On Convex Neural Networks

Mathematical Foundations of Data Science

Eloïse Berthier

November 30, 2018

1/2 page: What problem(s) is studied ? Why is it relevant ? What solution(s) is proposed ? Which contributions (theory, numerics, etc) ?

1 Introduction

While the use of neural networks has dramatically increased the performance of some recognition systems, there still lacks a clear mathematical understanding of this success. One of the key issues is that optimizing neural networks is a non-convex problem, hence training algorithms may not return a global minimum.

Yet in practice, training large neural networks often results in satisfying solutions, which are empirically considered *not too far* from a global minimum. Thus it suggests that there may be some implicit assumptions on the data or on the model that are usually verified and could explain this regularity, somehow making the problem *more convex* than at first sight.

There have been several attempts to study this phenomenon, heading in different directions.

Some rely on convex relaxations of the original optimization problem. [Zhang et al., 2016] defines a convex relaxation for training convolutional neural networks. They show that in the case of a two-layer CNN, the generalization error of the solution to the convex relaxed problem converges to that of the best possible two-layer CNN.

In [Haeffele and Vidal, 2017] are studied conditions under which the optimization landscape for the non-convex optimization problem is such that all critical points are either global minimizers or saddle points. [Li et al., 2018] provides a tool to visualize the loss surface along random directions near the optimal parameters.

There is also a series of some very recent contributions on the problem: [Du et al., 2018b], [Du et al., 2018a] and [Zou et al., 2018]. They rely on the analysis of the dynamics of the prediction space rather than of the parameter space, the former being governed by the spectral property of a Gram matrix. If the data is *not degenerate*, this Gram matrix's least eigenvalue is lower bounded, and randomly initialized gradient descent linearly converges to a global optimum.

In this report, we concentrate on another approach called *convex neural networks*. The core idea was introduced by [Bengio et al., 2006] and focuses on one hidden-layer neural networks. Instead of optimizing the weights of a network with a fixed number of hidden units, we could consider the equivalent problem of optimizing on the set of all the possible hidden unit functions. The training problem is convex with respect to the parameters of the model (which are now only the ouput weights).

Of course, while this approach solves the non-convexity issue, it introduces some other difficulties. These are extensively studied in [Bach, 2017], on which we will mainly focus in this report. A complete analysis of the generalization performance of convex neural networks is derived, in particular

showing that high-dimensional non-linear variable selection may be achieved, without any strong assumption regarding the data. Yet this approach requires to solve a convex problem in infinite dimension and is only possible if the non-convex subproblem of adding a new hidden unit can be solved efficiently. Finding a polynomial-time algorithm solving this subproblem induced by the conditional gradient method is still an open question.

[Chizat and Bach, 2018] focuses on solving the same convex problem, but avoids the potentially NP-hard problem of optimizing with conditional gradient. Instead, they discretize the unknown measure as a mixture of particles. This approach is inspired by optimal transport and brings strong asymptotical global optimality garantees for gradient flow optimization. It also provides simple numerical experiments.

Looking at all these different approaches, we may notice some recurrent patterns in the assumptions that are commonly made:

- *one hidden-layer models*: most of the articles focus on simple neural network models, that already include a wide variety of classical supervised learning configurations, but do not directly cover deep neural networks.
- *over parametrization*: all the good generalization properties are derived in the case of an over parametrized model, yet the key question is *how much* over parametrization is needed.
- *positive homogeneity*: most of the results are achieved in the case of positively homogeneous activation functions. It could explain the experimental advantage of using ReLU functions over others in deep learning.
- *proper initialization*: several articles rely on the conservation of an initial regularity through the optimizing process, for instance using gradient flows, and so emphasize the role of the initialization.

2 Defining a Convex Problem

2.1 General layout

We consider single hidden-layer neural networks. It defines a class of prediction functions on $\mathbb{R}^d$ parametrized as: $f(x) = \sum_{j=1}^{k} \eta_j \sigma(w_j^T x + b_j)$, where σ is a fixed activation function, k is the number of hidden units, and the $(\eta_j)_{j=1,\dots,k}$ are the weights of the output layer.

sigma positive homogeneous

2.2 From non-convexity to convexity

For a fixed value of k, training a neural network is an empirical risk minimization problem: $\min_{\eta,w,b} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \Omega(\eta, w, b)$, where ℓ is a smooth convex loss function and Ω a convex regularization function.
–> non convex problem

the convex problem
Penalization L1 / L2 sampling.
number of hidden units
sample complexity / curse of dimensionality

3 Theoretical Analysis

3.1 Errors decomposition

3 errors decomposition. Speak about the last two

3.2 Approxmiation error ?

3.3 Generalization error

4 An Optimization Problem

4.1 Frank Wolfe

4.2 NP hard

5 The Particle Gradient Flow Approach

A more general result, can be applied to this problem

6 Numerical Experiments

Main body (10 pages) : Presentation of the method(s). Theoretical guarantees. Numerics.

7 Conclusion and perspectives

Conclusion and perspective (1 page) Summary of the result obtained: pros and cons (limitation, problems, error in the articles, etc) Possible improvement/extension

References

[Bach, 2017] Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53.

[Bengio et al., 2006] Bengio, Y., Roux, N. L., Vincent, P., Delalleau, O., and Marcotte, P. (2006). Convex neural networks. In *Advances in neural information processing systems*, pages 123–130.

[Chizat and Bach, 2018] Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*.

[Du et al., 2018a] Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

[Du et al., 2018b] Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

[Haeffele and Vidal, 2017] Haeffele, B. D. and Vidal, R. (2017). Global optimality in neural network training.

[Li et al., 2018] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*.

[Zhang et al., 2016] Zhang, Y., Liang, P., and Wainwright, M. J. (2016). Convexified convolutional neural networks. *arXiv preprint arXiv:1609.01000*.

[Zou et al., 2018] Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*.

(a) Average return during the gradient ascent (b) Mean parameter during the gradient ascent