Second order methods for online stochastic optimization

With the advent of Big Data, new issues in statistical data processing has emerged. One of them is the real-time statistical processing of data when they arrive in continuous flow. In this case, data are not stored or they are sequentially acquired from files too large to be loaded into memory. Thus, traditional computational procedures are not adapted. Recursive methods are then required for the statistical processing of these data, among which we have stochastic algorithms. Stochastic algorithms, introduced in the 1950's by Robbins and Monro (1951), are partly used as tools of stochastic optimization to minimize a function, which is unknown but is written as the mathematical expectation of a known function depending on a random vector. To estimate the solution of this minimization problem, first-order algorithms such as stochastic gradient algorithms have become, in the context of Big Data, essential in recent years. They are indeed very fast in terms of computation time, they allow to perform machine learning tasks on large data sets and to process these data sequentially. From a theoretical point of view, these algorithms were first developed in the 90's (Polyak and Juditsky (1992) ; Duflo (1997) ; Pelletier (1998,2000)) and have reached their peak in recent years Bach and Moulines (2013) ; Godichon-Baggioni (2016) ; Gadat and Panloup (2017). Finally, in order to adapt these methods in the case where data are collected by different machines, parallelized versions of these algorithms have been recently studied Godichon-Baggioni and Saadane (2020) ; Patel and Dieuleveut (2019). However, these types of methods can encounter a number of problems. In particular, first-order algorithms only take into account the information given by the gradient of the function which we try to minimize, and they can be very sensitive to the case where the spectrum of Hessian of the function which we try to minimize is large, in the sense that the eigenvalues are at very different scales (see the example in Section 5.2 in Bercu et al. (2019) for evidence of this).

A solution to overcome the problems encountered by stochastic gradient algorithms is to implement second-order methods, i.e. taking into account the information given by the Hessian matrix of the function that we want to minimize. The first difficulty encountered by this method is that we do not necessarily know the Hessian matrix and it is therefore necessary to construct a recursive estimator of it in order to preserve the sequential character of the stochastic algorithm. The second difficulty is that we are more interested in its inverse, and in the spirit of online algorithms, we must be able to update the inverse of the Hessian estimator with a cost, in terms of computation time, as low as possible.

Different methods have been studied in recent years. A stochastic Quasi-Newton algorithm was introduced in Byrd et al. (2016), which can be viewed as a stochastic BFGS algorithm. This method aims to replace the inverse of the Hessian with a matrix which has similar behavior but can be easily estimated. However, this method does not allow us to obtain asymptotically efficient estimators, i.e. having an optimal performance when the sample size is large.
More recently in Bercu et al (2019), a stochastic Newton algorithm has been proposed, to estimate the parameters of a logistic regression model. The recursive estimator of the inverse of the Hessian matrix is updated using a very computationally inexpensive Ricatti formula (also called Sherman-Morrison formula). The authors have shown that the estimators thus obtained have an asymptotically optimal behavior, but this algorithm is only applicable to case of logistic regression.
In Leluc and Portier (2020), the authors introduce a "conditioned" stochastic gradient algorithm, which can be assimilated into stochastic Newton algorithms in some cases, and they show that the

estimators obtained can be asymptotically efficient. However, in this case, the inversion of the Hessian estimator is very expensive in terms of computation time.
Finally, a new second-order stochastic Gauss-Newton algorithm has been proposed particularly for the case of parametric nonlinear regression in Cénac et al. (2020). Thanks to an original approach, the convergence of this algorithm has been demonstrated and numerical experiments have confirmed the performance of this new second order stochastic algorithm.

PhD project.

The research work will start with the construction and study of second-order stochastic algorithms in the framework of constrained minimization problems of ridge regression. This work should lead to an article during the first year of the program, with an implementation on real data proposed by Atmo Normandie, the association in charge of the air quality monitoring in Normandy

To our knowledge, there is not yet a "universal" approach for stochastic Newton algorithms. The main problem comes from the implementation of online estimators of the inverse of the Hessian matrix. In Bercu et al. (2019) and Cénac et al. (2020), these estimators were constructed using the Riccati formula. However, the latter can only be applied if the Hessian has a very specific form and thus a generalization cannot be found via this approach.
Several objectives of work are envisaged in this project. A first objective would be to estimate directly the inverse of the Hessian matrix via a Robbins-Monro algorithm (Robbins and Monro (1951)) and/or its averaged version Polyak and Juditsky (1992). This approach could be applied for a very large spectrum of functions to be minimized, and the purpose would therefore be to show that the estimators thus obtained are asymptotically efficient. This work would lead to a possible publication during the second year of the program.

Another objective of the research would be not only to propose parallelized versions of the stochastic Newton algorithms, by focusing on machine-to-machine communication allowing to obtain better estimators of the Hessian, but also to ensure that the estimators thus obtained are always asymptotically efficient (inspired in particular by Godichon-Baggioni and Saadane (2020)). These methods will be tested on MNIST data which consist of handwritten images of digits. The purpose is to apply the parallelized Newton estimators on a training sample to estimate the parameter, before making predictions on the test sample. This work would lead to a possible publication during the third year of the program.

Another objective would be to give the speed of convergence in root mean square of the estimators, in order to guarantee finite time results. We will distinguish two cases : the case where the function is strongly convex, and we could then draw inspiration from Bach and Moulines (2013), and the case where the function is strictly convex, and we could then draw inspiration from Godichon-Baggioni (2016) ; Gadat and Panloup (2017). This work would lead to a possible publication at the end of the program. The rest goals are much more exploratory, and will depend on the progress of the PhD student.

Another purpose would be to adapt the different methods implemented during the research to obtain convergence results (with a less restrictive framework than the current one) for algorithms like Adagrad (Duchi et al. (2011)) or Adadelta (Zeiler (2012)). In fact, the latter, in some cases, can be considered as a compromise between stochastic gradient algorithms and stochastic Newton

algorithms, as they can take into account partial information of the Hessian, while being less expensive in terms of computation time. An interesting approach could be to build and study hybrid algorithms combining Adagrad and stochastic Newton or Gauss-Newton algorithms. The idea is to improve the performance of the latter when information of the Hessian matrix is available.

Numerical experiments will complete the theoretical study conducted on the proposed algorithms, in order to prove and test their performances on simulated and real data.

Scientific context of the program.

The supervision will be provided jointly by Bruno Portier (LMI, PR) and Antoine Godichon-Baggioni (LPSM, Sorbonne University, MCF). The PhD student will work within the LMI. He will participate in the Statistics working group organized by the LMRS of the University of Rouen. Other collaborations will be possible, especially for the application part of the PhD (Atmo Normandie).

References.

— Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In Advances in neural information processing systems, pages 773–781.

— Bercu, B., Godichon-Baggioni, A., and Portier, B. (2019). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. SIAM Journal on Control and Optimization, 58(1), 348-367.

— Bottou, L. Curtis, K., Nocedal, J. (2018) : Optimization Methods for Large-Scale Machine Learning, Siam Reviews, 60(2) :223-311, 2018.

— Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. (2016). A stochastic quasi-newton method for large-scale optimization. SIAM Journal on Optimization, 26(2) :1008–1031.

— Cénac, P., Godichon-Baggioni, A., and Portier, B. (2020). An efficient averaged stochastic Gauss-Newton algorithm for estimating parameters of non linear regressions models. arXiv preprint arXiv :2006.12920.

— Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7).

— Duflo, M. (1997). Random iterative models, volume 34 of Applications of Mathematics (New York). Springer-Verlag, Berlin. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

— Gadat, S. and Panloup, F. (2019). Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. arXiv preprint arXiv :1709.03342.

— Godichon-Baggioni, A. (2019). Lp and almost sure rates of convergence of averaged stochastic gradient algorithms : locally strongly convex objective. ESAIM : Probability and Statistics, 23, 841-873.

— Godichon-Baggioni, A., and Saadane, S. (2020). On the rates of convergence of parallelized averaged stochastic gradient algorithms. Statistics, 54(3), 618-635.

— Leluc, R. and Portier, F. (2020). Towards asymptotic optimality with conditioned stochastic gradient descent. arXiv preprint arXiv :2006.02745.

— Patel, K. K. and Dieuleveut, A. (2019). Communication trade-offs for synchronized distributed SGD with large step size. arXiv preprint arXiv :1904.11325.

— Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. Sto-

chastic processes and their applications, 78(2) :217–244.
— Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. SIAM J. Control Optim., 39(1) :49–72.
— Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. SIAM J. Control and Optimization, 30 :838–855.
— Robbins, H. and Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, pages 400–407.
— Zeiler, M. D. (2012). Adadelta : an adaptive learning rate method. arXiv preprint arXiv :1212.5701.