Non-parametric Inference under Local Differential Privacy

Cristina BUTUCEA

CREST, ENSAE, IP Paris

Data privacy protection is a major issue for our society nowadays due to the massive amounts of data collected and stored by many electronic devices at all times, on social networks, in medecine, in finance and so on. This leads to multiple sources of data concerning the same individuals (persons, funds, etc.) that can be easily aggregated in order to identify them. Therefore, privacy preserving mechanisms have to be applied to the data before their public release which implies to quantify the amount of privacy, but also to decide a priori whether collaboration between data holders is possible/authorized or unadvisable/forbidden.

Local differential privacy The concept of differential privacy (see Dinur and Nissim 2003, Dwork 2008, Dwork and Nissim 2004, Dwork *et al.* 2006, Evfimievski *et al.* 2003) provides a rigorous formalism to randomize data and quantify the amount of privacy. We consider n individuals with outcomes $X_1, \ldots, X_n$ supposed to be i.i.d. with common probability distribution $P \in \mathcal{P}$. This is called the original data and we need to produce a randomized observation Z that the statistician is allwoed to use in order to recover information about the distribution P. The conditional distribution of Z given $X = (X_1, \ldots, X_n)$ is denoted by Q and referred to as a channel distribution or a privatization scheme, i.e. $Pr(Z \in A|X = x) = Q(A|x)$.
We will introduce the notion of $\alpha-$differential privacy for some $\alpha \in (0, \infty)$. We distinguish global (or central) differential privacy when the privacy mechanism uses all the original data, vs. local differential privacy when each sample from the original data is privatized on the user's local machine before its release. In the sequel, we consider only the setup of local differential privacy (LDP) where slower rates are typically attained as compared to the optimal procedures that use the original data.

Estimation of the probability density function Next, we discuss recent results on non-parametric estimation of the common probability distribution f of P (Duchi et al., 2018, Rohde and Steinberger, 2019, Butucea et al. 2020). We construct $\alpha-$LDP privacy mechanisms and build projection estimators (histogram and wavelet estimators) that are minimax optimal (up to $\log$ factors) when the function f belongs to Besov $s-$smoothness classes. Optimality is shown with respect to all estimators but also with respect to all possible $\alpha-$LDP privacy mechanisms Q. Thus, sequentially interactive privacy mechanisms that use at each step i X_i together with the previously released $Z_1, ..., Z_{i-1}$, may provide more flexibility but do not improve over non-interactive ones, that use only X_i at each step i. We discuss adaptive methods to the smoothness s and new elbow effects.

Goodness-of-fit tests Finally, we address the non-parametric goodness-of-fit problem of testing whether $f \equiv f_0$ for some given smooth f_0 vs. $\|f - f_0\|_2 \geq \rho$. The goal is to determine the optimal separation rate ρ in the $\alpha-$LDP context, that is the value separating the set of functions f undistinguishable from f_0 from the set of functions f far enough from f_0 so that we can build tests with error probabilities tending to 0. A related problem that we also discuss is the estimation of the quadratic functional $\int f^2$.
We discuss in more details information theoretic inequalities and describe the lower bounds techniques (Duchi et al., 2013, Lam-Weil et al. 2020). It is highly surprising that sequentially interactive privacy mechanisms lead to faster minimax rates than non-interactive ones in these problems (Butucea et al. 2020). Similar behaviour can be found in the goodness-of-fit test of a given discrete distribution (Berrett and Butucea, 2020).

References

BERRETT, T. and BUTUCEA, C. (2020) Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms; arxiv:2005.12601

BUTUCEA C., DUBOIS A., KROLL M. and SAUMARD, A. (2019) Local differential privacy: elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli, to appear*.

BUTUCEA C., ROHDE, A. and STEINBERGER, L. (2020). Interactive versus non-interactive locally, differentially private estimation: Two elbows for the quadratic functional. arxiv:2003.04773

DINUR, I. and NISSIM, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–210. ACM.

DUCHI, J.C., JORDAN, M.I. and WAINWRIGHT, M.J. (2013) Local privacy and minimax bounds: sharp rates for probability estimation. *Advances in Neural Information Processing Systems.*

DUCHI, J.C., JORDAN, M.I. and WAINWRIGHT, M.J. (2018) Minimax optimal procedures for locally private estimation. J. Amer. Statist. Assoc. 113 182-201

DWORK, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 1–19. Springer.

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, ADAM (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265–284.

DWORK, C. and NISSIM, K. (2004). Privacy-preserving datamining on vertically partitioned databases. In *Annual International Cryptology Conference*, 528–544. Springer.

EVFIMIEVSKI, A., GEHRKE, J. and SRIKANT, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD- SIGACT-SIGART Symposium on Principles of Database Systems*, 211–222. ACM.

LAM-WEIL, J., LAURENT, B. and LOUBÈS, J.-M. (2020) Minimax optimal goodness-of-fit testing for densities under a local differential privacy constraint. arxiv:2002.04254

ROHDE, A. and STEINBERGER, L. (2019). Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.*, to appear.