Out-of-equilibrium sampling as a protocol to unveil locally-smooth neural network configurations in non-convex optimization

Clarissa Lauditi

Supervisors
Prof. Riccardo Zecchina
Asst. Prof. Enrico M. Malatesta

A Dissertation
submitted to the Faculty of Graduate Studies
in partial fulfillment of the requirements
for the Degree of
Doctor of Philosophy

Graduate Program in Physics
Politecnico of Turin University
Turin, Italy

Introduction

In this Doctoral thesis we address the analytical and computational problem of learning with neural networks by making use of static approaches and bias techniques on sampling arising from the toolbox of statistical physics. The main topic, which is to clarify what prevents algorithms from finding solutions to certain non-convex optimization problems, will be devoted to a perspective all about the geometric characterization of solutions in the optimization landscape. The following work is divided into three parts.

- Part I: Preliminaries. The first part aims to introduce the reader to the midst of literature and debates about neural networks (NNs) emerging behaviours, clarifying that many of the variables chosen when training NNs can be studied in isolation to quantify their impact on the learning dynamics.
    - ◇ After formalizing the differences between supervised and unsupervised tasks (1.1), Sec. 1.2 provides an overview on some of the regimes for which supervised learning in a non-convex, high-dimensional optimization landscape can be facilitated: the role of initialization for the learning speed (1.2.1), the impact of overparameterization in locally convexifying the landscape (1.2.1), and a rigorous definition of algorithmic hardness from a geometrical perspective (1.2.3).
    - ◇ Chapter 2 stands with the literature supporting high local entropy solutions as the ones attractive for the non-equilibrium dynamics of local search algorithms, from which all the protocols derived in the following parts, and the conclusions drawn, take their inspiration. Section 2.1 reports the case study of a minimal neural network model for which the optimization is high-dimensional and non-convex: the Perceptron with binary weights. The goal is to explain, using approaches borrowed from statistical physics of disordered systems, how it is possible in the typical case to characterize learning properties of simple models and how, depending on the probability measure from which solutions are sampled, one does or does not derive meaningful theoretical predictions compared to algorithmic results. Section 2.1.2 proposes an adequate measure to sample from the landscape attractive solutions for the algorithm dynamics; starting from that, section 2.2 illustrates how it is possible to optimize solvers to target high local entropy (i.e. dense clusters) configurations due to the correlation between flatness and performance of the corresponding minimizer.

- Part II: the benefits of robustness. Once it has been clarified that the non-trivial landscape where the learning process occurs is characterized by minima associated with different performance, and specifically that flat solutions are not only dynamically attractive but also robust to perturbations and performant for generalization, the second part of the manuscript focuses on introducing protocols to observe the structural change of these relevant configurations as the complexity of the learning task increases.

  - ◇ In Chapter 3 and for two models of networks with binary weights, the Perceptron and the Tree Committee machine, minima on randomly generated supervised tasks are characterized as configurations with a certain level of robustness (i.e., margin) and the first results on hierarchical organization of clusters as solutions arising from the coalescence of robust configurations is derived (sec.s 3.1 and 3.2). Then, the analysis specializes on an analytical method, entirely based on local geometry of minimizers and independent from the specific algorithmic strategy, to identify the constraint density threshold (i.e. proportional to the complexity of the tasks and the density of patterns in the dataset) that causes clusters to break and prevents algorithms from converging (3.1.2).

  - ◇ Chapter 4 adapts all the techniques in 3 to a mismatched teacher-student problem where the correlation among patterns visible to the student comes from a low-dimensional manifold of data. In Sec. 4.1, which defines the models, all the assumptions, which allows for the use of the Gaussian Equivalence Principle (4.1.1), are introduced in order to study the typical case student performance. In following sections, the geometry of minima of the learning tasks are studied by inspecting the fallout of overparameterization, keeping track of possible phase transitions when changing the expressivity of the student compared to the teacher model. As expected, prediction performances of the student get better when increasing the overparameterization, and a clear relationship between the robustness of a minimizer and the associated generalization is reported (4.2.2). As in 3, it is possible to derive a threshold for algorithmic performance degradation with the disappearance of robust solutions from the landscape, showing how efficient solvers cannot overcome the conjectured hard phase (4.2.3). The chapter ends with a digression on deeper models, supporting evidences for which, despite the overparameterization of bigger architectures, the optimization landscape still remains rough and non-convex, being the scope that of finding minima with an average non-trivial overlap among them (4.3).

  - ◇ Chapter 5 reports the first analytical study on the energetic barriers along the geodesic path connecting minimizers sampled independently from the equilibrium measure with different degrees of robustness. The model is a non-convex one-layer NN with continuous weights, i.e. the spherical negative Perceptron (5.1), and the goal of studying the training error on the

convex envelope spanned by y different solutions is that of deriving algorithmic implications of simple connectivity properties (5.5) and to characterize the geometry of the space of minima (5.3) more in detail with respect to 3 and 4. The existence in theory of a large geodesically convex component for sufficiently high degree of overparameterization and among solutions with high margin (5.3) reveals how it is possible that convex optimization strategies find in practice linearly-mode-connected solutions, and give rise to a star-shaped organization of the zero-energy manifold, where for sufficiently low density of constraints, even dominant configurations of a flat probability measure are linearly connected to a convex cluster of minima.

  - ◇ Chapter 6 refers to a work in preparation that uses hierarchical space organization of solutions with different margins to think of a practical application of averaging methods in non-convex optimization landscapes. The sections, each of them devoted to report evidences of how it is possible to play with the landscape structure to take advantage of negative margin solutions, explain how sometimes, given an optimization problem with some level of noise in the dataset, it is worthy to relax some constraints in the learning process, since averaging over under-performing model has the effect of targeting a regularized robust and well-performing barycenter in the end. Section 6.3 extends the analysis to a simple analytical proof that the generalization error along the geodesic between negative margin solutions shows a minimum in the middle of the path, which one has reason to think corresponds to an atypical robust solution.

- Part III: neuroscience comes into play. The present part of the manuscript talks about two simple examples of biological models to address questions in neuroscience through the use of optimization-like strategies. What is the memory capacity of a single neuron of a pyramidal cell (7)? And what if one can construct an associative memory model that, instead of simply storing prototypes i.i.d. patterns, plays with correlated inputs (8)?

  - ◇ Chapter 7 is a work in preparation that proposes use of a Tree Committee Machine with positive weights to model the activation scheme of an excitatory pyramidal cell that includes dendritic non-linearities. For the model, both theoretical critical capacity (7.2.1) and maximal algorithmic storage load (7.2.2) are derived, discussing how the simple addition of a non-linear activity layer makes the model faster to learn, more robust to noise, and with better prediction performances (7.4) with respect to its linear counterpart.
  - ◇ Chapter 8 is an extension of the Hopfield model with correlated patterns, i.e. the Random Feature Hopfield. Compared to the standard version, when trying to use the Hebb rule for the coupling synaptic matrix to store as prototypes not i.i.d. but correlated inputs, different phases appears in the usual phase diagram: a learning phase where the features generating

the inputs become attractive for the zero-temperature asynchronous dynamics (8.1), and a generalization phase (8.2) where mixtures of features let the model have new fixed point attractors related to patterns that it has never seen before.

- Part IV: conclusions. In this final Chapter we will discuss the results obtained and we will highlight several possible directions for future works.