

Research internship at Scool

(Inria center at the University of Lille)

Lower bound on the error of nonparametric estimation and tests on the mean with small sample-size

Keywords: Sequential statistics, Information theory, Robust statistics.

Supervision: The intern will be advised by Timothée Mathieu (timothee.mathieu@inria.fr) and Rémy Degenne from Inria team-project Scool.

Location: Inria Lille, 40 avenue Halley, 59650 Villeneuve d'Ascq, France.

Introduction of Internship's subject At the core of scientific experimental studies, there is the question of testing a hypothesis: is a new medical treatment better than a placebo? Is a new algorithm more efficient on a computational task than the state of the art? Because of random fluctuations, it is necessary to repeat experiments over several trials (to gather several *samples*) before making a decision. For those applications, we need statistical methods for hypothesis testing that come with formal guarantees on their probability of error. Those methods should also use as few trials as possible in order to be applicable to domains where gathering samples is long or costly. In order to use as little samples as possible, the tests and estimation procedures should adapt to the observations as they come. It is also crucial to design lower bound on the error as both an inspiration to construct estimators as well as a means to ascertain their optimality.

Mean estimation is a basic task in statistics and consists in finding an estimator of the expectation $\mathbb{E}_P[X]$ of a probability distribution P , given only access to $n \in \mathbb{N}$ independent samples $X_1, \dots, X_n \in \mathbb{R}$. When the distribution is Gaussian, the empirical mean is an efficient estimator of the mean and for any $\delta \in (0, 1)$, with probability larger than $1 - \delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}_P[X] \right| \leq \sqrt{2\text{Var}_P[X] \frac{\log(1/\delta)}{n}}. \quad (0.1)$$

where $\text{Var}_P[X]$ denotes the variance of P . This is a high probability guarantee on the rate of convergence to the mean.

On the other hand, following a now classical method of reduction to testing called Le Cam's method (see for example [5]), it is possible to design lower bounds on the estimation error that are nonparametric, non-asymptotic and true with high probability. Suppose that all distributions in \mathcal{D} have finite mean and suppose we want to design a lower bound on the error of an estimator $\hat{\mu}_n$. We have, for any error rate $x : \mathcal{D} \rightarrow \mathbb{R}_+$,

$$\sup_{P \in \mathcal{D}} \mathbb{P}_P (|\hat{\mu}_n - \mathbb{E}_P[X]| \geq x(P)) \geq \sup_{c \in \mathbb{R}} \max \left\{ \sup_{\substack{P \in \mathcal{D}, \\ \mathbb{E}_P[X] \leq c - x(P)}} \mathbb{P}_P(\hat{\mu}_n \geq c), \sup_{\substack{P \in \mathcal{D}, \\ \mathbb{E}_P[X] \geq c + x(P)}} \mathbb{P}_P(\hat{\mu}_n < c) \right\}, \quad (0.2)$$

Where $\mathbb{P}_P(E)$ denotes the probability of event E under the law P . Intuitively, this equation says that in order to estimate $\mathbb{E}_P[X]$ with a precision smaller than $x(P)$, we should be able to distinguish P with distribution that have a mean larger than $\mathbb{E}_P[X] + x(P)$ and we should be able to distinguish P with distribution that have a mean smaller than $\mathbb{E}_P[X] - x(P)$, for whatever value of $\mathbb{E}_P[X]$ can take (which introduces the c variable).

Using the Kullback-Leibler (KL) divergence, we can bound the right-hand-side of Equation (0.2). More precisely, in the case of finite variance, if an estimator $\hat{\mu}_n$ satisfies $\mathbb{P}(|\hat{\mu}_n - \mathbb{E}_P[X]| > x) \leq \delta$ where P has a variance $\text{Var}_P[X]$ which smaller than σ^2 and belong to some set of distributions \mathcal{G} that is invariant by translations, then necessarily [2],

$$\frac{\log(\frac{1}{4\delta})}{2n} \leq \inf_{P \in \mathcal{S}_0(\mathbb{R})} \inf \{ \text{KL}(P, G), G \in \mathcal{G} \text{ s.t. } \mathbb{E}_G[X] = x, \text{Var}_G[X] \leq \sigma^2 \}, \quad (0.3)$$

where $\mathcal{S}_0(\mathbb{R})$ are the probability distributions that are symmetric around 0. It has been shown [1, 3] that there exists estimators on the mean that achieve sub-Gaussian rate of convergence even when the data are not Gaussian but only posses a finite variance. More recent works show that there are estimators that can be more adaptive to P and have guarantees that depend on the KL [2, 4]. The main goal of this internship would be to study lower bounds on the estimation error in more general setting (multivariate setting, possibly with corruption) and to use this lower bound to construct an efficient estimator of the mean.

Internship’s main goal This internship focuses on theoretical research, and for this reason it requires a candidate with a strong background in mathematical statistics. In particular, we will look at testing on the mean of \mathbb{R}^d or Hilbert-space valued random variables, with or without corruption which has been a subject of a lot of research in recent years. The primary goal of the internship will be to learn how to construct lower bound using Information Theory methods like Le Cam’s method and to apply this method to construct information lower bound on mean estimation in \mathbb{R}^d for various setting (with/without corrupted data, with/without finite second moment...) in a way that adapts to the distribution. A secondary goal will be to construct and empirically test an estimator that have a rate of convergence which matches the lower bound. Depending on the output of the study, a publication in an international conference or journal will be considered.

Host institution and supervision The student will be hosted at Centre Inria de l’Université de Lille, in the Scool team. Scool (Sequential COntinual and Online Learning) is an Inria team-project. The research topic of Scool is the study of the sequential decision-making problem under uncertainty, most of our activities are related to either bandit problems, or reinforcement learning problems. More information, please visit <https://team.inria.fr/scool/projects>.

The advisors are focused mainly on sequential statistics and bandits, and although most of our work contain empirical validation of the techniques proposed we mainly work on the theoretical foundations of sequential learning and statistics. This internship is proposed in the context of the ANR project STRESS (Statistical Testing and Robust Estimation in the Small Sample regime) and may result in a Ph.D. if the internship went well.

References

- [1] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, nov 2012.
- [2] Rémy Degenne and Timothée Mathieu. Information lower bounds for robust mean estimation. *arXiv preprint arXiv:2403.01892*, 2024.
- [3] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695 – 2725, 2016.
- [4] Shivam Gupta, Jasper CH Lee, and Eric Price. Finite-sample symmetric mean estimation with fisher information rate. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4777–4830. PMLR, 2023.
- [5] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009.