can

be regarded as a benchmark result, following common practice in convex optimization for machine learning. It provides a convenient way to assess the mean-square behavior of a stochastic optimization algorithm in a convex landscape. Even if the bound does not appear to explicitly involve the ambient dimension of the problem, it is implicitly present both in the term  $\|a\|_{\text{exp},2}$  and in the fourth-order moment of the random variable  $Y$ . This result captures both the effect of the final horizon on the "convergence" rate and the dependence of the final bound on the dimension of the ambient space when the algorithm is averaged over all iterations. The appearance of both the "dimension" and the factor  $\sqrt{n}$  is not surprising, as it corresponds to the minimax rate of convergence in many stochastic optimization problems with convex landscapes (see, e.g., [40]).

Point *ii*) of the previous theorem concerns the case of a decreasing step sequence. The result holds for any finite simulation horizon but leads to a slightly weaker upper bound with an additional logarithmic factor in the bound of point *ii*).

In both cases, the quantity  $\|a\|_{\text{exp},\ell}$  captures the dimensional effect, *e.g.*, the number of unknown Sobol' coefficients to be estimated. In particular, it satisfies a universal upper bound, independently of the weight distribution  $a$ :

$$\|a\|_{\text{exp},\ell} \leq 2^{\ell p}.$$

Furthermore, for a fixed number of coefficients, the quantity  $\|a\|_{\text{exp},2}$  decreases as the distribution  $a$  increasingly favors subsets of smaller size. Nevertheless, one should not conclude that it is advisable to choose a distribution  $a$  such that  $\|a\|_{\text{exp},2}$  is very small, since the result concerns the value of  $\Phi^a$ , not of  $\bar{\mathbf{S}}^{\eta n}$  itself. Indeed, while it is possible to relate the proximity of  $\Phi^a(\bar{\mathbf{S}}^{\eta n})$  to its minimum with the proximity of  $\bar{\mathbf{S}}^{\eta n}$  to  $\mathbf{S}^*$ , this relation critically depends on the strong convexity constant of  $\Phi^a$  (associated with  $\nabla^2\Phi^a$ ), which is significantly degraded if  $a$  is chosen to produce a very small  $\|a\|_{\text{exp},2}$ .

Hence, the influence of  $a$  through  $\|a\|_{\text{exp},2}$  and the spectral properties of  $\nabla^2\Phi^a$  warrants a deeper mathematical investigation, which lies beyond the scope of the present paper and will be addressed in future work.

## 4 Experimental results

In this section, we explore the numerical performances of our algorithm.

### 4.1 Baselines and Competing Methods

Our study focuses on comparing our method with two existing estimation techniques for Sobol' indices: the standard Pick-Freeze estimator for individual Sobol' indices [32], and the kernel-based estimator of [10]. We consider both regular and irregular functions  $f$ . Let us first briefly discuss the advantages and limitations of these two methods.

- The asymptotically efficient version of the Pick-Freeze estimator for closed Sobol' indices studied in [32] (referred to as "PF1" below). For each subset  $u$ , it computes a Monte Carlo estimator

---

**Algorithm 1** Pick-Freeze Mirror algorithm
 

---

**Require:**  $N \geq 0$ , Step-size Sequence  $(\gamma_n)_{n \geq 1}$

$\hat{m}_0 \leftarrow 0$

$\hat{\mathbf{S}}_0^\theta \leftarrow 0$  and  $\hat{\mathbf{S}}_0^* \leftarrow \frac{\mathbf{1}_{q-1}}{q-1}$

▷ Initialization with a uniform distribution of weights over  $\Delta_q$

**for**  $n \leftarrow 1$  to  $N$  **do**

Sample  $(X_{n+1}, X'_{n+1}, U_{n+1}) \sim \mathbb{Q} \otimes \mathbb{Q} \otimes \mathcal{L}_Y^a$  and compute  $Y_{n+1}, Y_{n+1}^{(U)}$

Update the **estimator** of the expectation of  $Y$ :

$$\hat{m}_{n+1} = \hat{m}_n + \frac{1}{n+1}(Y_{n+1} - \hat{m}_n)$$

▷ Empirical mean estimation

Compute the estimator of  $\nabla \Phi^a$  at step  $n$ :

$$\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) = (Y_{n+1} - \hat{m}_n) M_{U_{n+1}}^{-1} \cdot \left[ (Y_{n+1} - \hat{m}_n) [M^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - (Y_{n+1}^{(U)} - \hat{m}_n) \right]$$

▷ Gradient estimation

Update the estimator of the collection of Sobol' indices:

$$\hat{\mathbf{S}}_{n+1} = \frac{\hat{\mathbf{S}}_n e^{-\eta_{n+1}} \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)}{\|\hat{\mathbf{S}}_n e^{-\eta_{n+1}} \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_1}$$

▷ Stochastic Mirror Descent

**end for**

---

$T_{u,N}$  of  $\Sigma_u$  based on a Pick-Freeze sample  $(Y_i, Y_i^{(u)})_{1 \leq i \leq N}$  :

$$T_{u,N} = \frac{\frac{1}{N} \sum_{i=1}^N Y_i Y_i^{(u)} - \left( \frac{1}{N} \sum \left[ \frac{Y_i + Y_i^{(u)}}{2} \right] \right)^2}{\frac{1}{N} \sum \left[ \frac{Y_i + Y_i^{(u)}}{2} \right]^2 - \left( \frac{1}{N} \sum \left[ \frac{Y_i + Y_i^{(u)}}{2} \right] \right)^2}$$

- Advantages: The Pick-Freeze estimators are consistent and satisfy a (joint) central limit theorem under the sole assumption that the output has a finite fourth-order moment.
- Drawbacks: To estimate all  $2^{p-1}$  Sobol' indices at a  $\sqrt{n}$  rate, a sample of size  $2^p n$  is required.
- The kernel-based estimator given in equation (20) of [10], using an Epanechnikov kernel of order 2 and 4 (see [19] for a definition), with the kernel bandwidth optimized via leave-one-out on the regression function ("Kernel 2" and "Kernel 4").
  - Advantages: A simple i.i.d. sample of size  $n$  is theoretically sufficient to estimate all Sobol' indices at the  $\sqrt{n}$  rate.

- Drawbacks: The input variables are assumed to be compactly supported and absolutely continuous with respect to the Lebesgue measure. Moreover, regularity assumptions on the regression function are also required. Although asymptotic normality is established for all Sobol’ indices, the computational cost of evaluating the estimator appears to grow exponentially with the order of the indices.

For all these techniques, 50 replicates of the estimators are computed up to step  $N_0 = 500$ . We compare these methods with our Pick-Freeze Mirror Descent algorithm (Algorithm 1) using a decreasing step sequence  $\eta_n = \eta_0/\sqrt{n+1}$ , where  $\eta_0$  is a tuning parameter. To fairly compare the results and account for the fact that the Pick-Freeze Mirror algorithm computes all indices simultaneously, we consider the algorithm at times  $N_0$  and  $2^p N_0$ , corresponding respectively to "Mirror\_PF-1" and "Mirror\_PF-2".

## 4.2 Numerical code used and comparison

### 4.2.1 Case of a smooth regular function: the Bratley function

We first consider the Bratley function, defined as

$$f_{\text{Bratley}}(X^1, \dots, X^p) = \sum_{i=1}^p (-1)^i \prod_{j=1}^i X^j,$$

where  $X^i \sim \mathcal{U}([0, 1])$  are i.i.d. and  $p = 5$ . For this function, [10] considered the performance of their algorithm for the estimation of both the first-order and total-order indices. Recall that the first-order indices are the Sobol’ indices associated with singletons, while the total-order indices are defined as

$$S_i^{\text{tot}} = 1 - \Sigma_{\{1, \dots, p\} \setminus \{i\}}^*.$$

In particular, since the total-order indices depend on the closed Sobol’ indices, we estimate them using our sequence  $(\hat{\mathbf{S}}_n)_{n \geq 1}$  and (8).

### 4.2.2 Case of a discontinuous function

We then consider a discontinuous function defined by

$$f_{\text{disc}}(X^{(1)}, X^{(2)}, X^{(3)}) = X^{(1)} \mathbb{1}_{X^{(3)} < 0} + (X^{(2)})^2 \mathbb{1}_{X^{(3)} \geq 0} + X^{(3)},$$

for  $X^i$  i.i.d  $\mathcal{N}(0, 1)$  random variables. This case is interesting because it does not satisfy the regularity assumptions required for the kernel based methods, nor the assumption on compactly supported random variables.

### 4.2.3 Comparison

We first observe that, as expected, the "Mirror\_PF-1" method performs worse than "Mirror\_PF-2". This is entirely reasonable, since the comparison essentially reflects the behavior of our algorithm

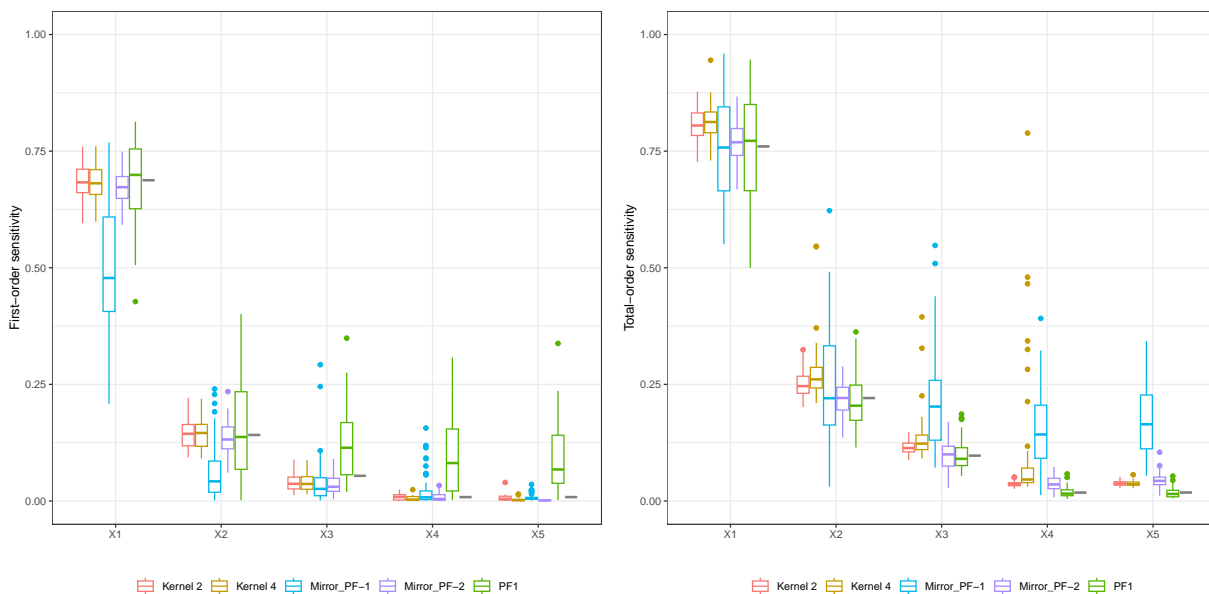


Figure 1: Estimators of the first-order indices (left panel) and total-order indices (right panel) of the Bratley function obtained with different methods. The reference value is shown as a gray line.

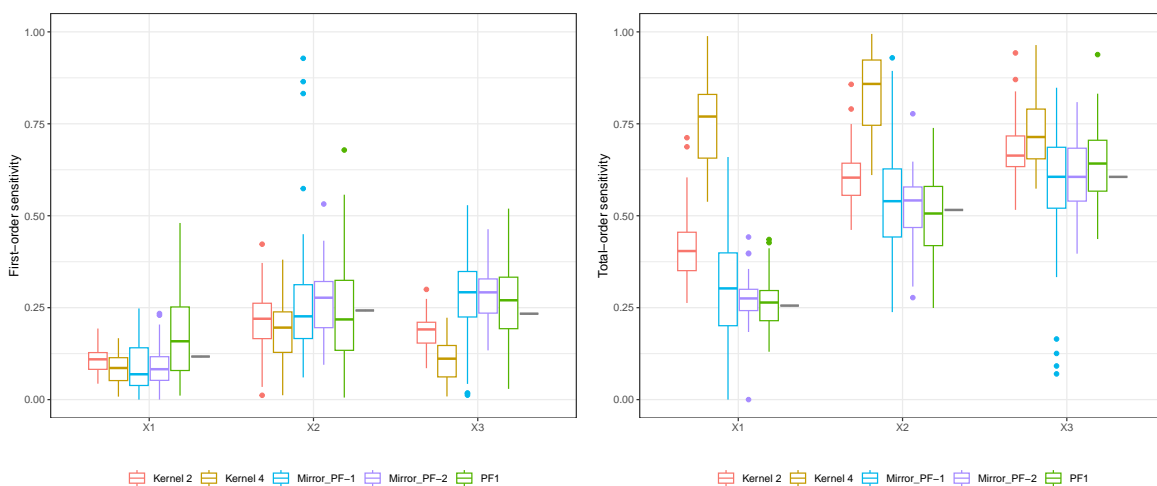


Figure 2: Estimators of the first-order indices (left panel) and the total order indices (right panel) for the function  $f_{disc}$  for the different methods. The reference value is represented with a gray line.

after 1 and  $N_0$  epochs, respectively (where one epoch corresponds to the full set of  $2^p$  variables in our problem). Nevertheless, in many cases, "Mirror\_PF-1", despite using an extremely small number

of iterations, still achieves reasonably good performance given its very low computational cost for estimating *all* Sobol' indices.

Next, it should be noted that the "Mirror\_PF-2" method consistently outperforms the baseline "PF1" method, which consists in independently repeating the Pick-Freeze procedure for each variable under consideration: in particular, the variance of "Mirror\_PF-2" are smaller than those of the Pick-Freeze method when considering the mirror algorithm at step  $2^p N_0$ . This clearly demonstrates the benefit of formulating our problem as a constrained optimization task, which allows all indices to be updated simultaneously. Moreover, the "Mirror\_PF-1" method with a single epoch sometimes achieves results comparable to the baseline "PF1" (this is in particular the case for the null indices of  $X^{(3)}$ ,  $X^{(4)}$  and  $X^{(5)}$  for  $f_{\text{Bratley}}$ ), thus efficiently exploiting the simplex constraint in the search for the Sobol' indices.

Finally, the comparison of our methods with the kernel-based methods Kernel2 and Kernel4 requires more nuance. Indeed, for the regular Bratley function, the two kernel methods produce results that are quite close to each other, while, for an equivalent number of iterations in the algorithm, they perform similarly to the "Mirror\_PF-2" method and clearly outperform both "Mirror\_PF-1" and PF1. However, this observation for the class of regular functions completely disappears in the case of non-regular functions. In this case, we observe that the kernel estimators fail to estimate the correct Sobol' indices, whereas the Mirror Pick-Freeze and the Pick-Freeze methods still provide good results. The rather excellent results of kernel methods for smooth functions and the very disappointing performance for discontinuous functions are entirely consistent with the behavior of the kernel-based methods, which rely on a function decomposition that is not valid when the functions are non-smooth. Finally, let us stress that a major drawback of kernel methods is the cost in memory required to run the estimation. On this aspect, our on-line algorithm performs at low memory cost even for a large number of input variables.

### 4.3 Investigating different sampling strategies

Finally, in this paragraph, we briefly describe the influence of the sampling strategy resulting from the choice of the distribution  $a$ . The distribution  $a$  impacts the way our algorithm will chose the different subsets all along the iterations. We chose to compare uniform sampling ("unif") with two adaptive strategies. In the latter, the distribution  $a$  is updated at each step of the algorithm based on the current value of the sequence  $\hat{\mathbf{S}}_n$ . Specifically, we set  $a \propto \hat{\mathbf{S}}_n$  in strategy "S" and  $a \propto 1/\hat{\mathbf{S}}_n$  in strategy "1/S". We also computed an averaged version of the algorithm in the uniform case, referred to as "avg".

We performed 100 repetitions of the algorithm for different time horizons

$$n \in (2^p) \times \{500, 1000, 1500, 2500, 3500, 5000\}.$$

We consider the discrete function  $f_{disc,2}$  defined by

$$f_{disc,2}(X^1, X^2, X^3) = X^1 \mathbf{1}_{\{X^3 > 0\}} + X^2 \mathbf{1}_{\{X^3 < 0\}} + X^3,$$

and a step sequence  $\eta_n = 0.3/\sqrt{n+1}$ .

The advantage of this choice lies in the fact that the true values of the Sobol' indices can be computed through an exact calculation. For each strategy, we compute the mean squared error of the estimated closed Sobol' indices, obtained as  $M^{-1}\hat{\mathbf{S}}_n$ . For comparison, we also computed the  $2^p$  Pick-Freeze estimators based on  $\{500, 1000, 1500, 2500, 3500, 5000\}$  observations. The mean squared errors obtained in this case correspond to the line labeled "PF" in Figure 3.

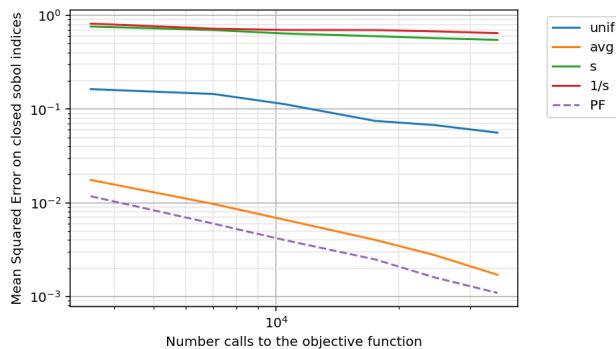


Figure 3: MSE on the  $2^5$  closed Sobol' indices computed over 100 repetitions of the Pick-freeze mirror algorithm for different choices of the sampling distribution  $a$ .

Interestingly, we observe that the uniform strategy performs better than the adaptive strategies in this case. Furthermore, when considering an averaged version of the algorithm, we recover a convergence rate of  $1/n$ , similar to that of the Pick-Freeze method. Finally, this last method, using a uniform weight for  $a$  combined with Cesàro averaging, appears to achieve the best results among all our strategies.

These simulations illustrate that the choice of  $a$  has a direct impact on the quality of our numerical results. Similarly, the choice of  $a$  also likely affects the performance of the averaged algorithm. The theoretical study of this influence is not addressed in the present work, as it requires nontrivial developments that go well beyond this initial study. We therefore leave this analysis for future work, in particular by studying the asymptotic behavior of our methods as a function of  $a$ , through a central limit theorem with a variance that likely depends on this distribution  $a$ .

## 5 Proof of the theoretical results

This technical section is devoted to the proof of Theorem 1.2 stated in the early introduction of the paper. The result contains two parts that are then detailed in Theorem 3.3 and in Theorem 3.4.

## 5.1 Gradient of $\Phi^a$

*Proof of Proposition 3.2.* For  $s \in \Delta_q$  and  $h \in \mathbb{R}^q$  we verify that:

$$\begin{aligned}
\Phi^a(s+h) - \Phi^a(s) &= \sum_{u \in \Upsilon} a_u \mathbb{E}[(Y - \mathbb{E}[Y])[M^{-1}h]_u \left( (Y - \mathbb{E}[Y])[M^{-1}s]_u - (Y^u - \mathbb{E}[Y]) \right)] + o(|h|^2) \\
&= \sum_{u \in \Upsilon} a_u \mathbb{E} \left[ (Y - \mathbb{E}[Y]) \sum_i M_{u,i}^{-1} h_i \left( (Y - \mathbb{E}[Y])[M^{-1}s]_u - (Y^u - \mathbb{E}[Y]) \right) \right] + o(|h|^2) \\
&= \sum_{u \in \Upsilon} a_u \left\langle \mathbb{E} \left[ (Y - \mathbb{E}[Y]) M_{u,:}^{-1} \left( (Y - \mathbb{E}[Y])[M^{-1}s]_u - (Y^u - \mathbb{E}[Y]) \right) \right], h \right\rangle + o(|h|^2) \\
&= \langle \nabla \Phi^a(s), h \rangle + o(|h|^2)
\end{aligned}$$

where  $M_{u,:}^{-1}$  refers to line  $u$  of the matrix  $M^{-1}$ . Let us finally remark that if  $U$  is a discrete  $\Upsilon$ -valued random variable distributed according to the probability distribution  $a$ , then we can rewrite  $\nabla \Phi^a$  as:

$$\begin{aligned}
\nabla \Phi^a(s) &= \sum_u a_u \mathbb{E} \left[ (Y - \mathbb{E}[Y]) M_{u,:}^{-1} \left( (Y - \mathbb{E}[Y])[M^{-1}s]_u - (Y^u - \mathbb{E}[Y]) \right) \right] \\
&= \mathbb{E} \left[ (Y - \mathbb{E}[Y]) M_{U,:}^{-1} \left( (Y - \mathbb{E}[Y])[M^{-1}s]_U - (Y^U - \mathbb{E}[Y]) \right) \right].
\end{aligned}$$

□

## 5.2 Proof of Theorem 3.3 (almost sure convergence)

The proof of almost sure convergence follows the general approach of [33], with specific attention given to the presence of an estimation bias in the algorithm due to the online approximation of  $\mathbb{E}[Y]$ . This bias is handled in a manner quite similar to that of [6], although the control of this bias term is, of course, specific to the context considered in the present work. Finally, although the approach based on relative smoothness introduced in [2] offers a powerful framework for analyzing mirror descent methods (even in the stochastic cases, see [19, 18]), it cannot be employed in our setting, as the conditions required to apply the results of [19, 18] are not satisfied. Indeed, while the function  $\Phi^a$  exhibits uniform Hessian bounds over the entire simplex (see in particular Appendix B), the entropy function  $h$  given in Equation (16) defining our Bregman divergence is not differentiable on the boundary of the simplex so that the metric induced by  $\mathcal{D}_h$  cannot be upper-bounded by the one induced by the Hessian of  $\Phi^a$ .

*Proof of Theorem 3.3.*

The proof is organized into three parts. The first part leverages the variational formulation of mirror descent to establish a one-step inequality for the Bregman divergences. The second part provides a precise, quantitative estimate of the bias between the estimated drift at iteration  $n$  and the theoretical drift, which arises from replacing  $\mathbb{E}[Y]$  with the empirical mean  $\hat{m}_n$  in the gradient. In the final part, the convergence of the sequence is established using the Robbins–Siegmund lemma, together with several technical estimates.

**Step 1: Obtaining a recursion on  $\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n)$**  Recall the definition of  $\hat{\mathbf{S}}_{n+1}$

$$\hat{\mathbf{S}}_{n+1} = \arg \min_{s \in \Delta_q} \left\{ \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), s - \hat{\mathbf{S}}_n \rangle + \frac{1}{\eta_{n+1}} \mathcal{D}_h(s, \hat{\mathbf{S}}_n) \right\},$$

then the first order condition reads

$$\forall s \in \Delta_q : \quad \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), s - \hat{\mathbf{S}}_{n+1} \rangle + \langle \nabla_1 \mathcal{D}_h(\hat{\mathbf{S}}_{n+1}, \hat{\mathbf{S}}_n), s - \hat{\mathbf{S}}_{n+1} \rangle \geq 0,$$

where  $\nabla_1 \mathcal{D}_h$  refers to the gradient with respect to the first variable. As  $\nabla_1 \mathcal{D}_h(x, y) = \nabla h(x) - \nabla h(y)$  (Lemma A.2) we deduce that:

$$\begin{aligned} \forall s \in \Delta_q : \quad \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_{n+1} - s \rangle &\leq \langle \nabla h(\hat{\mathbf{S}}_{n+1}) - \nabla h(\hat{\mathbf{S}}_n), s - \hat{\mathbf{S}}_{n+1} \rangle \\ &\leq \mathcal{D}_h(s, \hat{\mathbf{S}}_n) - \mathcal{D}_h(s, \hat{\mathbf{S}}_{n+1}) - \mathcal{D}_h(\hat{\mathbf{S}}_{n+1}, \hat{\mathbf{S}}_n), \end{aligned}$$

where the second inequality derives from the three point lemma A.1. Therefore, using simple algebra and Inequality (17), we obtain that:

$$\mathcal{D}_h(s, \hat{\mathbf{S}}_{n+1}) \leq \mathcal{D}_h(s, \hat{\mathbf{S}}_n) - \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_{n+1} - s \rangle - \frac{1}{2} \|\hat{\mathbf{S}}_{n+1} - \hat{\mathbf{S}}_n\|_2^2.$$

We can now substitute  $\hat{\mathbf{S}}_{n+1} - s$  by  $\hat{\mathbf{S}}_n - s$  in the middle term and write that:

$$\eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_{n+1} - s \rangle = \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - s \rangle + \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_{n+1} - \hat{\mathbf{S}}_n \rangle.$$

The Young inequality induces:

$$\eta_{n+1} |\langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_{n+1} - \hat{\mathbf{S}}_n \rangle| \leq \frac{1}{2} \eta_{n+1}^2 \|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 + \frac{1}{2} \|\hat{\mathbf{S}}_{n+1} - \hat{\mathbf{S}}_n\|_2^2$$

This finally leads to the key inequality:

$$\mathcal{D}_h(s, \hat{\mathbf{S}}_{n+1}) \leq \mathcal{D}_h(s, \hat{\mathbf{S}}_n) - \eta_{n+1} \langle \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - s \rangle + \frac{1}{2} \eta_{n+1}^2 \|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2. \quad (25)$$

**Step 2: Comparison between  $\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)$  and  $\nabla \Phi^a(\hat{\mathbf{S}}_n)$ .** Recall that from Proposition 3.2

$$\Phi^a(\hat{\mathbf{S}}_n) = \mathbb{E} \left[ \nabla H(\hat{\mathbf{S}}_n, Y_{n+1}, Y_{n+1}^{(U)}, U_{n+1}) | \mathcal{F}_n \right],$$

we can then write

$$\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) = \nabla \Phi^a(\hat{\mathbf{S}}_n) + \Delta M_{n+1} + R_{n+1}, \quad (26)$$

where  $\Delta M_{n+1}$  is a  $(\mathcal{F}_n)$ -martingale increment defined by:

$$\Delta M_{n+1} = \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) - \mathbb{E} \left[ \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) | \mathcal{F}_n \right], \quad (27)$$

and  $R_{n+1}$  is a rest/bias term that comes from the estimation of  $\mathbb{E}[Y]$  with  $\hat{m}_n$  in Equation (21):

$$R_{n+1} = \mathbb{E} \left[ \widehat{\nabla\phi}_{n+1}(\hat{\mathbf{S}}_n) | \mathcal{F}_n \right] - \nabla\Phi^a(\hat{\mathbf{S}}_n). \quad (28)$$

Using Equation (26) in (25), we deduce that:

$$\begin{aligned} \mathcal{D}_h(s, \hat{\mathbf{S}}_{n+1}) &\leq \mathcal{D}_h(s, \hat{\mathbf{S}}_n) - \eta_{n+1} \langle \nabla\Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - s \rangle - \eta_{n+1} \langle \Delta M_{n+1}, \hat{\mathbf{S}}_n - s \rangle \\ &\quad - \eta_{n+1} \langle R_{n+1}, \hat{\mathbf{S}}_n - s \rangle + \frac{1}{2} \eta_{n+1}^2 \|\widehat{\nabla\phi}_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 \end{aligned}$$

Let us choose  $s = \mathbf{S}^*$  in the previous inequality to obtain

$$\begin{aligned} \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) &\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n) - \eta_{n+1} \langle \nabla\Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle - \eta_{n+1} \langle \Delta M_{n+1}, \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle \\ &\quad - \eta_{n+1} \langle R_{n+1}, \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle + \frac{1}{2} \eta_{n+1}^2 \|\widehat{\nabla\phi}_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 \end{aligned} \quad (29)$$

Now from the Cauchy-Schwarz inequality and then the Young inequality, we have

$$\begin{aligned} -\eta_{n+1} \langle R_{n+1}, \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle &\leq \eta_{n+1} \|R_{n+1}\|_2 \|\hat{\mathbf{S}}_n - \mathbf{S}^*\|_2 \\ &\leq \eta_{n+1} \|R_{n+1}\|_2 \frac{1 + \|\hat{\mathbf{S}}_n - \mathbf{S}^*\|_2^2}{2}. \end{aligned} \quad (30)$$

We now use Inequality (30) in (29) together with the strong convexity of  $h$  in inequality (17) to obtain:

$$\begin{aligned} \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) &\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n) \left(1 + \frac{1}{2} \eta_{n+1} \|R_{n+1}\|_2\right) - \eta_{n+1} \langle \nabla\Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle - \eta_{n+1} \langle \Delta M_{n+1}, \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle \\ &\quad + \frac{1}{2} \eta_{n+1} \|R_{n+1}\|_2 + \frac{1}{2} \eta_{n+1}^2 \|\widehat{\nabla\phi}_{n+1}(\hat{\mathbf{S}}_n)\|_2^2. \end{aligned}$$

We observe that  $R_{n+1}$  is  $\mathcal{F}_n$ -measurable and computing the conditional expectation with respect to  $\mathcal{F}_n$  leads to:

$$\begin{aligned} \mathbb{E}[\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) | \mathcal{F}_n] &\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n) \left(1 + \frac{1}{2} \eta_{n+1} \|R_{n+1}\|_2\right) - \eta_{n+1} \langle \nabla\Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle \\ &\quad + \frac{1}{2} \eta_{n+1} \|R_{n+1}\|_2 + \frac{1}{2} \eta_{n+1}^2 \mathbb{E}[\|\widehat{\nabla\phi}_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 | \mathcal{F}_n] \end{aligned} \quad (31)$$

**Step 3: Conclusion of the proof.** Using Lemma 5.1 and our set of assumptions on the step-size sequence (23), we deduce that:

$$\sum \eta_{n+1} \mathbb{E}[\|R_n\|_2] < \infty.$$

Notice that since all the terms are non-negative,

$$\mathbb{E} \left[ \sum \eta_{n+1} \|R_n\|_2 \right] \leq \sum \eta_{n+1} \mathbb{E}[\|R_n\|_2],$$

and it shows that  $\sum \eta_{n+1} \|R_n\|_2$  is almost surely finite and  $\prod_{n \geq 0} (1 + \eta_{n+1} \|R_n\|_2) < \infty$  almost surely as well. In the meantime, Lemma 5.2 associated with conditions (23) also shows that:

$$\sum_n \mathbb{E}[\eta_{n+1}^2 \|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2] < \infty$$

We can now apply the Robbins-Siegmund Lemma (stated in Lemma A.4) and deduce that a random variable  $D_\infty$  exists such that:

$$\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) \longrightarrow D_\infty \in L^1 \quad \text{and} \quad \sum \eta_{n+1} \langle \nabla \Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle < \infty, \quad \text{a.s.} \quad (32)$$

Using the convexity of  $\Phi^a$  we obtain

$$\sum \eta_{n+1} \left( \Phi^a(\hat{\mathbf{S}}_n) \right) \leq \sum \eta_{n+1} \langle \nabla \Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle < +\infty, \quad \text{a.s.} \quad (33)$$

The rest of the proof proceeds now in a standard way in stochastic approximation theory. From Equation (33) and  $\sum \eta_{n+1} = +\infty$ , we know that a subsequence  $(\hat{\mathbf{S}}_{n_k})_{k \geq 1}$  exists such that:

$$\lim_{k \rightarrow +\infty} \langle \nabla \Phi^a(\hat{\mathbf{S}}_{n_k}), \hat{\mathbf{S}}_{n_k} - \mathbf{S}^* \rangle = 0, \quad \text{a.s.}$$

The strong convexity of  $\Phi^a$  may be translated into the following inequality:

$$\forall s \in \Delta_q \quad \langle \nabla \Phi^a(s), s - \mathbf{S}^* \rangle \geq \rho_a \|s - \mathbf{S}^*\|^2,$$

where  $\rho_a > 0$  refers to the lowest eigenvalue of  $\nabla^2 \Phi^a$  over  $\Delta_q$  (see Proposition B.2 in the appendix). It then implies that

$$\lim_{k \rightarrow +\infty} \hat{\mathbf{S}}_{n_k} = \mathbf{S}^*, \quad \text{a.s.}$$

In the meantime, we also deduce from Equation (32) that  $\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n_k}) \longrightarrow D_\infty$  a.s. and the continuity of the Bregman divergence  $\mathcal{D}_h$  yields  $D_\infty = 0$  a.s. Finally, the lower bound (17) implies that the entire sequence  $\hat{\mathbf{S}}_n$  converges towards  $\mathbf{S}^*$ . □

### 5.3 Proof of Theorem 3.4 ( non-asymptotic upper bound)

The following proof follows an approach recently proposed in [6] to obtain a non-asymptotic bound for biased stochastic mirror methods. Nevertheless, care must be taken to meticulously track the sequence of inequalities in order to preserve the dimension-dependent constants that have a significant influence on the final bound. Moreover, although the proof strategy is similar to that of [6], it still requires bias controls that are specific to the model considered here, as well as precise inequalities on the spectra of the change-of-variable matrices  $M$  used.

*Proof of Theorem 3.4.* The starting point of the non-asymptotic bound is Equation (31) that can be written as

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) | \mathcal{F}_n] &\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) - \eta_{n+1} \langle \nabla \Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle \\
&\quad + \frac{\eta_{n+1} \|R_{n+1}\|_2}{2} (1 + \|\hat{\mathbf{S}}_n - \mathbf{S}^*\|_2^2) + \frac{1}{2} \eta_{n+1}^2 \mathbb{E}[\|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 | \mathcal{F}_n] \\
&\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1}) - \eta_{n+1} (\Phi^a(\hat{\mathbf{S}}_n) - \Phi^a(\mathbf{S}^*)) \\
&\quad + \frac{5}{2} \eta_{n+1} \|R_{n+1}\|_2 + \frac{1}{2} \eta_{n+1}^2 \mathbb{E}[\|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2 | \mathcal{F}_n],
\end{aligned}$$

where we used in the last line the convexity of  $\Phi^a$  that yields  $-\langle \nabla \Phi^a(\hat{\mathbf{S}}_n), \hat{\mathbf{S}}_n - \mathbf{S}^* \rangle \leq -(\Phi^a(\hat{\mathbf{S}}_n) - \Phi^a(\mathbf{S}^*))$  and the compactness of  $\Delta_q$  that yields  $\|\hat{\mathbf{S}}_n - \mathbf{S}^*\|_2^2 \leq \|\hat{\mathbf{S}}_n - \mathbf{S}^*\|_1^2 \leq 4$ . We now compute the overall expectation, use Lemma 5.1 and Lemma 5.2, and obtain that:

$$\begin{aligned}
\eta_{n+1} \mathbb{E}[\Phi^a(\hat{\mathbf{S}}_n) - \Phi^a(\mathbf{S}^*)] &\leq \mathbb{E}[\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n) - \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1})] \\
&\quad + 5\eta_{n+1} \mathbb{E}[\|R_{n+1}\|_2] + \frac{\eta_{n+1}^2}{2} \mathbb{E}[\|\widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n)\|_2^2] \\
&\leq \mathbb{E}[\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_n) - \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_{n+1})] \\
&\quad + 5\|a\|_{\text{exp}} \text{Var}(Y) \frac{\eta_{n+1}}{n} + C \left(1 + \sqrt{\|a\|_{\text{exp},2}}\right) \mathbb{E}[Y^4] \frac{\eta_{n+1}^2}{2}.
\end{aligned}$$

The rest of the proof then proceeds following a standard argument with Cesaro mean and telescopic sums:

$$\begin{aligned}
\sum_{k=1}^n \eta_{k+1} (\mathbb{E}[\Phi^a(\hat{\mathbf{S}}_k)] - \Phi^a(\mathbf{S}^*)) &\leq \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1) + 5\|a\|_{\text{exp}} \text{Var}(Y) \sum_{k=1}^n \frac{\eta_{k+1}}{k} \\
&\quad + C \left(1 + \sqrt{\|a\|_{\text{exp},2}}\right) \mathbb{E}[Y^4] \sum_{k=1}^n \frac{\eta_{k+1}^2}{2}
\end{aligned}$$

Defining now the weighted Cesaro average as:

$$\bar{\mathbf{S}}_n^\eta = \frac{\sum_{k=1}^n \eta_{k+1} \hat{\mathbf{S}}_k}{\sum_{k=1}^n \eta_{k+1}},$$

then, the convexity of  $\Phi^a$  yields:

$$\begin{aligned} \mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a &\leq \frac{\sum_{k=1}^n \eta_{k+1} (\mathbb{E}[\Phi^a(\hat{\mathbf{S}}_k)] - \Phi^a(\mathbf{S}^*))}{\sum_{k=1}^n \eta_{k+1}} \\ &\leq \frac{\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1)}{\sum_{k=1}^n \eta_{k+1}} + 5\|a\|_{\text{exp}} \text{Var}(Y) \frac{\sum_{k=1}^n \eta_{k+1} k^{-1}}{\sum_{k=1}^n \eta_{k+1}} + C \left(1 + \sqrt{\|a\|_{\text{exp},2}}\right) \mathbb{E}[Y^4] \frac{\sum_{k=1}^n \eta_{k+1}^2}{2 \sum_{k=1}^n \eta_{k+1}} \end{aligned} \quad (34)$$

It remains to optimize both terms of the right hand side of Inequality (34) by choosing an appropriate step-size sequence  $(\eta_{k+1})_{k \geq 0}$ . There exists two classical choices:

- Either  $(\eta_{k+1})_{k \geq 0}$  is piecewise constant (depending on the final number of performed simulations):

$$\eta_{k+1} = \begin{cases} \eta^* & \forall k \leq n-1 \\ 0 & \forall k > n-1. \end{cases}$$

, In that case, Inequality (34) reads:

$$\mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a \leq \frac{\mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1)}{n\eta^*} + 5\|a\|_{\text{exp}} \text{Var}(Y) \frac{\log(n)}{n} + C \left(1 + \sqrt{\|a\|_{\text{exp},2}}\right) \mathbb{E}[Y^4] \eta^*$$

We finally optimize the choice of  $\eta^*$ , which leads to:

$$\eta^* = \sqrt{\frac{1}{(1 + \sqrt{\|a\|_{\text{exp},2}})n}}.$$

Our final upper bound is then:

$$\mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a \leq \sqrt{1 + \sqrt{\|a\|_{\text{exp},2}}} \left( \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1) + C\mathbb{E}[Y^4] \right) n^{-1/2} + 5\|a\|_{\text{exp}} \text{Var}(Y) \frac{\log(n)}{n}.$$

- Or  $(\eta_{k+1})_{k \geq 0}$  is a decreasing step-size sequence of the form  $\eta_{k+1} = \eta_0(k+1)^{-\alpha}$  with  $\alpha \in [1/2, 1)$ . In that case the right hand side of Inequality (34) decays as  $(\sum_{k=1}^n \eta_{k+1})^{-1} \leq \eta_0(n+1)^{-\alpha+1}$ . The situation becomes somewhat more technical and tedious to describe precisely, and only the dependence on the number of iterations remains straightforward to characterize. In particular, one can verify that choosing  $\alpha = 1/2$ , and  $\eta_0 \asymp 1$  leads to:

$$\mathbb{E}[\Phi^a(\bar{\mathbf{S}}_n^\eta)] - \min \Phi^a \leq C \left( \mathcal{D}_h(\mathbf{S}^*, \hat{\mathbf{S}}_1) + \|a\|_{\text{exp},2} + \mathbb{E}[Y^4] \right) \frac{\log n}{\sqrt{n}},$$

which constitutes a slightly weaker bound than the inequality obtained in the previous case.  $\square$

## 5.4 Controls on the bias terms

**Lemma 5.1. Control of  $\mathbb{E}[\|R_{n+1}\|_2]$**  For any  $n \geq 1$ , one has:

$$\mathbb{E}[\|R_{n+1}\|_2] \leq \frac{\|a\|_{\text{exp},1/2} \text{Var}[Y]}{n}.$$

*Proof of Lemma 5.1.* We recall that the bias term involved in  $R_{n+1}$  is given by:

$$R_{n+1} = \mathbb{E} \left[ \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) | \mathcal{F}_n \right] - \nabla \Phi^a(\hat{\mathbf{S}}_n)$$

According to definition (28), we are led to compute:

$$\mathbb{E} \left[ \widehat{\nabla} \phi_{n+1}(\hat{\mathbf{S}}_n) | \mathcal{F}_n \right] = \mathbb{E} \left[ (Y_{n+1} - \hat{m}_n) \left( (Y_{n+1} - \hat{m}_n) [\mathbf{M}^{-1} s]_{U_{n+1}} - (Y_{n+1}^{(U)} - \hat{m}_n) \right) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right]$$

Let us use the decompositions:

$$Y_{n+1} - \hat{m}_n = (Y_{n+1} - \mathbb{E}[Y]) + (\mathbb{E}[Y] - \hat{m}_n) \quad \text{and} \quad Y_{n+1}^{(U)} - \hat{m}_n = (Y_{n+1}^{(U)} - \mathbb{E}[Y]) + (\mathbb{E}[Y] - \hat{m}_n)$$

and the simple algebra:

$$(\alpha + \epsilon) \left( (\alpha + \epsilon)C - (\beta + \epsilon) \right) = \alpha(\alpha C - \beta - \alpha) + \epsilon(2\alpha C - \alpha - \beta) + \epsilon^2(C - 1),$$

we deduce that

$$\begin{aligned} \widehat{\nabla} \phi_{n+1}(s) &= \nabla H(s, Y_{n+1}, Y_{n+1}^{(U)}, U_{n+1}) \\ &+ (\mathbb{E}[Y] - \hat{m}_n) \left( 2(Y_n - \mathbb{E}[Y]) [\mathbf{M}^{-1} s]_{U_{n+1}} - (Y_n - \mathbb{E}[Y]) - (Y_{n+1}^{(U)} - \mathbb{E}[Y]) \right) [\mathbf{M}^{-1}]_{U_{n+1},:}^T \\ &+ (\mathbb{E}[Y] - \hat{m}_n)^2 \left( [\mathbf{M}^{-1} s]_{U_{n+1}} - 1 \right) [\mathbf{M}^{-1}]_{U_{n+1},:}^T. \end{aligned}$$

We then compute the conditional expectation with respect to  $\mathcal{F}_n$  of the previous expression and combine our last decomposition with Equation (21) to get:

$$\begin{aligned} R_{n+1} &= (\mathbb{E}[Y] - \hat{m}_n) \mathbb{E} \left[ \left\{ 2(Y_n - \mathbb{E}[Y]) [\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - (Y_n - Y_{n+1}^{(U)}) \right\} [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &+ (\mathbb{E}[Y] - \hat{m}_n)^2 \mathbb{E} \left[ \left( [\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - 1 \right) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_k \right] \end{aligned} \quad (35)$$

We first consider the conditional expectation:

$$\begin{aligned} &\mathbb{E} \left[ \left\{ 2(Y_n - \mathbb{E}[Y]) [\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - (Y_n - Y_{n+1}^{(U)}) \right\} [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= \mathbb{E} \left[ 2(Y_n - \mathbb{E}[Y]) [\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &\quad - \mathbb{E} \left[ (Y_n - Y_{n+1}^{(U)}) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \end{aligned}$$

For the first term remark that  $Y_n$  is independent of  $U_{n+1}$  and of  $\mathcal{F}_n$ , therefore

$$\begin{aligned} & \mathbb{E} \left[ 2(Y_n - \mathbb{E}[Y]) [M^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= 2E[Y_n - \mathbb{E}[Y]] \mathbb{E} \left[ [M^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= 0 \end{aligned}$$

Similarly, we can rewrite the second term as

$$\begin{aligned} & E \left[ (Y_n - Y_{n+1}^{(U)}) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= E \left[ ((Y_n - \mathbb{E}[Y]) + (E[Y] - Y_{n+1}^{(U)})) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= -E \left[ (Y_{n+1}^{(U)} - \mathbb{E}[Y]) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \\ &= -E \left[ (Y^{(U)} - \mathbb{E}[Y]) [\mathbf{M}^{-1}]_{U,:}^T \right] \end{aligned}$$

Let us consider a coordinate of the above column vector

$$\mathbb{E}[(Y^{(U)} - \mathbb{E}[Y]) \mathbf{M}_{U,j}^{-1}] = \mathbb{E} \left[ \mathbb{E}[(Y^{(U)} - \mathbb{E}[Y]) | U] \mathbf{M}_{U,j}^{-1} \right] = 0.$$

As a consequence, the bias term simply reduces to the following  $q$ -dimensional vector:

$$R_{n+1} = (\mathbb{E}[Y] - \hat{m}_n)^2 \mathbb{E} \left[ ([\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - 1) [\mathbf{M}^{-1}]_{U_{n+1},:}^T | \mathcal{F}_n \right] \quad (36)$$

We are led to compute for  $s \in \Delta_q$ , and  $U$  the discrete random variable distributed according to  $a$  the following vector:  $\mathbb{E}[(\mathbf{M}^{-1} s)_U - 1] [\mathbf{M}^{-1}]_{U,:}^T$ . At this stage, we use the value of  $\mathbf{M}^{-1}$  given in Equation (9), the triangle inequality and get:

$$\|R_{n+1}\|_2 \leq (\mathbb{E}[Y] - \hat{m}_n)^2 \mathbb{E} \left[ |[\mathbf{M}^{-1} \hat{\mathbf{S}}_n]_{U_{n+1}} - 1| \left\| \mathbf{M}_{U_{n+1},:}^{-1} \right\|_2 | \mathcal{F}_n \right]$$

Since for any  $s \in \Delta_q$ , we have  $[\mathbf{M}^{-1} s]_U = \sum_{v \subset U} s_v \in [0, 1]$  and that

$$\left\| \mathbf{M}_U^{-1} \right\|_2^2 = \sum_{V \in \Upsilon} \mathbf{1}_{V \subset U} = 2^{|U|},$$

we then use our definition of  $\|a\|_{\text{exp},1/2}$  given in Equation (24) and get:

$$\|R_{n+1}\|_2 \leq \|a\|_{\text{exp},1/2} (\mathbb{E}[Y] - \hat{m}_n)^2.$$

We finally obtain the desired result while considering the global expectation:

$$\mathbb{E}[\|R_n\|_2] \leq \|a\|_{\text{exp},1/2} \mathbb{E}[(\mathbb{E}[Y] - \hat{m}_n)^2] \leq \|a\|_{\text{exp},1/2} \frac{\text{Var}[Y]}{n}.$$

□

**Lemma 5.2. Control of  $\mathbb{E}[\|\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n)\|^2]$**  A constant  $C$  independent of  $p$  exists such that:

$$\mathbb{E} \left[ \|\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n)\|^2 \right] \leq C \left( 1 + \sqrt{\|a\|_{\text{exp},2}} \right) \mathbb{E}[Y^4]$$

*Proof of Lemma 5.2.* Let us recall that

$$\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n) = (Y_{n+1} - \hat{m}_n) \left( (Y_{n+1} - \hat{m}_n)[\mathbf{M}^{-1}\hat{\mathbf{S}}_n]_{U_{n+1}} - (Y_{n+1}^{(U)} - \hat{m}_n) \right) [\mathbf{M}^{-1}]_{U_{n+1},:}$$

We can then compute the square norm and obtain:

$$\begin{aligned} \|\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n)\|^2 &= (Y_{n+1} - \hat{m}_n)^2 \left( (Y_{n+1} - \hat{m}_n)[\mathbf{M}^{-1}\hat{\mathbf{S}}_n]_{U_{n+1}} - (Y_{n+1}^{(U)} - \hat{m}_n) \right)^2 \|\mathbf{M}^{-1}\|_{U_{n+1},:}^2 \\ &\leq \left\{ 2(Y_{n+1} - \hat{m}_n)^4 ([\mathbf{M}^{-1}\hat{\mathbf{S}}_n]_{U_{n+1}})^2 + 2(Y_{n+1} - \hat{m}_n)^2 (Y_{n+1}^{(U)} - \hat{m}_n)^2 \right\} \|\mathbf{M}^{-1}\|_{U_{n+1},:}^2 \end{aligned}$$

We apply the Cauchy-Schwarz inequality and obtain:

$$\begin{aligned} \mathbb{E} \left[ \|\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n)\|^2 \right] &\leq 2\mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 ([\mathbf{M}^{-1}\hat{\mathbf{S}}_n]_{U_{n+1}})^2 \right] \\ &\quad + 2\sqrt{\mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 \|\mathbf{M}^{-1}\|_{U_{n+1},:}^4 \right]} \sqrt{\mathbb{E} \left[ (Y_{n+1}^{(U)} - \hat{m}_n)^4 \right]} \\ &\leq 2\mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 ([\mathbf{M}^{-1}\hat{\mathbf{S}}_n]_{U_{n+1}})^2 \right] \\ &\quad + 2\sqrt{\mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 \right]} \sqrt{\mathbb{E} \left[ \|\mathbf{M}^{-1}\|_{U_{n+1},:}^4 \right]} \sqrt{\mathbb{E} \left[ (Y_{n+1}^{(U)} - \hat{m}_n)^4 \right]}, \end{aligned}$$

where in the last line we used the independence between  $Y_{n+1}$  and  $U_{n+1}$ . From the same argument as the one used in Lemma 5.1, we have for any  $s \in \Delta_q$ :

$$[\mathbf{M}^{-1}s]_U = \sum_{v \subset U} s_v \in [0, 1] \quad \text{and} \quad \|\mathbf{M}^{-1}\|_{U,:}^4 = \left( \sum_{v \in \mathcal{Y}} \mathbf{1}_{v \subset U} \right)^4 = 2^{2|U|}. \quad (37)$$

We then deduce that:

$$\begin{aligned} \mathbb{E} \left[ \|\widehat{\nabla}\phi_{n+1}(\hat{\mathbf{S}}_n)\|^2 \right] &\leq 2\mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 \right] + \mathbb{E} \left[ (Y_{n+1} - \hat{m}_n)^4 \right] \sqrt{\mathbb{E} \left[ \|\mathbf{M}^{-1}\|_{U_{n+1},:}^4 \right]} \\ &\leq 16 \left( \mathbb{E}[(Y_{n+1} - \mathbb{E}[Y])^4] + \mathbb{E}[(\mathbb{E}[Y] - \hat{m}_n)^4] \right) \left( 1 + \sqrt{\mathbb{E} \left[ \|\mathbf{M}^{-1}\|_{U_{n+1},:}^4 \right]} \right) \\ &\leq 16 \left( \mathbb{E}[(Y_{n+1} - \mathbb{E}[Y])^4] + C_4 \frac{\mathbb{E}[Y^4]}{n^2} \right) \left( 1 + \sqrt{\|a\|_{\text{exp},2}} \right) \end{aligned}$$

where the last bound follows from Lemma A.3. Combining all these inequalities proves the result.  $\square$

# Appendices

## A State of the art results

We list below some technical results that will be useful for our purpose in the proof of Theorem 1.2 (a) and (b).

**Properties of Bregman divergence** We recall some standard results that are valid for any Bregman divergence  $\mathcal{D}_h$ . We refer to [41] for further details.

**Lemma A.1** (Three points lemma). *For any triple of points  $(x, y, z)$ , one has:*

$$\mathcal{D}_h(x, z) = \mathcal{D}_h(x, y) + \mathcal{D}_h(y, z) - \langle \nabla h(z) - \nabla h(y), x - y \rangle.$$

**Lemma A.2** (Gradient of the Bregman divergence). *For any pair of points  $(x, y)$ , one has:*

$$\nabla_x \mathcal{D}_h(x, y) = \nabla h(x) - \nabla h(y).$$

### Moments of sums of random variables

**Proposition A.3.** [Dharmadhikari and Jogdeo, see [44], Example 16 p. 60] *Assume that  $X_1, \dots, X_n$  are centered independent random variables with finite moments of order  $p \geq 2$ . Then an explicit constant  $C_p$  exists such that:*

$$\mathbb{E}[|X_1 + \dots + X_n|^p] \leq C_p n^{p/2-1} \sum_{k=1}^n \mathbb{E}|X_k|^p,$$

Moreover,  $C_p$  is universal and given by:

$$C_p = \frac{p(p-1)}{2} (1 \vee 2^{p-3}) \left(1 + \frac{2}{p} K_{2m}^{(p-2)/2m}\right) \quad \text{with } m = \lfloor p/2 \rfloor \quad \text{and} \quad K_{2m} = \sum_{r=1}^{2m} \frac{r^{2m-1}}{(r-1)!}.$$

**Robbins Sigmund Lemma** We shall state one of the most useful result on stochastic algorithms. This result is known as the Robbins-Sigmund Lemma and stated below.

**Lemma A.4** (Robbins-Sigmund Lemma, see [46]). *Consider a filtration  $(\mathcal{F}_n)_{n \geq 0}$  and 4 sequences of random variables  $(A_n)_{n \geq 0}$ ,  $(B_n)_{n \geq 0}$ ,  $(\alpha_n)_{n \geq 0}$  and  $(\beta_n)_{n \geq 0}$  that are  $\mathcal{F}_n$ -measurables, non-negatives and integrables such that these sequences enjoy the next assumptions:*

- (1)  $(\alpha_n)_{n \geq 0}$ ,  $(A_n)_{n \geq 0}$  and  $(\beta_n)_{n \geq 0}$  are  $\mathcal{F}_n$ -predictables.
- (2)  $\sup_{\omega \in \Omega} \prod_{n \geq 1} (1 + \alpha_n(\omega)) < +\infty$  and  $\sum_{n \geq 0} \mathbb{E}[\beta_n] < +\infty$ .
- (3)  $\mathbb{E}[B_{n+1} | \mathcal{F}_n] \leq (1 + \alpha_{n+1})B_n + \beta_{n+1} - A_{n+1}$ .

Then,

(i)  $B_n \rightarrow B_\infty$  in  $L^1$  where  $B_\infty$  is an integrable random variable and  $\sup_{n \geq 1} \mathbb{E}[B_n] < +\infty$

(ii)  $\sum_{n \geq 0} \mathbb{E}[A_n] < \infty$  and  $\sum_{n \geq 0} A_n < +\infty$  almost surely.

## B Properties on the matrix $\mathbf{M}$ and on the spectrum of $\nabla^2 \Phi^a$

This section is devoted to obtaining properties of the matrix  $\mathbf{M}$  depending on the dimension. Here, we will stress the dependency by writing

$$\Upsilon_p = \{u \subset \{1, \dots, p\}\}$$

and  $\mathbf{M}_p$  the matrix such that for all  $u, v \in \Upsilon_p$

$$[\mathbf{M}_p]_{u,v} = (-1)^{|u|-|v|} \delta_{v \subset u}.$$

Let us remark that the elements of  $\Upsilon_p$  can be ordered using the lexicographic order on numbers written in binary. The correspondence writes

$$\begin{aligned} \emptyset &= 0, & \{1\} &= 1 \\ \{2\} &= 10, & \{1, 2\} &= 11 \\ \{3\} &= 100, & \{1, 3\} &= 101, & \{2, 3\} &= 110, & \{1, 2, 3\} &= 111 \\ & \dots & & & & & & \end{aligned}$$

Furthermore

$$\Upsilon_{p+1} = \Upsilon_p \cup \{u \cup \{p+1\}, u \in \Upsilon_p\}$$

This recursion allows to write a recursion on matrices  $\mathbf{M}_p$  when lines and column are ordered as above:

$$\mathbf{M}_{p+1} = \begin{pmatrix} \mathbf{M}_p & 0 \\ -\mathbf{M}_p & \mathbf{M}_p \end{pmatrix}. \quad (38)$$

**Invert of  $\mathbf{M}$**  Upon closer examination of the matrix  $\mathbf{M}$ , the general term  $\mathbf{M}_{u,v}$  corresponds to the Möbius function for subsets ordered by inclusion<sup>2</sup>., denoted by  $\mu_\Upsilon(v, u) = (-1)^{|u|-|v|} \mathbf{1}_{v \subset u}$ . This, in turn, allows for the use of the Rota–Möbius inversion formula [47] that states that for two functions  $f$  and  $g$ :

$$\forall v \in \Upsilon : \quad g(v) = \sum_{v \subset u} f(u) \iff \forall u \in \Upsilon : \quad f(u) = \sum_{v \subset u} \mu_\Upsilon(v, u) g(v).$$

As a consequence

$$\forall (u, v) \in \Upsilon \times \Upsilon \quad \mathbf{M}_{u,v}^{-1} = \mathbf{1}_{v \subset u}.$$

---

<sup>2</sup>This observation was kindly shared with us by P. Rochet, to whom we extend our sincere thanks.

**Eigenvalues of  $\mathbf{M}_p \mathbf{M}_p^T$**  From (38), we deduce that if  $V_p = \mathbf{M}_p \mathbf{M}_p^T$ , then

$$V_{p+1} = \begin{pmatrix} V_p & -V_p \\ -V_p & 2V_p \end{pmatrix}$$

**Lemma B.1.** For any  $p \geq 1$  the eigenvalues of  $V_p = \mathbf{M}_p \mathbf{M}_p^T$  can be obtained by a recursion as

$$Sp(V_0) = \{1\}, \quad Sp(V_{p+1}) = \left\{ \frac{3+\sqrt{5}}{2}\lambda, \frac{3-\sqrt{5}}{2}\lambda, \lambda \in Sp(V_p) \right\}.$$

In particular the largest eigenvalue of  $V_p$  equals  $\left(\frac{3+\sqrt{5}}{2}\right)^p$  and its smallest is  $\left(\frac{3-\sqrt{5}}{2}\right)^p = \left(\frac{3+\sqrt{5}}{2}\right)^{-p}$

*Proof.* Let  $\lambda$  be an eigenvalue of  $V_p$  and  $u$  an eigenvector, then for  $\mathbf{a} \in \mathbb{R}$

$$V_{p+1} \begin{pmatrix} u \\ \mathbf{a}u \end{pmatrix} = \begin{pmatrix} V_p & -V_p \\ -V_p & 2V_p \end{pmatrix} \begin{pmatrix} u \\ \mathbf{a}u \end{pmatrix} = \begin{pmatrix} \lambda(1-\mathbf{a})u \\ \lambda(-1+2\mathbf{a})u \end{pmatrix} = \begin{pmatrix} \lambda(1-\mathbf{a})u \\ \lambda \frac{(-1+2\mathbf{a})}{\mathbf{a}} \mathbf{a}u \end{pmatrix}$$

As a consequence,  $\begin{pmatrix} u \\ \mathbf{a}u \end{pmatrix}$  is an eigenvector of  $V_{p+1}$  if and only if

$$1 - \mathbf{a} = \frac{(-1 + 2\mathbf{a})}{\mathbf{a}} \iff \mathbf{a}^2 + \mathbf{a} - 1 = 0.$$

This amounts to choose

$$\mathbf{a} = \frac{-1 \pm \sqrt{5}}{2}.$$

We deduce from this reasoning that if  $\lambda$  is an eigenvalue of  $V_p$  then  $\frac{3 \pm \sqrt{5}}{2} \lambda$  are eigenvalues of  $V_{p+1}$ . By recursion we obtain all the eigenvalues of  $V_{p+1}$  which concludes the proof.  $\square$

**Proposition B.2.** For any distribution  $a$  and any  $s \in \Delta_q$  the Hessian Matrix  $\nabla^2 \Phi^a(s)$  verifies:

$$\nabla^2 \Phi^a(s)_{i,j} = \text{Var}(Y) \mathbb{E}[\mathbf{M}_{U,i}^{-1} \mathbf{M}_{U,j}^{-1}] = \text{Var}(Y) \left( \sum_{u \in \{1, \dots, p\}} a_u \mathbf{M}_{u,i}^{-1} \mathbf{M}_{u,j}^{-1} \right), \quad \forall i, j \in \Upsilon.$$

Moreover, the lowest eigenvalue of  $\nabla^2 \Phi^a(s)$  verifies:

$$\inf \{ \lambda : \lambda \in Sp(\nabla^2 \Phi^a(s)) \} \geq \text{Var}(Y) \min_{u \in \Upsilon} a_u \left( \frac{3 - \sqrt{5}}{2} \right)^p.$$

*Proof.* The formula for the Hessian matrix derives from a simple computation. Concerning the smallest eigenvalue, let us remark that in the case of a uniform distribution  $a$ , the computation can be carried out explicitly:

$$\nabla^2 \Phi^{(unif)}(s) = \frac{\text{Var}(Y)}{2^p} (\mathbf{M} \mathbf{M}^T)^{-1}.$$

In the general case with a lower bounded discrete probability distribution  $a$ , we can apply Lemma B.1 and obtain that the lowest eigenvalue is  $\text{Var}(Y) \times \min_{u \in \Upsilon} a_u \times \left(\frac{3-\sqrt{5}}{2}\right)^p$ .  $\square$

## References

- [1] A. Barakat, P. Bianchi, W. Hachem, and S. Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. Electronic Journal of Statistics, 15(2):3892 – 3947, 2021.
- [2] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. Mathematics of Operations Research, 42(2):330–348, 2017.
- [3] J. D. Betancourt, F. Bachoc, T. Klein, D. Idier, R. Pedreros, and J. Rohmer. Gaussian process metamodeling of functional-input code for coastal flood hazard assessment. Reliability Engineering and System Safety, 198, June 2020.
- [4] B. Broto, F. Bachoc, and M. Depecker. Variance reduction for estimation of shapley effects and adaptation to unknown input distribution. SIAM/ASA Journal on Uncertainty Quantification, 8(2):693–716, 2020.
- [5] S. Bubeck. Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning, (8), 2015.
- [6] M. Costa, S. Gadat, and L. Huang. Cvar penalized portfolio optimization with biased stochastic mirror descent. Finance and Stochastics, 29:609–664, 2025.
- [7] S. Da Veiga and F. Gamboa. Efficient estimation of sensitivity indices. Journal of Nonparametric Statistics, 25(3):573–595, 2013.
- [8] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. Basics and Trends in Sensitivity Analysis: Theory and Practice in R. SIAM, 2021.
- [9] S. Da Veiga, F. Gamboa, A. Lagnoux, T. Klein, and C. Prieur. New estimation of sobol’indices using kernels. arXiv preprint arXiv:2303.17832, 2023.
- [10] S. Da Veiga, F. Gamboa, A. Lagnoux, T. Klein, and C. Prieur. Efficient estimation of Sobol’ indices of any order from a single input/output sample. 2024.
- [11] S. Da Veiga, F. Wahl, and F. Gamboa. Local polynomial estimation for sensitivity analysis on models with correlated inputs. Technometrics, 51(4):452–463, 2009.
- [12] E. De Rocquigny, N. Devictor, and S. Tarantola. Uncertainty in industrial practice. Wiley Online Library, 2008.
- [13] L. Devroye, P. G. Ferrario, L. Györfi, and H. Walk. Strong universal consistent estimate of the minimum mean squared error. In Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, pages 143–160. Springer, 2013.

- [14] L. Devroye, L. Györfi, G. Lugosi, and H. Walk. A nearest neighbor estimate of the residual variance. Electronic Journal of Statistics, 12(1):1752–1778, 2018.
- [15] L. Devroye, D. Schäfer, L. Györfi, and H. Walk. The estimation problem of minimum mean squared error. Statistics & Decisions, 21(1):15–28, 2003.
- [16] A. Doury, S. Somot, and S. Gadat. On the suitability of a convolutional neural network based rcm-emulator for fine spatio-temporal precipitation. Climate Dynamics, 62(9):8587–8613, 2024.
- [17] A. Doury, S. Somot, S. Gadat, A. Ribes, and L. Corre. Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. Climate Dynamics, 60(5):1751–1779, 2023.
- [18] R. Dragomir, A. Taylor, A. d’Aspremont, and J. Bolte. Optimal complexity and certification of bregman first-order methods. Mathematical Programming, pages 41–83, 2022.
- [19] R. A. Dragomir, M. Even, and H. Hendriks. Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. In M. Meila and T. Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 2815–2825. PMLR, 18–24 Jul 2021.
- [20] M. Duffo. Random iterative models, volume 34. Springer Science & Business Media, 2013.
- [21] B. Efron and C. Stein. The Jackknife Estimate of Variance. Annals of Statistics, 9(3):586–596, 1981.
- [22] S. Gadat, T. Klein, and C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. Annals of Statistics, 44(3):982–1009, 2016.
- [23] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. Electronic Journal of Statistics, 12:461–529, 2016.
- [24] F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. Bernoulli, 28(4):2345–2374, 2022.
- [25] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. Electronic Journal of Statistics, 8:575–603, 2014.
- [26] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol Pick-Freeze Monte Carlo method. Statistics, 50(4):881–902, 2016.
- [27] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013.
- [28] Godichon-Baggioni, Antoine. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. ESAIM: PS, 23:841–873, 2019.

- [29] L. Györfi and H. Walk. On the asymptotic normality of an estimate of a regression functional. J. Mach. Learn. Res., 16:1863–1877, 2015.
- [30] M. B. Heredia, C. Prieur, and N. Eckert. Nonparametric estimation of aggregated sobol’indices: application to a depth averaged snow avalanche model. Reliability Engineering & System Safety, 212:107422, 2021.
- [31] W. Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Statistics, 19:293–325, 1948.
- [32] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. ESAIM: Probability and Statistics, 18:342–364, 1 2014.
- [33] G. Lan, A. Nemirovskij, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. Math. Program., Ser. A, (134):425–458, 2012.
- [34] E. Liitiäinen, F. Corona, and A. Lendasse. On nonparametric residual variance estimation. Neural Processing Letters, 28(3):155–167, 2008.
- [35] E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. Journal of Multivariate Analysis, 101(4):811–823, 2010.
- [36] J. Liu and Y. Yuan. Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. Journal of Machine Learning Research, 25(271):1–40, 2024.
- [37] A. Marrel, B. Iooss, S. Da Veiga, and M. Ribatet. Global sensitivity analysis of stochastic computer models with joint metamodels. Statistics and Computing, 22(3):833–847, 2012.
- [38] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. W. Jones, editors, Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications, chapter 4, pages 55–99. Elsevier, 2006.
- [39] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Advances in neural information processing systems, 24, 2011.
- [40] A. Nemirovsky and D. Yudin. Problem Complexity and Method Efficiency in Optimization. John Wiley, 1983.
- [41] A. Nemirovsky, D. Yudin, and E. Dawson. Wiley-interscience series in discrete mathematics, 1983.
- [42] K. Pearson. On the partial correlation ratio. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 91(632):492–498, 1915.

- [43] N. Peteilh, T. Klein, T. Druot, N. Bartoli, and R. P. Liem. Challenging Top Level Aircraft Requirements based on operations analysis and data-driven models, application to take-off performance design requirements. In AIAA AVIATION 2020 FORUM, AIAA AVIATION 2020 FORUM, Reno, NV, United States, June 2020. American Institute of Aeronautics and Astronautics, American Institute of Aeronautics and Astronautics.
- [44] V. Petrov. Sums of independent random variables. Springer Verlag, 1975.
- [45] E. Plischke and E. Borgonovo. Fighting the curse of sparsity: Probabilistic sensitivity measures from cumulative distribution functions. Risk Analysis, 40(12):2639–2660, 2020.
- [46] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Optimizing methods in statistics. Elsevier, pages 233–257, 1971.
- [47] G.-C. Rota. On the foundations of combinatorial theory I. theory of Möbius functions. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 2(4):340–368, 1964.
- [48] A. Saltelli, K. Chan, and E. Scott. Sensitivity analysis. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.
- [49] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. Global sensitivity analysis: the primer. Wiley Online Library, 2008.
- [50] T. J. Santner, B. Williams, and W. Notz. The Design and Analysis of Computer Experiments. Springer-Verlag, 2003.
- [51] O. Sebbouh, R. M. Gower, and A. Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In M. Belkin and S. Kpotufe, editors, Proceedings of Thirty Fourth Conference on Learning Theory, volume 134 of Proceedings of Machine Learning Research, pages 3935–3971. PMLR, 15–19 Aug 2021.
- [52] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. Math. Modeling Comput. Experiment, 1(4):407–414 (1995), 1993.
- [53] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1-3):271–280, 2001.
- [54] M. Solís. Non-parametric estimation of the first-order sobol indices with bootstrap bandwidth. Communications in Statistics-Simulation and Computation, 50(9):2497–2512, 2021.
- [55] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering & System Safety, 93(7):964–979, 2008.