Master thesis internship in mathematical statistics
Online estimation of Sobol indices, stochastic miror descent optimisation algorithm

Student profile: Master 2 student in Applied Probability, Statistics, Data Science. Programming skills in Python will be appreciated.

Keywords: Machine Learning, Applied Probability and Statistics, Sensitivity analysis

Advisor and contact:

- Thierry Klein, thierry.klein@math.univ-toulouse.fr
- Sébastien Gadat, sebastien.gadat@tse-fr.eu

Working place: Institut de Mathématiques de Toulouse or ENAC

Length: 5 to 6 months, starting in February/March 2025

Remuneration: Standard internship grant : 550 euros/month.

Context This internship proposal is part of the research project funded by the ANR grant "*Gatsbii*", a pluri-disciplinary research project between Enac, Université Paul Sabatier, EDF R&D and Thales, Paris Dauphine and l'ENSAE. The project deals with the use of applied mathematics, computer science, game theory and A.I. on sensitivity analysis (S.A. for short below) for applications in various industrial fields, including the ones of EDF and Thales.

Objectives and Contents of the Internship:

We briefly describe below some tools of S.A. before describing the goal of the internship.

Mathematics of S.A. Assuming that $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d$ is a random vector with d *independent* components and G an application from $\mathbb{R}^d$ to $\mathbb{R}$. We define $Y := G(\mathbf{X})$ and assume that Y is a square-integrable random variable. For $D = \{1, \ldots, d\}$, we denote by $\mathcal{P}(D)$ the set of all subsets of D. A classical and important result is the Hoeffding decomposition of the variance [2, 1], that leads to

$$\mathrm{Var}(Y) = \mathrm{Var}(G(\mathbf{X})) = \sum_{u \in \mathcal{P}(D),\ u \neq \emptyset} V_u\,, \tag{1}$$

with

$$V_u = \sum_{v \subset u} (-1)^{|u|-|v|}\, \mathrm{Var}([Y|X_i, i \in v])$$

where $|\cdot|$ denotes the cardinality of a subset. The Sobol, closed Sobol and Total Sobol indices with respect to $(X_i)_{i \in u}$ are respectively defined respectively by

$$S_u = \frac{V_u}{Var(Y)},\ S_u^{cl} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i,\ i \in u])}{Var(Y)},\ S_u^T = 1 - S_{\overline{u}}^{cl}$$

where $\overline{u} = \{1, \ldots, p\} \setminus u$. Hence dividing both side of (1) by $\mathrm{Var}(Y)$ we get $1 = \sum_{u \subset \{1,\ldots,p\} \setminus \emptyset} S_u$.

Let $\mathbf{S} = (S_u)_{u \subset \{1,\ldots,p\} \setminus \emptyset}$, and $\mathbf{S}^{cl} = (S_u^{cl})_{u \subset \{1,\ldots,p\} \setminus \emptyset}$ be the vectors of size $q = 2^p - 1$ of respectively all Sobol indices and closed Sobol indices as soon as we have ordered the subsets of $\{1, \ldots, p\}$ by the graphical lexical order. Now the square matrix M of size q such that

$$\mathbf{S} = M\mathbf{S}^{cl}$$

is triangular with diagonal terms equal to 1. More precisely from (1) and (**??**), we have that $\forall u, v \subset \{1, \cdots, p\} \setminus \emptyset$

$$M_{u,v} = (-1)^{|u|-|v|} 1_{v \subset u}$$

These indices are in practice used to detect influent parameters in complex models , unfortunately their exact computations are impossible and the practitioner needs to contruct estimators of these unknown quantities.

Over the last thirty years, considerable effort has been put on providing efficient estimation strategies for these indices, as well as strong theoretical guarantees for estimators mostly obtained from various sampling strategies (for Monte Carlo sampling, Pick-Freeze schemes or space-filling designs, [1, 3] for a review). The aim of this internship is to compare the classical Pick freeze method, with a new method based specific optimisation algorithms

Pick freeze representation of Sobol's indices Let $(X'_1, \ldots, X'_p)$ be an independant copy of $(X_1, \ldots, X_p)$ For any $u \subset \{1, \ldots, p\} \setminus \emptyset$ one has

$$\mathrm{Var}(\mathbb{E}[Y|X_i, \ i \in u]) = \mathrm{Cov}(Y, Y^u)$$

where

$$Y^u = f(X_1^u, \ldots, X_p^u)$$

with $X_i^u = X_i$ if $i \in u$ and $X_i^u = X'_i$ if $i \notin u$. Hence

$$S_u^{cl} = \frac{\mathrm{Cov}(Y, Y^u)}{\mathrm{Var}(Y)} \tag{2}$$

Sobol indices as solutions of an optimisation problem. Define the function from $\mathbb{R}^q$ to : $x = (x_u)_{u \subset \{1,\ldots,p\} \setminus \emptyset}$ by

$$\Psi^a(x) = \frac{1}{2} \sum_{u \subset \{1,\ldots,p\} \setminus \emptyset} a_u \mathbb{E}[((Y - \mathbb{E}[Y]) x_u - (Y^u - \mathbb{E}[Y]))^2] \tag{3}$$

where $a = (a_u)_{u \subset \{1,\ldots,p\} \setminus \emptyset}$ is a probability distribution. The function Ψ^a is minimised for $x = S^{cl}$, and we can write

$$\hat{\mathbf{S}}^{cl} = \mathrm{argmin}_{s \in {}^q} \left\{ \Psi^a(s), \sum_{u \in \{1, \cdots, p\} \setminus \emptyset} [Ms]_u = 1 \right\}.$$

then we have

$$\hat{\mathbf{S}} = \mathrm{argmin}_{s \in \Delta_q} \left\{ \Psi^a(M^{-1} y), \sum_{u \in \{1, \cdots, p\} \setminus \emptyset} [s]_u = 1 \right\}. \tag{4}$$

where Δ_q is the simplex $\Delta_q = \left\{ x = (x_u)_{u \in \{1, \cdots, p\} \setminus \emptyset}, \sum_{i \in \{1, \cdots, p\} \setminus \emptyset} x_u = 1 \right\}$.

The pick freeze representation and Equation (4) lead to two natural methods to construct estimators of Sobol indices. The first one called the Pick freeze estimator his simply the empirical version of (2) and has been widely studied in the literature The second one is up to our knowledge new.

Aim of the internship

- To study the classical Pick freeze estimator: construction , convergences properties (onsistancy, rate of convergence, efficiency).
- To build stochastic gradient descent alogorithm to solve the optimisation problem given by (4). In particular to understand how one can use mirror descent algorithm to take into account the constraints.

Associated to the internship project, a PhD grant may be foreseen according to the evolution of the research project, the different opportunities of fundings in Toulouse and the motivation of the student. Any student interested in this research project shall contact us to plan a short zoom meeting interview.

References

[1] S. **Da Veiga**, F. Gamboa, **B. Iooss**, and C. Prieur. *Basics and trends in sensitivity analysis: Theory and practice in R.* SIAM, 2021.

[2] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Math. Statistics*, 19(3):293–325, 1948.

[3] C. Prieur and S. Tarantola. Variance-based sensitivity analysis: theory and estimation algorithms. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of UQ*, chapter 35, pages 1217–1239. Springer International Publishing, 2017.