JMLR: Workshop and Conference Proceedings vol 40:1–5, 2015 28th Annual Conference on Learning Theory

Open Problem: The landscape of the loss surfaces of multilayer networks

Anna Choromanska ACHOROMA@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University, New York

Yann LeCun YANN@CS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University, and Facebook Research, New York

Gérard Ben Arous BENAROUS@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University, New York

Editor: Under Review for COLT 2015

Abstract

Deep learning has enjoyed a resurgence of interest in the last few years for such applications as image and speech recognition, or natural language processing. The vast majority of practical applications of deep learning focus on supervised learning, where the supervised loss function is minimized using stochastic gradient descent. The properties of this highly non-convex loss function, such as its landscape and the behavior of critical points (maxima, minima, and saddle points), as well as the reason why large- and small-size networks achieve radically different practical performance, are however very poorly understood. It was only recently shown that new results in spin-glass theory potentially may provide an explanation for these problems by establishing a connection between the loss function of the neural networks and the Hamiltonian of the spherical spin-glass models. The connection between both models relies on a number of possibly unrealistic assumptions, yet the empirical evidence suggests that the connection may exist in real. The question we pose is whether it is possible to drop some of these assumptions to establish a stronger connection between both models.

Keywords: multilayer networks, deep learning, spherical spin-glass model, Hamiltonian, non-convex optimization

1. Introduction

The vast majority of practical applications of deep learning use multi-stage architectures composed of alternated layers of linear transformations and max functions (most often Rectified Linear Units, e.g. Nair and Hinton (2010)), and focus on supervised learning, where the loss function that needs to be minimized is most often cross entropy or hinge loss.

Several researchers experimenting with larger networks had noticed that, while multilayer nets do have many local minima, the result of multiple experiments consistently give very similar performance. This suggests that all those local minima are more or less equivalent in terms of error. It was also previously noticed that the problem of training deep learning systems resides with avoiding saddle points and quickly "breaking the symmetry" by picking sides of saddle points and choosing a suitable attractor LeCun et al. (1998); Saxe et al. (2014); Dauphin et al. (2014).

Earlier theoretical analysis, conveniently reviewed in Dauphin et al. (2014), suggest the existence of a certain structure of critical points of random Gaussian error functions on high dimensional continuous spaces. They imply that critical points whose error is much higher than the global minimum are exponentially likely to be saddle points with many negative and approximate plateau directions whereas all local minima are likely to have an error very close to that of the global minimum. Their work establishes a strong empirical connection between neural networks and the theory

of random Gaussian fields by providing experimental evidence that the cost function of neural networks exhibits the same properties as the Gaussian error functions on high dimensional continuous spaces. Nevertheless they provide no theoretical justification for the existence of this connection.

2. The connection between multilayer networks and spin-glass models

We next discuss the assumptions that were made in Choromanska et al. (2015) to establish a connection between the loss function of neural networks and the Hamiltonian of the spherical spin-glass models (for detailed explanations see Choromanska et al. (2015)). The assumptions are numbered and marked with letter resp. 'p' or 'u' denoting whether the assumption is resp. plausible, i.e. it can be satisfied in practice or else it can be imposed on the network without significantly changing its performance, or obviously unrealistic, e.g 'A1p' denotes the first assumption, which is plausible.

It can be shown that the loss function of a typical multilayer network with ReLUs can be expressed as a polynomial function of the weights in the network, whose degree is the number of layers, and whose number of monomials is the number of paths (denoted as Ψ) from inputs to outputs. As the weights (or the inputs) vary, some of the monomials are switched off and others become activated. Consider a simple model of a fully-connected feed-forward neural network with $H-1$ hidden layers (n_i denotes the number of units in the i^{th} hidden layer, where input layer has index $i=0$ and output layer has index $i=H$), and having a single output (consider binary classification problem). Let $\Lambda = \sqrt[H]{\Psi}$, and we assume $\Lambda \in \mathbb{Z}^+$. Let X_i be the random input of the i^{th} path of a network. Then the normalized output of the network can be expressed as

$$Y = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i=1}^{\Psi} X_i A_i \prod_{k=1}^{H} w_i^{(k)},$$

where $w_i^{(k)}$ is the weight of the k^{th} segment of the i^{th} path (this segment connects layer $(k-1)$ with layer k of the network), and A_i is a Bernoulli random variable denoting whether the i^{th} path is active ($A_i=1$) or not ($A_i=0$). Consider hinge loss $L(\mathbf{w}) = \max(0, 1-Y_t Y)$, where Y_t is a random variable corresponding to the true data labeling taking values 1 or -1, and $\mathbf{w}$ denotes all network weights. Recall that max operator is often modeled as Bernoulli random variable taking values 0 or 1. Denote this random variable as M and its expectation as $\rho^{'}$. Therefore

$$L(\mathbf{w}) = M(1-Y_t Y) = M + \frac{1}{\Lambda^{(H-1)/2}} \sum_{i=1}^{\Psi} Z_i I_i \prod_{k=1}^{H} w_i^{(k)}, \tag{1}$$

where $Z_i = -Y_t X_i$, and $I_i = M A_i$ is a Bernoulli random variable taking values 0 or 1. Assume random variables $I_1, I_2, \ldots, I_\Psi$ have the same probability of success (**A1p**), and thus they have the same expectation denoted as ρ. Also assuming that each X_i is a standard Gaussian random variable (**A2p**), it follows that Z_i is also a standard Gaussian random variable.

For large-size networks large number of network parameters are redundant Denil et al. (2013) and can either be learned from a very small set of unique parameters or not learned at all with almost no loss in prediction accuracy. Assume that Λ is the maximal number of non-redundant (unique) parameters (**A3p**), and that they are uniformly distributed on the graph of connections of the network (**A4p**), i.e. every H-length product of unique weights appears in Equation 1 (the set of all products is $\{w_{i_1} w_{i_2} \ldots w_{i_H}\}_{i_1,i_2,\ldots,i_H=1}^{\Lambda}$). Thus re-indexing the terms gives

$$L(\mathbf{w}) = M + \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1,i_2,\ldots,i_H=1}^{\Lambda} Z_{i_1,i_2,\ldots,i_H} I_{i_1,i_2,\ldots,i_H} w_{i_1} w_{i_2} \ldots w_{i_H}.$$

Assuming (**A5u**) the independence of $Z_{i_1,i_2,\ldots,i_H}$ and $I_{i_1,i_2,\ldots,i_H}$ one obtains

$$\mathbb{E}_{M,I_1,I_2,\ldots,I_\Psi}[L(\mathbf{w})] = \rho^{'} + \rho\frac{\mathbf{1}}{\mathbf{\Lambda^{(H-1)/2}}}\sum_{\mathbf{i_1},\mathbf{i_2},\ldots,\mathbf{i_H}=\mathbf{1}}^{\mathbf{\Lambda}}\mathbf{Z_{i_1,i_2,\ldots,i_H}w_{i_1}w_{i_2}\ldots w_{i_H}}.$$

It is also assumed that Z's are independent (**A6u**) . Finally, the spherical assumption (**A7p**) imposes that $\frac{1}{\Lambda}\sum_{i=1}^{\Lambda} w_i^2 = 1$.

Note that the term in bold is a Hamiltonian of the spherical spin-glass model Auffinger et al. (2010). It was recently shown Auffinger et al. (2010) that the Hamiltonian of this model has interesting properties when the size of the model (Λ) goes to ∞. We next list these properties along with the possible interpretation for neural networks: (i) critical points form an ordered structure such that there exists an *energy barrier* (a certain value of the Hamiltonian) below which with overwhelming probability one can find only low-index[1] critical points, most of which are concentrated close to the barrier (this would explain why in case of large networks recovered local minima are typically corresponding to the same test performance which is not the case for small networks, (ii) Recovering the ground state, i.e. global minimum, takes exponentially long time, (iii) with overwhelming probability one can find only high-index saddle points above energy $E_{-\infty}$ and there are exponentially many of those (this would explain the importance of saddle points in the optimization problem), (iv) low-index critical points are 'geometrically' lying closer to the ground state than high-index critical points (this would explain why recovering poor quality local minima, which are 'far' from the global minimum, is more likely for small-size networks than for large-size networks).

Open problem: Is it possible to establish a connection between the loss function of the neural networks and the Hamiltonian of the spherical spin-glass models under milder assumptions? The central problem is to eliminate unrealistic assumptions of variable independence (**A5-6u**). Note that assumption **A5u** implies that the activation mechanism of any path (for the i^{th} path it is denoted as I_i) is independent of the input data, which clearly cannot be true. Similarly, assumption **A6u** implies all paths have independent inputs, which cannot be true since many paths share the same input. Alternatively, it would also be desired to find network architectures for which the connection to spin-glass models can be established explicitly with only mild (plausible), if any, assumptions.

References

A. Auffinger, G. Ben Arous, and J. Cerny. Random matrices and complexity of spin glasses. *arXiv:1003.1129*, 2010.

A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

Y. Dauphin, R. Pascanu, Ç. Gülçehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*. 2014.

M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. D. Freitas. Predicting parameters in deep learning. In *NIPS*. 2013.

Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In *Neural Networks: Tricks of the trade*. Springer, 1998.

V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*. 2014.

1. Index of $\nabla^2 L$ at $\boldsymbol{w}$ is the number of negative eigenvalues of the Hessian $\nabla^2 L$ at $\boldsymbol{w}$. Local minima have index 0.

Appendix A. Empirical evidence

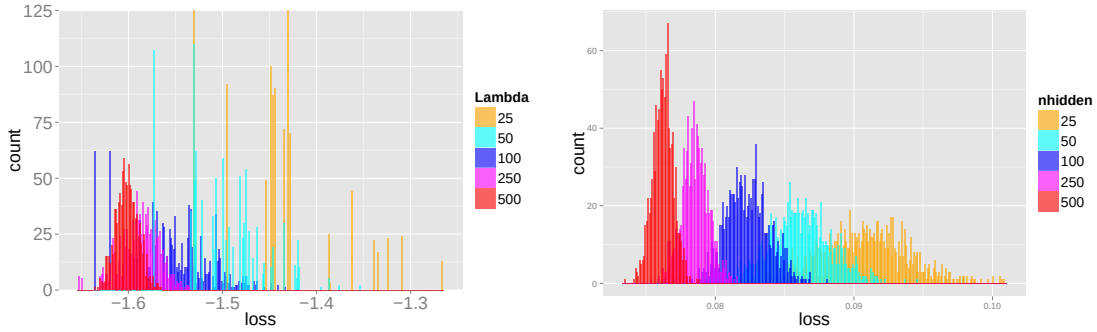

Figure 1: Distributions of the scaled test losses for the spin-glass (left) and the neural network (right) experiments.

In this section we briefly summarize a subset of results from Choromanska et al. (2015) showing the similarity between the loss function of the neural networks and the Hamiltonian of the spherical spin-glass models. The spin glass model was simulated for Λ from 25 to 500, where for each value of Λ, the distribution of minima was obtained by sampling 1000 initial points on the unit sphere and performing stochastic gradient descent (SGD) to find a minimum energy point. The neural network model was simulated using a scaled-down version of MNIST, where each image was downsampled to size 10×10. 1000 networks were trained with one hidden layer and $\mathsf{nhidden} = \{25, 50, 100, 250, 500\}$ hidden units, each one starting from a random set of parameters sampled uniformly within the unit cube. All networks were trained for 200 epochs using SGD with learning rate decay. The distribution of the scaled test losses[2] is compared in Figure 1 for both models. We see that for small values of Λ and $\mathsf{nhidden}$, we obtain poor local minima[3] on many experiments. For larger values of Λ and $\mathsf{nhidden}$, the variance of losses decreases, and the distribution becomes increasingly concentrated around the energy barrier where local minima have high quality. This indicates that (i) getting stuck in poor local minima is a major problem for smaller networks but becomes gradually of less importance as the network size increases, and (ii) in case of larger networks recovered local minima are typically corresponding to the same test performance, which is not the case for small networks.

Appendix B. Spherical spin-glass model

Figure 2 captures exemplary plots of the distributions of the mean number of critical points, local minima and low-index saddle points. Clearly local minima and low-index saddle points are located in the band $(-\Lambda E_0(H), -\Lambda E_\infty(H))$, where $-\Lambda E_\infty(H)$ is the energy barrier and $-\Lambda E_0(H)$ corresponds to the ground state (global minimum), whereas high-index saddle points can only be found above the energy barrier $-\Lambda E_\infty(H)$. This 'geometric' structure, if it is also true for multilayer neural networks, plays a crucial role in the optimization problem. The optimizer, e.g. SGD, often

2. To observe qualitative differences in behavior for different values of Λ (for spin-glass model) and $\mathsf{nhidden}$ (for neural network), it is necessary to rescale the loss values to make their expected values approximately equal. For spin-glasses, the expected value of the loss at critical points scales linearly with Λ, therefore the losses have to be divided by Λ, whereas for neural networks, the expected value of the loss at critical points was empirically found to scale with $\mathsf{nhidden}$ according to power law $\mathbb{E}[L] \propto e^{\alpha \mathsf{nhidden}^\beta}$ (α and $beta$ are coefficients), therefore the losses were divided by $L/e^{\alpha \mathsf{nhidden}^\beta}$.
3. Almost all recovered solutions were local minima with index equal to 0 (while computing the index of solutions, all eigenvalues less than 0.001 in magnitude were set to 0).

easily avoids the band of high-index critical points, which have many negative curvature directions, and descends to the band of low-index critical points which lie closer to the global minimum. Thus finding bad-quality solution, i.e. the one far away from the global minimum, is highly unlikely for large-size networks (it is also confirmed by the experimental results in Figure 1). Furthermore, as shown in Figure 2, low-index critical points are mostly concentrated close to the energy barrier ('peaked' distribution), which would potentially explain why in case of large networks recovered local minima are typically corresponding to the same test performance which is not the case for small networks.

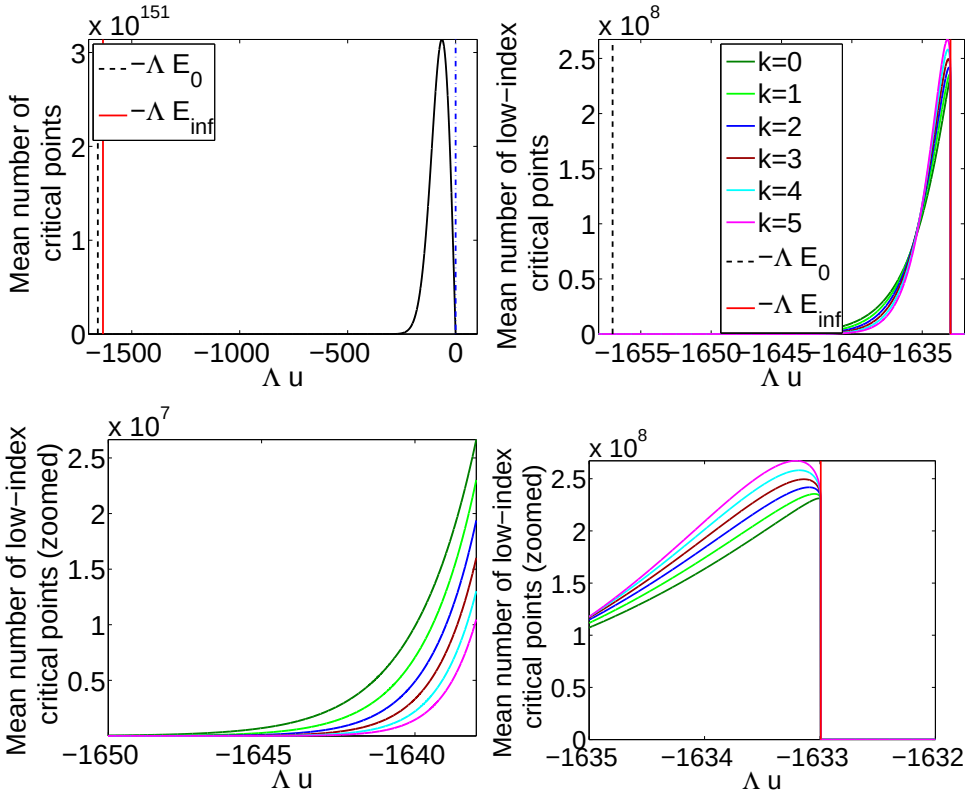

Figure 2: Distribution of the mean number of critical points, local minima and low-index saddle points (original and zoomed; k denotes the index). Parameters H and Λ were set to $H = 3$ and $\Lambda = 1000$. Black line: $u = -\Lambda E_0(H)$, red line: $u = -\Lambda E_\infty(H)$. $-\Lambda E_0$ corresponds to ground state (global minimum). Figure must be read in color.