Memory–Gradient Trade-offs in Large-Scale Finite-Sum Optimization

Master internship proposal

This master internship revolves around stochastic optimization with application to machine learning. The goal is to design new optimization methods, implement them numerically and prove their performance theoretically.

Research question

Training machine learning models is usually done by solving large-scale optimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x),$$

where N corresponds to the number of training samples (usually millions, $N \geq 10^6$) and $x \in \mathbb{R}^d$ (often $d \geq 10^6$) is the parameter of the model. Algorithms that remain practical in this setting rely on first-order stochastic approximations of ∇f by computing only a few gradient terms ∇f_i per iteration. This is the mechanism behind stochastic gradient descent (SGD) or popular variants like ADAM. While sometimes efficient, these algorithms often suffer from oscillations or do not converge exactly to a solution of the problem, when used with fixed step-sizes [3]. This is fixed by variance reduction (VR) techniques [2, 4] that reduce oscillations by storing previously computed stochastic gradient terms or by computing additional ones.

Despite the notable gain that VR provides in several applications, those methods are not broadly used, and in particular not popular for training neural networks, mostly because they require an additional cost proportional to $N \times d$ either in storage [2] or computation [4]. This makes them impractical for large-scale applications. For large-scale problems, there exist VR approaches that store only one or a few extra stochastic gradients [1], but the performance is sometimes far behind that of expensive VR techniques (when those are applicable).

The goal of this master internship is to theoretically characterize the memory/compute trade-off in VR techniques and propose new methods that find the best compromise between the two limitations. More precisely, the intern will:

- analyze the memory/compute trade-off of existing VR methods,
- design new methods with an intermediate memory budget,
- prove their convergence, starting with the convex setting,
- implement them and test them numerically, including for training neural networks.

Candidate profile

The candidate is expected to have a background in applied mathematics, with interests in optimization and machine learning. The recruited intern is expected to also be familiar with basic concepts of deep learning and Python programming, notably with `numpy` and/or `pytorch` modules.

Practical information

Supervision The internship will be jointly supervised by Camille Castera (camille.castera at math.u-bordeaux.fr) and Jean-François Aujol (jean-francois.aujol at math.u-bordeaux.fr) and will be hosted at the Institut de Mathématiques de Bordeaux (Talence, France).

Date The approximate starting date is March 2027, and will last approximately 5 month. The date and duration are flexible.

PhD opportunity Continuing with a PhD thesis will be considered if the successful candidate and the supervisors jointly agree.

Application process Please contact the supervisors directly by e-mail. Attach a resume and write a few lines describing your interest and qualifications for the position.

References

[1] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.10018.

[2] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://arxiv.org/abs/1407.0202.

[3] Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods, 2023. URL https://arxiv.org/abs/2301.11235. arXiv:2301.11235.

[4] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Abstract.html.