Parametric complexity analysis for a class of first-order Adagrad-like algorithms

- Serge Gratton (Université de Toulouse, INP, IRIT, Toulouse, France)
- **Sadok Jerad** (Université de Toulouse, INP, IRIT, Toulouse, France)
- Philippe L. Toint (NAXYS, University of Namur, Namur, Belgium)

Keywords: Stochastic Optimization, Non-Convex Optimization, Algorithmic Complexity Analysis, Deep Learning.

Abstract: Methods where the objective function or a noisy proxy are not evaluated such as Adagrad [3] or Adam [4] have become very popular in the context of finite-sum minimization, where noise in the evaluation arises from sampling among a very large number of terms. That such methods can be provably convergent to first-order stationary points is quite remarkable. Moreover, several authors have been able to prove global convergence rates, including the recent contribution by [2], where an improved (compared to earlier analysis) such rate was proved for the Adagrad algorithm. We build on the results of [2] and propose two classes of algorithms for optimization in the presence of noise that do not require the evaluation of the objective function. We achieve several goals:

1. The global rate of convergence result of [2] is extended to a parametric class of methods comprising the Adagrad algorithm. These rates give the best complexity to the Adagrad among all algorithms in the class.
2. An improved asymptotic rate is also derived for these methods under an additional conditional variance condition, indicating that the results of [2] cannot be sharp if this condition holds. In this case, the parameter choice yielding the best bounds no longer corresponds to Adagrad and the improvement is asymptotic and implicit. This new complexity bound is independent from gradient "sparsity" and allows an essentially arbitrary choice of the learning rate. It therefore provides an alternative to that of [5].
3. A new class of methods is proposed in a stochastic and non-convex settings. This new class is reminiscent of the "divergent stepsize" subgradient method for non-smooth convex optimization (see [1] and references therein). Under the additional conditional variance condition, its global rate of convergence is shown to be very close to that of methods using (exact) function evaluations.

Numerical illustrations of the discussed methods on simple examples in the Deep Learning setting indicate that methods of the second class have some merits, but also that, at least in our examples, there remains some distance from the above theory to real behaviour.

References:

[1] Beck, A. *First-order Methods in Optimization*. Number 25 in MOS-SIAM Optimization Series. SIAM, Philadelphia, USA, 2017.

[2] Défossez, A., Bottou, L., Bach, F., and Usunier, N. A simple convergence proof for Adam and Adagrad. arXiv:2003.02395v2, 2020.

[3] Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, July 2011.

[4] Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings in the International Conference on Learning Representations (ICLR)*, 2015.

[5] Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. In *Proceedings of OPT2020: 12th Annual Workshop on Optimization for Machine Learning*, 2020.