An Inertial Newton Algorithm for Deep Learning with Convergence Guarantees

Camille Castera[1,*] Jérôme Bolte[2] Cédric Févotte[1] Edouard Pauwels[1,3]

[1]IRIT, Université de Toulouse, CNRS, France, *corresponding author: camille.castera@irit.fr
[2]Toulouse School of Economics, Université Toulouse 1 Capitole, France
[3]DEEL, IRT Saint Exupery, Toulouse, France

Keywords: Nonsmooth Nonconvex Stochastic Optimization, Deep Learning, Dynamical Systems

We introduce a new second-order inertial method for machine learning called INDIAN [1], exploiting the geometry of the loss function while requiring only stochastic approximations function values and generalized gradients. This makes the method fully implementable and adapted to large scale optimization problems such as the training of a deep neural network.

The algorithm combines both gradient-descent and Newton features as well as inertia. It is a discretized and split version of a dynamical system introduced in [2]. We provide a strong meaning to each hyperparameter of the model by making a connection to Newton's second law.

We prove the convergence of INDIAN to critical points for almost any *classical* deep learning problems. To do so, we provide a well suited framework to analyze deep learning losses, involving tame optimization and Clarke subdifferential [3, 4]. In this framework we provide a step by step proof recipe combining continuous dynamical system analysis together with discrete stochastic approximations in the lines of [5].

From an empirical point of view the algorithm shows promising results on popular benchmark problems, as well as some appealing generalization properties.

References

[1] Camille Castera, Jérôme Bolte, Cédric Févotte and Edouard Pauwels, An Inertial Newton Algorithm for Deep Learning, *arXiv preprint*, arXiv:1905.12278, 2019.

[2] Felipe Alvarez, Hédy Attouch, Jérôme Bolte and Patrick Redont, A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics, *Journal de Mathématiques Pures et Appliquées*, 81(8):747-779, 2002.

[3] Frank H. Clarke, Optimization and nonsmooth analysis, SIAM, 1990.

[4] Jérôme Bolte, Aris Daniilidis, Adrian Lewis and , Masahiro Shiota, Clarke subgradients of stratifiable functions *SIAM Journal on Optimization*, 18(2)556–572, 2007.

[5] Michel Benaïm, Josef Hofbauer and Sylvain Sorin, Stochastic approximations and differential inclusions, *SIAM Journal on Control and Optimization*, 44(1):328-348, 2005.