MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

Deterministic optimisation in Deep Learning from a continuous and energetic point of view

B. Bensaid
Université de Bordeaux

Supervisor(s): R.Turpault (Bordeaux-INP), G.Poette (CEA-CESTA)

PhD expected duration: Oct. 2021 - Oct. 2024

Address: 351 Cr de la Libération, 33400 Talence

E-mail: bilel.bensaid@u-bordeaux.fr

Abstract:

Deep learning models have recently been used to substitute parts of simulation codes with neural networks [3]. In this context, huge networks do not seem appropriate from a computational point of view. With this in mind, this work aims at building efficient shallow networks. The guiding principle of this research starts from this assessment: some recent papers [4] point the optimizer out as a lever to improve performances. There are stochastic and deterministic optimizers. Before analysing stochastic algorithms it is imperative to deeper understand the deterministic ones (as stochastic optimizers are often based on deterministic ones).

The first contribution concerns a validation methodology. Usually to green a new optimizer light it is tested on machine learning benchmarks: MNIST, CIFAR-10,... In such a case, there is no clue about critical points localization and their values. As a result, the optimizer may give a saddle point with a low value and without any further information the algorithm is said to be satisfactory. An original and more thorough way of testing is suggested: build small networks which make it possible to compute exactly the critical points and their nature. Several examples are derived changing the asymptotic properties (singular hessian, steep gradients etc.) around the minimums.

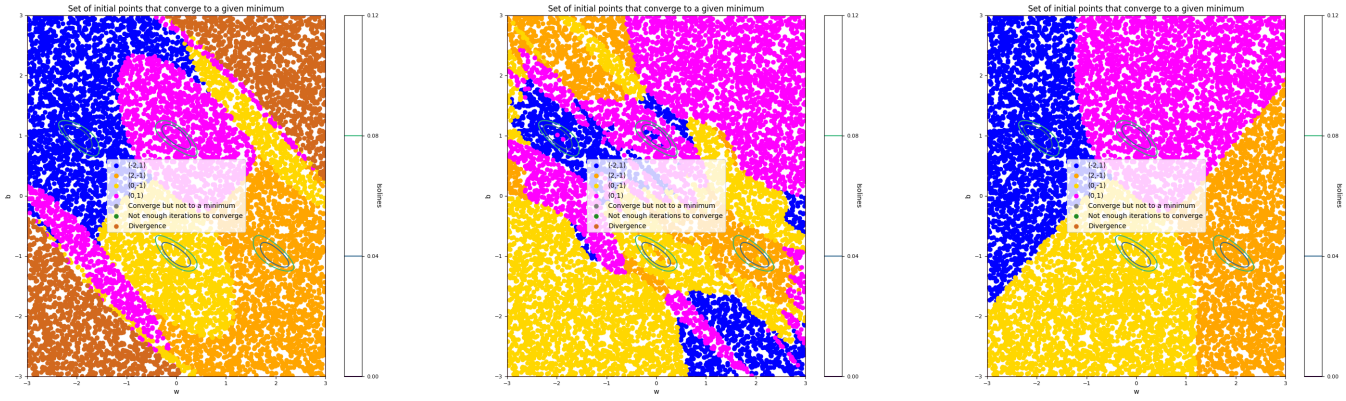
The most used deterministic optimizers are applied on these examples: Gradient Descent(GD), Momentum, Adam, Levenberg-Marquardt(LM) [2] and their variants. Two aspects are interesting:

1. A statistical study on the probability of finding a given minimum is made. It is remarkable that the usual algorithms of order one do not get any maximums or saddle points without adding stochastic perturbations;
2. For a given minimum θ^*, it is worthy to consider the set of all initial points that converge toward θ^*. The display of these regions reveals some surprising behaviors 1b. The most impressive one is the significant initial conditions sensitivity of Adam algorithm. The situation is even worse: starting very close to a minimum the optimizer may converge to another one. This is obviously a non-desirable property.

To explain what is observed, continuous versions of the optimizers - that is to say differential equations - are derived so that a simple discretization of these ODEs (Euler most of the time) gives back the classic algorithms. In this framework, Lyapunov stability plays an essential role:

1. Maximums and saddle points are unstable so the trajectory can always escape their neighbourhoods [1];
2. Concerning Momentum and Adam, the form of Lyapunov functions indicates that even close to a minimum the potential energy is not necessary zero. So an update is still possible although the algorithm should stop.

This framework opens a large set of possibilities about the discretization scheme to use:



(a) GD (b) Adam (c) Euler Richardson modified GD

Figure 1: Sensitivity to the initial conditions

1. Adaptive schemes like Euler-Richardson are relevant in this field. Indeed, in practice a huge amount of time is spent to tune the hyperparameters: width, depth, activation functions, learning rate, coefficients of the moving average,... These new schemes enable the practitioner to get rid of all the hyperparameters involved in the optimization. As well as save time, this kind of algorithm admits convex regions 1c.
2. The Lyapunov theory leads quite naturally to energetic considerations. Starting from the ODE, some quantities that should decrease over time are identified. New adaptive strategies for the learning rate can be designed by requiring that the discrete equivalents of these quantities also decrease. This ensures convergence whereas a lot of trajectories diverge for the canonical versions (see the large amount of brown diverging trajectories for GD 1a whereas its corrected version 1c remains stable in the whole domain).

In this work, benchmarks more suitable for a meticulous analysis of deep learning optimisation are built. The observed behaviours lead to identify some valuable mathematical properties. This brings to original ways for adapting optimizers' parameters with a moderate sensitivity to the initializer.

References

[1] N. G. (Nikolaĭ Gurevich) Chetaev. *The stability of motion.* Pergamon Press, New York, [2d rev. ed.] translated from the russian, by morton nadler. translation editors: a. w. babister [and] j. burlak. edition, 1961.

[2] M.T. Hagan and M.B. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994.

[3] G. Kluth, K. D. Humbird, B. K. Spears, J. L. Peterson, H. A. Scott, M. V. Patel, J. Koning, M. Marinak, L. Divol, and C. V. Young. Deep learning for nlte spectral opacities. *Physics of Plasmas*, 27(5):052707, 2020.

[4] Paul Novello, Gaël Poëtte, David Lugato, and Pietro M Congedo. Explainable Hyperparameters Optimization using Hilbert-Schmidt Independence Criterion. working paper or preprint, June 2021.

Short biography – Engineer student from Mines Saint-Etienne specialized in Data Science I am also graduated of a master degree in Applied Mathematics (MAEA from Université Lyon Bernard). My thesis at Institut de Mathématiques de Bordeaux is financed by CEA-CESTA within the scope of LRC-Anabase.