Internship Proposal: "Fast Algorithmic Methods for Optimization and Learning"

Samir Adly and Olivier Prot
Department of Mathematics, University of Limoges
Email: samir.adly@unilim.fr and olivier.prot@unilim.fr
http://www.unilim.fr/pages_perso/samir.adly/
https://www.unilim.fr/pages_perso/olivier.prot/

1 Context of the internship

In a real Hilbert space $\mathcal{H}$, we will consider and compare accelerated gradient algorithms for solving the unconstrained optimization problem:

$$\min\{f(x) : x \in \mathcal{H}\}, \tag{1.1}$$

where $f : \mathcal{H} \to \mathbb{R}$ is a smooth function such that its gradient ∇f is L-Lipschitz continuous and $\operatorname{argmin} f \neq \emptyset$. Both convex and nonconvex cases will be considered.

• **Historical Aspects.** For the past two decades, we have been experiencing a digital revolution where the data is at the center of the models. It concerns the extraction of relevant information from a large database of digital data using adapted machine learning algorithms. As the data is very large, we talk about "big data", which has as a direct consequence the formulation of models in high dimensional spaces. Mathematics, computer science and statistics have adapted to propose a new generation of algorithms capable of solving these problems in a reasonable time. Machine learning, has been developed recently and is used in many fields with concrete applications. Supervised learning algorithms learn from labeled training data, and help predict the outcome of unexpected data. Algorithms based on neural networks have gained popularity in recent times due to their efficiency boosted by the increase in computing capabilities of computers. In the various stages of data processing, the passage through the optimization of a function is necessary. This optimization represents the minimization of the errors, i.e. the difference between the studied data and the mathematical model describing these data. The Gradient Descent Method (GDM) is one of the most popular used to minimize a function, due to its simplicity. First-order methods have gained popularity in recent years due to their importance in solving large scale optimization problems in Machine Learning and Data Science by only having access to the gradient of the function. One of the drawbacks of the GDM is its slowness (zig-zag pattern convergence on quadratic functions). An improvement of the GDM was proposed in 1964 by B. Polyak where he considered a momentum term associated with a gradient descent step. The associated continuous surrogate is known as the heavy ball with friction (HBF):

$$\text{(HBF)} \qquad \ddot{x}(t) + \alpha\dot{x}(t) + \nabla f(x(t)) = 0. \tag{1.2}$$

This is an inertial system with a fixed viscous damping coefficient $\alpha > 0$. From a mechanical point of view, it could be interpreted as the motion of a material point subject to viscous friction

damping and conservative potential forces. The (HBF) is a second order (in time) dissipative system where the presence of inertia allows the system to overcome some known drawbacks of the (GDM) and acts to accelerate the convergence. We note that the (HBF) is not a descent method and the convergence of the trajectories towards a critical point of the potential to be minimized is well-known under various assumptions like convexity or analyticity. For a strongly smooth convex function and a viscous damping coefficient judiciously chosen, (HBF) provides convergence at exponential rate. For a general convex function, the asymptotic convergence rate of (HBF) is $\mathcal{O}(\frac{1}{t})$ (in the worst case). This is however not better than the steepest descent.

An other momentum method was introduced by Nesterov in 1983, known in the literature as Nesterov Accelerated Gradient (NAG). To obtain a continuous surrogate of (NAG), a decisive step was taken by Su-Boyd-Candès [18] with the introduction of an Asymptotic Vanishing Damping (AVD) coefficient of the form $\frac{\alpha}{t}$, with $\alpha > 0$. For a general convex function f, it provides a continuous version of the NAG:

$$(\mathrm{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0. \tag{1.3}$$

For $\alpha \geq 3$, each trajectory $x(\cdot)$ of $(\mathrm{AVD})_\alpha$ satisfies the asymptotic convergence rate of the values $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(1/t^2\right)$ as $t \to +\infty$. The introduction of the Hessian-driven damping in [6] allows to damp the transversal oscillations which can occur with (HBF). Recent studies have been devoted to inertial dynamics that combines asymptotic vanishing damping with Hessian-driven damping. In fact, the corresponding algorithms involve a correcting term in the NAG which reduces the oscillatory aspects [8, 7, 18].

Due to its performance in wide range of nonconvex problems in machine learning, the stochastic gradient descent (SGD) has attracted many researchers nowadays. The objective function to be minimized is of the form of a stochastic programming: (SP) $f(x) = \mathbb{E}_\xi f(x, \xi)$ (where the expectation is taken over the randomness ξ). A typical example is given where the loss function is averaged over n data points. The SGD with momentum and NAG with noise could be studied from the perspective of a continuous Stochastic Differential Equation (SDE) (see the recent work [15]). Adding to the continuous SDE a Hessian-driven damping and possibly dry friction could be of great interest to propose new stable and robust stochastic algorithms.

Our approach in the current project follows the sequence of works in the literature on the link between continuous dynamical systems (both deterministic and stochastic) and its interpretation for the design of fast and efficient first-order optimization algorithms.

• **Scientific context.** This internship is based on the close links between dissipative dynamical systems and the associated optimization algorithms obtained by temporal discretization. In recent years, an in-depth study was carried out linking the NAG method to inertial dynamics with an AVD coefficient. Recently a decisive step was obtained to improve (NAG) which consists in introducing into the dynamics a damping which is controlled by the Hessian of the function to be minimized. This geometric damping has a drastic effect on the attenuation of oscillations, which is beneficial from the point of view of optimization. Precisely, our study will be based on the asymptotic behavior, when the time t tends to infinity, of the following general damped inertial differential systems (with or without dry friction).

2 Objectives of the internship

The objectives of this internship are as follows:

- To carry out a state of the art and a quick bibliographic study to identify recent research in this field.
- Code the algorithms and compare them with others in the literature to measure their efficiency, robustness and speed of convergence. The programming language that will be used is either Matlab or Python.
- Test these numerical optimization algorithms on examples from the fields of machine learning, deep learning or signal /image processing.

3 Research profile of the candidate

Student in Master 2 Applied Mathematics, with in particular :

- Basic knowledge in the field of convex analysis and optimization.
- Basic knowledge in the field of machine learning.
- Basic knowledge of computer programming (Matlab/python).

4 Duration of the internship and practical aspects

The duration of this internship is 4 to 5 months, starting in February 2022. It will take place at the laboratory XLIM of the University of Limoges (123, avenue Albert Thomas - 87060 LIMOGES CEDEX).
This internship may lead to a PhD thesis in September 2022.

References

[1] S. Adly, H. Attouch, *First-order inertial algorithms involving dry friction damping*, Math. Prog. Ser. A (2021).

[2] S. Adly, H. Attouch, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30(3) (2020), 2134–2162.

[3] S. Adly, H. Attouch, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, J. Conv. Anal., 28 (2) (2021), pp. 281–310.

[4] S. Adly, H. Attouch, M.H. Le, *First-order inertial optimization algorithms with threshold effects associated with dry friction*, Available in HAL CNRS (2021).

[5] F. Alvarez, H. Attouch, *The heavy ball with friction dynamical system for convex constrained minimization problems*, Optimization (Namur, 1998), Lecture Notes in Econom. and Math. Systems, Springer-Verlag, 481(1-2) (2000), 25–35.

[6] F. Álvarez, H. Attouch, J. Bolte, P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics*, J. Math. Pures Appl., 81(8) (2002), 747–779.

[7] H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First-order optimization algorithms via inertial systems with Hessian driven damping*, *Math. Program.*, (2020) https://doi.org/10.1007/s10107-020-01591-1, preprint available at hal-02193846.

[8] H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), pp. 5734–5783.

[9] H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).

[10] A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), No. 1, pp. 183–202.

[11] Y. Nesterov, *A method of solving a convex programming problem with convergence rate O(1/k2)*, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.

[12] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[13] B.T. Polyak, *Some methods of speeding up the convergence of iterative methods*, Z. Vylist Math. Fiz., 4 (1964), pp. 1–17.

[14] B.T. Polyak, *Introduction to optimization*. New York: Optimization Software. (1987).

[15] B. Shi, W. J. Su, M. I. Jordan, , *On learning rates and Schrödinger Operators* , arXiv:2004.06977v1 (2020).

[16] B. Shi, S. S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program, (2021), https://doi.org/10.1007/s10107-021-01681-8, preprint available at arXiv:submit/2440124[cs.LG] Oct 2018.

[17] W. J. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*. Neural Information Processing Systems 27 (2014), 2510–2518.

[18] W. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method*, Journal of Machine Learning Research, 17 (2016), pp. 1–43.