Report for:
Exact full-RSB SAT/UNSAT transition in infinitely wide two-layer neural networks

Summary of the manuscript

This work considers the problem of how many Gaussian random patterns $\xi^\mu \sim \mathcal{N}(0, I_N)$, $\mu \in [P]$ a two-layer neural network can asymptotically classify within a margin $\kappa \in \mathbb{R}$ when the corresponding binary labels are random $y^\mu \sim \mathrm{Rad}(1/2)$ (a.k.a. the *storage capacity problem*). The analysis focus on a particular two-layer architecture with fixed second layer weights and K hidden-units that partition the input data ξ^μ in K disjoint patches ξ_l^μ of size N/K, a.k.a. as the *tree committee machine*:

$$\hat{y} = \mathrm{sign}\left(\frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l \varphi\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} w_{li}\xi_{li}\right)\right) \tag{1}$$

where $c_l \in \{-1, +1\}$ denote the (frozen) second-layer weights, φ a (generic) activation function and $w_{li} \in \mathbb{R}$ the first-layer weights, which are also assumed to be normalized $\sum_{li} w_{li}^2 = N/K$. This architecture is particularly amenable to a theoretical treatment since the K pre-activations are independent, making the analysis close to the single layer-case $K = 1$ (a.k.a. *the perceptron*). Nevertheless, different from the latter, in full generality ($K \geq 1$, $\varphi \neq \mathrm{id}$, $\kappa \in \mathbb{R}$), this problem is non-convex on the weights w_{li}, which is what makes the analysis challenging. From the statistical physics perspective, the lack of convexity translates into a potential lack of replica symmetry (RS), requiring a replica symmetry breaking treatment (RSB) of the corresponding Gardner volume of solutions which correctly classify the random labels.

Previous literature on this problem has mostly focused in the RS and 1-RSB approximations of the Gardner volume. The main contribution of this work is to carry out a full-RSB analysis, and to precisely characterize the transitions between the different levels of RSB as a function of the sample complexity $\alpha = P/N$ and margin κ. This allows to exactly compute the capacity threshold $\alpha_c(\kappa)$ of maximum number of random patterns the network can correctly classify.

Comments, questions and suggestions

1. Page 3 in the Introduction:

   > *For continuous weight models instead, the picture is not as clear: the same tools used for binary models provide an algorithmic threshold that can be easily overcomed by simple algorithms [19].*

   I understand what you mean, but I would not call an "algorithmic threshold" a threshold which can be "easily overcomed".

2. Bottom of Page 5:

   > *Interestingly, this algorithm can be proven to reach capacity, provided that the typical states exhibit no overlap gap, i.e. the overlap distribution of typical states is with a compact support.*

Can you be more precise about your definition of nOG? To my knowledge, the standard definition of OGP (e.g. from [25]) is about about the disconnectedness of the support of the overlap distribution, rather than being compact or not (note that a finite set is compact).

3. Can you further elaborate on the difference between nOG and OGP? This is not very clear from the brief discussion in the end of Page 5. Giving an explicit example of a problem / overlap distribution which has nOG but no OGP would be useful.

4. The central limit theorem argument in Section 3.2 is a key part in the derivation, but I feel some important points are not properly discussed. For instance

   (a) You take the large-width $K \to \infty$ limit *before* the high-dimensional limit $N, P \to \infty$ at fixed $\alpha = P/N$. Unless these two limits commute, this means you effectively assume K grows faster than N. Is it clear these two limits commute here? In either case, this should be stressed both in this section and before, stating that all your capacity results hold strictly for the infinite width case.

   (b) How important is the $1/\sqrt{K}$ scaling in the second-layer for this argument? This seems to be closely related to the fact you get a NNGP in the following. I understand that $\hat{y}$ is invariant under this scaling due to the sign, but the stability Δ^μ upon which depends the loss function is not. For instance, would the argument hold in a "mean-field" scaling $1/K$? This should be discussed.

   (c) Can you be be more explicit in the passage from the first to the second equality in eq. (18)? I didn't find this in the Appendix, and would encourage you to detail it better there.

5. From your argument, it looks like the critical storage capacity α_c is independent of the width K (differently from the fully-connected committee machine). Is this right?

6. You refer to the "Gardner transition" several times in the text, but this is never explicitly defined.

7. Last sentence of the Conclusion:

   > *Finally, we compared our estimates of the SAT-UNSAT threshold with the performance of Gradient Descent. In all cases analyzed we have given evidence that Gradient Descent stops finding solutions before the exact SAT/UNSAT threshold that we computed, thus implying the presence of an algorithmic gap.*

   The wording of this sentence is too strong - the numerical evidence does not "imply" but "suggest" the presence of an algorithmic gap.

8. Bottom of Page 17:

   > *When training a tree committee machine we have empirically observed that the first method leads to a larger probability of finding a solution, while for the negative perceptron the second method works best.*

   This is curious. Do you have an intuition why? In particular, I am not sure the second method really makes sense in the context of the analysis. Indeed, the unconstrained risk function will be different from the constrained one — why would you expect the same algorithmic threshold?

9. Overall, I think the paper lacks a clear "take home message". The authors start by motivating that the current RS and 1RSB results provide only an upper bound of the storage capacity of these networks. But what do we learn from correcting a 10^{-2} digit in this constant (in the worst case discussed)? For instance, it would be nice to have a discussion on how the activation function impact the capacity. From Table 1, it seems it does not change much — is this always the case or one can build examples with larger gaps?

Small typos and suggestions

- Broken reference [?] in the top of Page 3.
- There is a φ too much between eqs. (1) and (2). From the calculation that follows, I think you meant no φ in eq. (2).
- For the sake of clarity, state that throughout the manuscript you assume that N/K is an integer.
- The notation Θ for the Heaviside step function in eq. (6) is standard in Physics, but not in CS. Since this work can also interest people in this community, I suggest the authors to define it explicitly or say it in words in/around eq. (6).
- Page 8, "semidefite" $\rightarrow$ "semidefinite".
- Sometimes you write "tree committee machine", sometimes "tree-committee machine" and sometimes "tree committee-machine".