Master 2 research internship 2018-2019

Statistical Inference under Differential Privacy

Béatrice Laurent, Institut de Mathématiques de Toulouse, INSA de Toulouse
beatrice.laurent@insa-toulouse.fr

Sébastien Gerchinovitz, Institut de Mathématiques de Toulouse, Université Paul Sabatier
sebastien.gerchinovitz@math.univ-toulouse.fr

Context In modern statistical inference problems, the existence of big datasets is a good news for statistical performances, but it also raises important privacy issues. Case studies have indeed shown that it is sometimes possible to identify an individual within an anonymous dataset if we compare these data to another non-anonymous dataset that shares some common data; see, e.g., [1]. An important mathematical challenge is thus to understand to what extent we can modify statistical methods (for, e.g., parametric or non-parametric estimation, confidence intervals, tests, etc) in order to protect the privacy of each individual in the dataset while keeping good statistical performances.

For this internship, we consider the formal definition of privacy known as *differential privacy*. It was introduced in the cryptography literature and later considered, for example, in Dwork [3] or in Wasserman [5].

It has been shown by Duchi et al [2] that under differential privacy, one can quantify the achievable performances of some statistical estimation procedures. More precisely, they provide minimax bounds for various statistical estimation problems under the notion of differential privacy. They show that in some cases, even a simple parametric problem such as the estimation of the mean of a n sample leads to very poor minimax rates under privacy. In short: privacy constrainsts can deteriorate the statistical performances. A key point to get such results is to prove sharp bounds for information theoretic quantities, that rely the Kullback-Leibler divergence between two private samples with the total variation distance between the initial distributions.

Goal of the intership The performances of testing procedures under privacy have been less studied. Simulation results for goodness-of-fit and independence chi-square tests are given in Gaboardi et al. [4].

The aim of the internship is to study the papers cited in the References and to deepen the theoretical properties of testing problems under differential privacy constraints.

Research environment This internship will take place at INSA de Toulouse. The intern will receive the monthly French 'gratification' and will be supervised by Béatrice Laurent and Sébastien Gerchinovitz. Please do not hesitate to contact us if you have any questions.

References

[1] Abowd, J.M., Schmutte, I. M. (2018) *An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices*, To appear in American Economic Review.

[2] Duchi, John C., Jordan, Michael I., Wainwright, Martin, J. (2017) *Local privacy, Data processing inequalities and Minimax rates*, https://arxiv.org/pdf/1302.3203.pdf

[3] Dwork, C. and Smith, A., (2009), *Differential privacy for statistics : what we know and what we want to learn*, Journal of Privacy and Confidentiality 1, Number 2, pp. 135-154.

[4] Gaboardi, M., Lim, H.W., Rogers, R. and Vadhan, S.P. (2016) *Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing* https://arxiv.org/pdf/1602.03090.pdf

[5] Wasserman, L., Zhou, S., (2009) *A statistical framework for differential privacy*, https://arxiv.org/pdf/0811.2501.pdf