Energy Conservation in Infinitely Wide Neural-Networks

Shu Eguchi(✉) and Takafumi Amaba

Fukuoka University, 8-19-1 Nanakuma, Jônan-ku, Fukuoka 814-0180, Japan
fmamaba@fukuoka-u.ac.jp

Abstract. A three-layered neural-network (NN), which consists of an input layer, a wide hidden layer and an output layer, has three types of parameters. Two of them are pre-neuronal, namely, thresholds and weights to be applied to input data. The rest is post-neuronal weights to be applied after activation. The current paper consists of the following two parts. First, we consider three types of stochastic processes. They are constructed by summing up each of parameters over all neurons at each epoch, respectively. The neuron number will be regarded as another time different to epochs. In the wide neural-network with a neural-tangent-kernel- (NTK-) parametrization, it is well known that these parameters are hardly varied from their initial values during learning. We show that, however, the stochastic process associated with the post-neuronal parameters is actually varied during the learning while the stochastic processes associated with the pre-neuronal parameters are not. By our result, we can distinguish the type of parameters by focusing on those stochastic processes. Second, we show that the variance (sort of "energy") of the parameters in the infinitely wide neural-network is conserved during the learning, and thus it gives a conserved quantity in learning.

Keywords: Wide neural-networks · Cumulative sum of parameters · Energy conservation

1 Introduction

In recent years, great developments have been made in understanding the mechanisms of training of a neural networks when the width of the network is large. The first step was given in Neal [7], where it was shown that for any NTK-parametrized NN, the output before training converges to a Gaussian process on the space of inputs as the width increases. This means that even in the case of a neural network with nonlinear transformations, Bayesian regression with this Gaussian process as its prior distribution is tractable when we take the limit of width to infinity (Williams [12] and Goldberg et al. [2]). This idea has been extended to deep neural-networks by Lee et al. [5].

The Bayesian regression and training by gradient method have been linked by Jacot et al. ([4]). They found that gradient method in a NTK-parametrized

© Springer Nature Switzerland AG 2021
I. Farkaš et al. (Eds.): ICANN 2021, LNCS 12894, pp. 177–189, 2021.
https://doi.org/10.1007/978-3-030-86380-7_15

NN with the large width is equivalent to kernel learning with the neural tangent kernel (NTK), and found a connection between the kernel and the maximum-a-posteriori estimator in Bayesian inference. They and Lee et al. ([6]) also showed that as the NTK-parametrized NN becomes wider, the model becomes linearized along the gradient descent or flow as the training, and the parameters become harder to be changed. This "lazy" regime appears, as shown in Chizat et al. [1], not only in over-parametrized neural-networks, but also in more abstract settings depending on the choice of scaling and initialization.

Due to the universal nature discovered in [4] and [6], we have not been able to distinguish whether they are pre- or post-neuronal if we focus on the behavior of the parameters. In this paper, we show that, during the learning, the behaviors of the cumulative sums of parameters over all neurons are different from each other according to their types of parameters. This implies that it is possible to distinguish whether the parameters are pre- or post-neuronal. When the width of the network tends to infinity, we also show that the "energy" of the cumulative sum is conserved (Theorem 2).

2 Related Works

Integral Representation of Mean-Field Parametrized NN. A mean-field parametrized NN forms like a Riemann sum, and thus has an integral representation when the width tends to infinity. In Sonoda-Murata [8] and Murata [11], the relationship between the distribution of parameters and the output is described via ridgelet transformation and their reconstruction theorem. On the other hand, in the case of our NTK-parametrized NN, the output before training is given by a *stochastic integral* when the network is infinitely wide. It would be of independent interest to investigate the reconstruction theorem in this situation.

Dynamics of Infinitely Wide Mean-Field Parametrized NN. For training of mean-field parametrized NN, another method for training is the stochastic gradient descent. It is described as a stochastic differential equation in the parameter space, in particular, it gives a gradient Langevin dynamics. When the width of the network is infinite, the parameter space is infinite-dimensional. Then the corresponding dynamics is described by an infinite-dimensional Langevin dynamics in a reproducing kernel Hilbert space, which appears as a collection of features. This infinite-dimensional model contains all models of finite width, and thus allows us to analyze them universally among all models with finite width. The convergence of this learning and the generalization error are discussed in Suzuki [9] and Suzuki-Akiyama [10].

3 Our Contribution

We consider the following NTK-parametrized NN of the width m:

$$f(x;\theta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} b_j \sigma(a_j x + a_{0,j}).$$

Here, the input $x \in \mathbb{R}$ is one-dimensional and the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be non-negative and Lipschitz continuous. We denote the coordinates of the parameter $\theta = (\boldsymbol{a}_0, \boldsymbol{a}, \boldsymbol{b})$ as follows.

- Pre-neuronal thresholds: $\boldsymbol{a}_0 = (a_{0,1}, a_{0,2}, \ldots, a_{0,m}) \in \mathbb{R}^m$,
- Pre-neuronal weights: $\boldsymbol{a} = (a_1, a_2, \ldots, a_m) \in \mathbb{R}^m$,
- Post-neuronal weights: $\boldsymbol{b} = (b_1, b_2, \ldots, b_m) \in \mathbb{R}^m$.

Given a training data $\{(x_i, y_i)\}_{i=1}^n$, we put $\hat{y}_i(\theta) := f(x_i; \theta)$ and define a loss function by

$$L(\theta) := \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i(\theta) - y_i)^2.$$

The solution to the associated gradient flow equation $\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\frac{1}{2}(\nabla_\theta L)(\theta(t))$ is denoted by $\theta(t) = (\boldsymbol{a}_0(t), \boldsymbol{a}(t), \boldsymbol{b}(t)) = (\{a_{0,j}(t)\}_{j=1}^m, \{a_j(t)\}_{j=1}^m, \{b_j(t)\}_{j=1}^m)$, where we set its initialization by $\theta(0) = (\boldsymbol{a}_0(0), \boldsymbol{a}(0), \boldsymbol{b}(0)) \sim \mathrm{N}(\boldsymbol{0}, I_{3m})$. Here, I_{3m} is the identity matrix of order $3m$.

It is known that when the width m of the network is sufficiently large and training is performed, the optimal parameters are obtained as values close to the initial ones (Jacot et al. [4]). In this paper, we further investigate behaviors of the parameters. Specifically, we consider cumulative sums of the parameters over all neurons at each epoch, which are normalized by a scale depending on the width m. We focus on what arises when we take the normalized cumulative sums along the gradient flow, even the values of parameters are hardly varied. It is enough to consider only two cumulative sums $\sum_{j=1}^m a_j(0)$ and $\sum_{j=1}^m b_j(0)$ associated with pre- and post-neuronal weights respectively since thresholds have the same role as pre-neuronal weights by considering $\{(x_i, 1)\}_{i=1}^n$ as a two-dimensional input.

To compare their behaviors among different widths during the training, we have to consider which scale is appropriate to normalize the cumulative sums of the parameters. The initialization gives us a hint. At the initialization, variances of the cumulative sums are given by $\sum_{j=1}^m \mathrm{Var}(a_j(0)) = \sum_{j=1}^m \mathrm{Var}(b_j(0)) = m$. Thus it would be natural to normalize $\sum_{j=1}^m a_j(0)$ and $\sum_{j=1}^m b_j(0)$ by scaling of $\sqrt{m}$. Moreover, we embed them into the space of continuous functions on the interval $[0, 1]$ as follows. On the m-equidistant partition $\{s_k := \frac{k}{m}\}_{k=0}^m$ of the interval, we set $A_{s_k}^{(m)}(t) := \frac{1}{\sqrt{m}} \sum_{j=1}^k a_j(t)$ and $B_{s_k}^{(m)}(t) := \frac{1}{\sqrt{m}} \sum_{j=1}^k b_j(t)$ and then we extend them onto subintervals $[s_{k-1}, s_k]$ by linear interpolations:

$$\begin{aligned} A_s^{(m)}(t) &:= \frac{A_{s_k}^{(m)}(t) - A_{s_{k-1}}^{(m)}(t)}{s_k - s_{k-1}}(s - s_{k-1}) + A_{s_{k-1}}^{(m)}(t), \\ B_s^{(m)}(t) &:= \frac{B_{s_k}^{(m)}(t) - B_{s_{k-1}}^{(m)}(t)}{s_k - s_{k-1}}(s - s_{k-1}) + B_{s_{k-1}}^{(m)}(t) \end{aligned} \qquad \text{if } s_{k-1} \le s \le s_k.$$

For each width m and time t of the gradient flow, these embedded functions $A^{(m)}(t) = \{A_s^{(m)}(t)\}_{0 \le s \le 1}$ and $B^{(m)}(t) = \{B_s^{(m)}(t)\}_{0 \le s \le 1}$ are random continuous-functions on $[0, 1]$, namely, stochastic processes.

With this embedding, it will be necessary that they do not diverge when $m \to \infty$ in order to compare them appropriately among various widths. At the initialization, by the so-called Donsker's invariance principle, which is well known in probability theory, the stochastic processes $\{(A^{(m)}(0), B^{(m)}(0))\}_{m=1}^{\infty}$ converge to a two-dimensional Brownian motion. In general, for any time t of the gradient flow, the following is valid.

Theorem 1. *The family* $\{(A^{(m)}(t), B^{(m)}(t))\}_{m=1}^{\infty}$ *is tight.*

This implies that a certain subsequence $\{(A^{(m_k)}(t), B^{(m_k)}(t))\}_{k=1}^{\infty}$ converges almost surely (by replacing the probability space appropriately if necessary). In what follows, we denote the subsequence again by $\{(A^{(m)}(t), B^{(m)}(t))\}$ for simplicity of notations. The limit $(A(t), B(t))$ of this subsequence gives a dynamics on the infinite-dimensional Banach space $C([0,1] \to \mathbb{R}^2)$ and then it would be another interest to describe the dynamics. In terms of $B(t) = \{B_s(t)\}_{0 \le s \le 1}$, we have

$$f(x_i;\theta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sigma(a_j x_i + a_{0,j}) b_j \to \int_0^1 \sigma(a_s x + a_{0,s}) \mathrm{d}B_s(0) =: \hat{y}_i^{(\infty)}$$

in probability as $m \to \infty$, and this limit is called a stochastic integral. In the above, $\{a_s\}_{0 \le s \le 1}$ and $\{a_{0,s}\}_{0 \le s \le 1}$ are mutually independent Gaussian processes on $[0,1]$ with a zero mean and the covariance function given by $\mathbf{E}[a_s a_u] = \mathbf{E}[a_{0,s} a_{0,u}] = \mathbf{1}_{\{0\}}(u - s)$. Here, $\mathbf{1}_{\{0\}}$ is the indicator function of the singleton $\{0\}$. These are also independent of $B(0)$. Although it can be smoothly expected that the dynamics of $\{(A(t), B(t))\}_{t \ge 0}$ is described by the neural tangent kernel, since $C([0,1] \to \mathbb{R}^2)$ is a non-Hilbert Banach space, it is difficult to employ the concepts of their gradient and kernel that depend on the inner product structure.

Now, among NTK-parametrized NNs of various widths, we can compare the dynamics for the cumulative sum at an "appropriate scale". Figure 1 shows outputs of neural networks widths of $m = 100, 1000, 10000$ after training. The training data are indicated by points, and we have used gradient descent. The following Figs. 2, 3, 4 and 5 show the changes of the parameters and their cumulative sums during the training. Each line in Figs. 2 and 4 represents how the corresponding parameter is varied during the training.

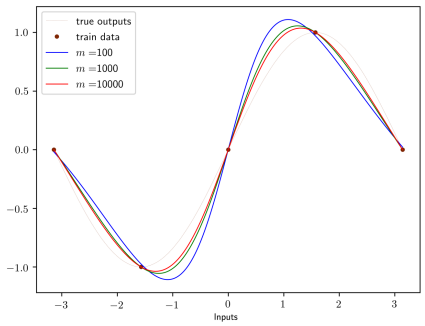

Fig. 1. Outputs after training

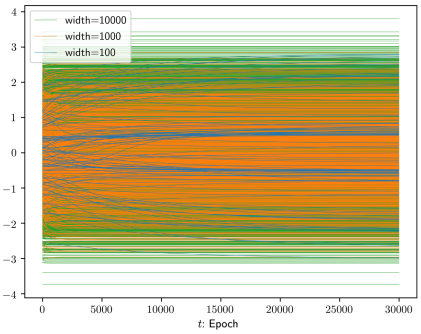

Fig. 2. Changes of parameters a_j during the training

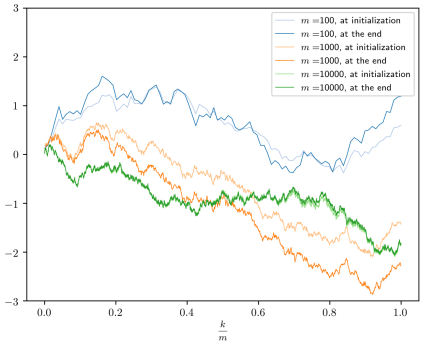

Fig. 3. Cumulative sums of parameters a_j before/after the training

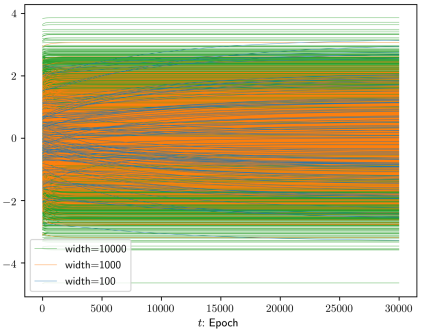

Fig. 4. Changes of parameters b_j during the training

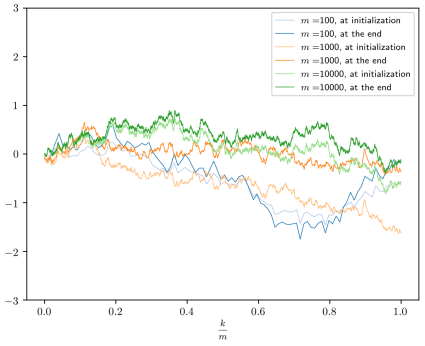

Fig. 5. Cumulative sums of parameters b_j before/after the training

From the figures, as width increases, the variation of cumulative sum becomes smaller for parameters a, while we can see it is actually varied for parameters b.

In fact, when $t = 0$ and $m \to \infty$, by the law of large numbers, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} A_{s_m}^{(m)}(t) &= -\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i(\theta(0)) - y_i\big)\frac{1}{m}\sum_{j=1}^{m}\sigma'\big(a_j(0)x_i + a_{0,j}(0)\big)x_i b_j(0) \\ &\to -\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i^{(\infty)} - y_i\big)\mathbf{E}\big[\sigma'\big(a_1(0)x_i + a_{0,1}(0)\big)x_i\big]\mathbf{E}[b_1(0)] = 0.\end{aligned}$$

On the other hand, since the activation function σ is non-negative and non-zero,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} B_{s_m}^{(m)}(t) &= -\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i(\theta(0)) - y_i\big)\frac{1}{m}\sum_{j=1}^{m}\sigma\big(a_j(0)x_i + a_{0,j}(0)\big) \\ &\to -\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i^{(\infty)} - y_i\big)\mathbf{E}\big[\sigma\big(a_1(0)x_i + a_{0,1}(0)\big)\big] \neq 0.\end{aligned}$$

As above, we observed numerically that the cumulative sum of the parameters b is varied along the gradient flow. It can be shown, however, that the following "energy" is conserved along the gradient flow.

Theorem 2. *We have* $\lim_{m\to\infty} \frac{1}{m}\sum_{j=1}^{m} \big(b_j(t) - \mathbf{E}[b_j(t)]\big)^2 = 1$ *for all* $t \geq 0$.

Here, $\mathbf{E}$ denotes the expectation operator. The same for $a_{0,j}(t)$ and $a_j(t)$.

Figures 6 and 7 below confirm Theorem 2 in the learning shown in Fig. 1. The expectations have been simulated with using Monte Carlo methods.

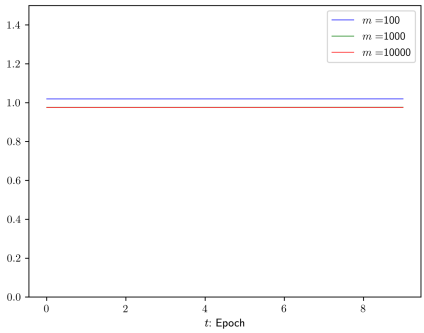

Fig. 6. Graph of $\frac{1}{m}\sum_{j=1}^{m}(a_j(t)-\mathbf{E}[a_j(t)])^2$

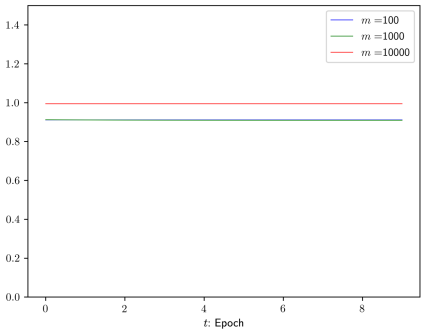

Fig. 7. Graph of $\frac{1}{m}\sum_{j=1}^{m}(b_j(t) - \mathbf{E}[b_j(t)])^2$

4 Conclusion

In this paper, we showed that in a three-layer wide neural-network, the cumulative sum of pre-neuronal parameters is hardly varied along the gradient flow, while it is varied for post-neuronal parameters. This allowed us to find a critical difference among the behaviors of the pre- and post-neuronal parameters, this is a first trial to distinguish them, which has not been so far. Furthermore, we showed that the energy is conserved along the gradient flow.

Acknowledgments. The authors would like to express their appreciation to Professor Masaru Tanaka and Professor Jun Fujiki who provided valuable comments and advices.

A Proof of Theorem 1 and Theorem 2

Recall that the activation function σ has been assumed to be non-negative and Lipschitz continuous. Then σ is differentiable almost everywhere and the Lipschitz constant can be expressed as $\|\sigma'\|_\infty := \operatorname{ess\,sup}|\sigma'|$, where σ' is the almost-everywhere-defined derivative of σ. We shall put $|\mathcal{X}| := \max_{i=1,2,\ldots,n}|x_i|$, where

$\{x_i\}_{i=1}^m$ is the input data. Note that the loss function $L(\theta) = \frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta) - y_i\big)^2$ depends on the width m as does so for the outputs $\hat{y}_i(\theta) = \frac{1}{\sqrt{m}}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})b_j$.

A.1 Equipments About Gradient Flow $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta(t)} = -\frac{1}{2}(\boldsymbol{\nabla_\theta L})(\boldsymbol{\theta(t)})$

Lemma 1. *Along the gradient flow, we have $L(\theta(t)) \le L(\theta(0))$ for $t \ge 0$.*

In the coordinate $\theta(t) = (\boldsymbol{a}_0(t), \boldsymbol{a}(t), \boldsymbol{b}(t)) = (\{a_{0,j}(t)\}_{j=1}^m, \{a_j(t)\}_{j=1}^m, \{b_j(t)\}_{j=1}^m)$, the gradient flow $\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\frac{1}{2}(\nabla_\theta L)(\theta(t))$ can be written as follows: for $j = 1, 2, \ldots, m$ and $t \in \mathbb{R}$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}a_{0,j}(t) &= -\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta(t)) - y_i\big)\sigma'\big(a_j(t)x_i + a_{0,j}(t)\big)\frac{b_j(t)}{\sqrt{m}},\\
\frac{\mathrm{d}}{\mathrm{d}t}a_j(t) &= -\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta(t)) - y_i\big)\sigma'\big(a_j(t)x_i + a_{0,j}(t)\big)x_i\frac{b_j(t)}{\sqrt{m}},\\
\frac{\mathrm{d}}{\mathrm{d}t}b_j(t) &= -\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta(t)) - y_i\big)\sigma\big(a_j(t)x_i + a_{0,j}(t)\big)\frac{1}{\sqrt{m}}.
\end{aligned} \tag{1}$$

Proposition 1. *For $m = 1, 2, 3, \ldots$, $j = 1, 2, \ldots, m$ and $t \ge 0$, we have*

$$F_j(t) \le \left(F_j(0) + \frac{\sigma(0)t}{\sqrt{m}}\sqrt{L(\theta(0))}\right)\mathrm{e}^{\frac{\|\sigma'\|_\infty(|\mathcal{X}|+1)}{\sqrt{m}}\sqrt{L(\theta(0))}\,t},$$

where $F_j(t) := |a_{0,j}(t)| + |a_j(t)| + |b_j(t)|$.

Proof. We begin with estimating $a_j(t)$. Let $\dot{a}_j(s) := \frac{\mathrm{d}}{\mathrm{d}s}a_j(s)$. By fundamental theorem of calculus, the triangle inequality and (1), we have

$$\begin{aligned}
|a_j(t)| &\le |a_j(0)| + \int_0^t \left|\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta(s)) - y_i\big)\sigma'\big(a_j(s)x_i + a_{0,j}(s)\big)x_i\frac{b_j(s)}{\sqrt{m}}\right|\mathrm{d}s\\
&\le |a_j(0)| + \int_0^t \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\left(\frac{1}{n}\sum_{i=1}^n |\hat{y}_i(\theta(s)) - y_i|\right)|b_j(s)|\mathrm{d}s.
\end{aligned}$$

Since it holds that

$$\frac{1}{n}\sum_{i=1}^n |\hat{y}_i(\theta(s)) - y_i| \le \sqrt{L(\theta(s))} \le \sqrt{L(\theta(0))} \tag{2}$$

by virtue of Jensen's inequality and Lemma 1, we obtain

$$|a_j(t)| \le |a_j(0)| + \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\sqrt{L(\theta(0))}\int_0^t |b_j(s)|\mathrm{d}s. \tag{3}$$

Similarly, we have

$$|a_{0,j}(t)| \le |a_{0,j}(0)| + \frac{\|\sigma'\|_\infty}{\sqrt{m}}\sqrt{L(\theta(0))}\int_0^t |b_j(s)|\mathrm{d}s. \tag{4}$$

For $b_j(t)$, by estimating in a manner similar to $|a_j(t)|$, we get

$$|b_j(t)| \le |b_j(0)| + \int_0^t \frac{1}{n}\sum_{i=1}^n |\hat{y}_i(\theta(s)) - y_i| \cdot \sigma\big(a_j(s)x_i + a_{0,j}(s)\big)\frac{1}{\sqrt{m}}\mathrm{d}s.$$

By using a estimate: $\sigma\big(a_j x_i + a_{0,j}\big) \le \sigma(0) + \|\sigma'\|_\infty(|\mathcal{X}||a_j| + |a_{0,j}|)$ and (2),

$$|b_j(t)| \le |b_j(0)| + \frac{\sigma(0)t}{\sqrt{m}}\sqrt{L(\theta(0))} + \frac{\|\sigma'\|_\infty t(|\mathcal{X}|+1)}{\sqrt{m}}\sqrt{L(\theta(0))}\int_0^t (|a_j(s)| + |a_{0,j}(s)|)\mathrm{d}s. \tag{5}$$

By putting estimates (3), (4) and (5) together, we have

$$F_j(t) \le F_j(0) + \frac{\sigma(0)t}{\sqrt{m}}\sqrt{L(\theta(0))} + \frac{\|\sigma'\|_\infty(|\mathcal{X}|+1)}{\sqrt{m}}\sqrt{L(\theta(0))}\int_0^t F_j(s)\mathrm{d}s.$$

Now, by applying Grönwall's inequality, we reach the conclusion.

Proposition 2. *For every $j = 1, 2, \ldots, m$, we have*

(i) $\displaystyle\int_0^t F_j(u)\mathrm{d}u \le G_j(t)$,

(ii) $\displaystyle\int_0^t \max\big\{|\dot{a}_{0,j}(u)|, |\dot{a}_j(u)|, |\dot{b}_j(u)|\big\}\mathrm{d}u \le \sqrt{\frac{L(\theta(0))}{m}}\big\{\|\sigma'\|_\infty(|\mathcal{X}|+1)G_j(t) + \sigma(0)t\big\}$,

where $F_j(u) := |a_{0,j}(u)| + |a_j(u)| + |b_j(u)|$ and

$$G_j(t) = \left(F_j(0) + \frac{\sigma(0)}{\|\sigma'\|_\infty(|\mathcal{X}|+1)}\right) t \cdot \mathrm{e}^{\frac{2\|\sigma'\|_\infty(|\mathcal{X}|+1)}{\sqrt{m}}\sqrt{L(\theta(0))}\,t}. \tag{6}$$

Note that each $G_j(t)$ depends on the width m of the network.

Proof. (i) Put $c_1 = \frac{\|\sigma'\|_\infty(|\mathcal{X}|+1)}{\sqrt{m}}\sqrt{L(\theta(0))}$ and $c_2 = \frac{\sigma(0)}{\sqrt{m}}\sqrt{L(\theta(0))}$. Then by Proposition 1, we have

$$\int_0^t F_l(u)\mathrm{d}u \le \int_0^t (F_l(0) + c_2 u)\,\mathrm{e}^{c_1 u}\mathrm{d}u \le F_l(0)\frac{\mathrm{e}^{c_1 t} - 1}{c_1 t}t + \frac{c_2}{c_1}t \cdot \mathrm{e}^{c_1 t}.$$

Since it holds that $\frac{\mathrm{e}^x - 1}{x} \le \mathrm{e}^{2x}$ for $x > 0$, we obtain

$$\int_0^t F_l(u)\mathrm{d}u \le F_l(0)\mathrm{e}^{2c_1 t} \cdot t + \frac{c_2}{c_1}t \cdot \mathrm{e}^{c_1 t} \le \left(F_l(0) + \frac{c_2}{c_1}\right) t \cdot \mathrm{e}^{2c_1 t} = G_j(t).$$

(ii) We show only for $\int_0^t |\dot{b}_j(u)|\mathrm{d}u$. The same is for the other parameters. By (1) and (2), we get $\int_0^t |\dot{b}_j(u)|\mathrm{d}u \le \sqrt{\frac{L(\theta(0))}{m}}\int_0^t \sigma\big(a_j(u)x_i + a_{0,j}(u)\big)\mathrm{d}u$. Then by using that $\sigma\big(a_j(u)x_i + a_{0,j}(u)\big) \le \|\sigma'\|_\infty(|\mathcal{X}|+1)F_j(u) + \sigma(0)$ and by (i), we have the conclusion.

Proposition 3. *For all $p > 0$, we have the following:* $\limsup_{m\to\infty} \mathbf{E}[(\sqrt{L(\theta(0))})^p] < \infty$, $\limsup_{m\to\infty} \mathbf{E}[G_j(t)^p] < \infty$ *and* $\limsup_{m\to\infty} \mathbf{E}[(\sqrt{L(\theta(0))}\, G_j(t))^p] < \infty$.

Proof. The last estimate follows from the first two estimates and Cauchy-Schwarz' inequality. Since the first estimate is obvious, we show only the second. For this, it is sufficient to show that

$$\limsup_{m\to\infty} \mathbf{E}[\mathrm{e}^{\frac{p}{\sqrt{m}}\sqrt{L(\theta(0))}}] < \infty. \tag{7}$$

In the following, we write $a_{0,j}(0) = a_{0,j}$, $a_j(0) = a_j$ and $b_j(0) = b_j$. First, we note that $\sqrt{L(\theta(0))} \le \frac{1}{\sqrt{n}} \sum_{i=1}^n |\hat{y}_i(\theta(0)) - y_i| \le \frac{1}{\sqrt{n}} \sum_{i=1}^n |\hat{y}_i(\theta(0))| + \frac{1}{\sqrt{n}}|y_i|$. Then by using Hölder's inequality, we get

$$\begin{aligned}
\mathbf{E}[\mathrm{e}^{\frac{p}{\sqrt{m}}\sqrt{L(\theta(0))}}] &\le \mathrm{e}^{\frac{p}{\sqrt{nm}}\sum_{i=1}^n |y_i|} \Big(\prod_{i=1}^n \mathbf{E}\big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}|\hat{y}_i(\theta(0))|}\big] \Big)^{1/n} \\
&\le \mathrm{e}^{\frac{p}{\sqrt{nm}}\sum_{i=1}^n |y_i|} \max_{i=1,2,\ldots,n} \mathbf{E}\big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}|\hat{y}_i(\theta(0))|}\big] \le \mathrm{e}^{\frac{p}{\sqrt{nm}}\sum_{i=1}^n |y_i|} \sum_{i=1}^n \mathbf{E}\big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}|\hat{y}_i(\theta(0))|}\big].
\end{aligned}$$

Since we have $(\hat{y}_i(\theta(0)) \mid \boldsymbol{a}_0, \boldsymbol{a}) \sim \mathrm{N}\big(0, \frac{1}{m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2\big)$,

$$\begin{aligned}
\mathbf{E}\big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}|\hat{y}_i(\theta(0))|}\big] &= \sqrt{\frac{2}{\pi}} \int_0^\infty \mathbf{E}\Big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}\left(\frac{1}{m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2\right)^{1/2} w}\Big] \mathrm{e}^{-\frac{w^2}{2}}\,\mathrm{d}w \\
&\le \sqrt{\frac{2}{\pi}} \int_{-\infty}^\infty \mathbf{E}\Big[\mathrm{e}^{\frac{p\sqrt{n}}{\sqrt{m}}\left(\frac{1}{m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2\right)^{1/2} w}\Big] \mathrm{e}^{-\frac{w^2}{2}}\,\mathrm{d}w = 2\mathbf{E}\Big[\mathrm{e}^{\frac{p^2 n}{2m}\frac{1}{m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2}\Big].
\end{aligned}$$

Furthermore, by Jensen's inequality and independence,

$$\mathbf{E}\Big[\mathrm{e}^{\frac{p^2 n}{2m}\frac{1}{m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2}\Big] \le \mathbf{E}\Big[\mathrm{e}^{\frac{p^2 n}{2m}\sum_{j=1}^m \sigma(a_j x_i + a_{0,j})^2}\Big]^{1/m} = \mathbf{E}\Big[\mathrm{e}^{\frac{p^2 n}{2m}\sigma(a_1 x_i + a_{0,1})^2}\Big].$$

We can show that $\sigma(a_1 x_i + a_{0,1})^2 \le 16\{\|\sigma'\|_\infty(|\mathcal{X}| + 1)\}^2\{(a_{0,1})^2 + (a_1)^2\} + (\sigma(0))^2$. Hence

$$\mathbf{E}\Big[\mathrm{e}^{\frac{p^2 n}{2m}\sigma(a_1 x_i + a_{0,1})^2}\Big] \le \mathrm{e}^{\frac{p^2 n}{2m}(\sigma(0))^2} \cdot \mathbf{E}\Big[\mathrm{e}^{\frac{8p^2 n\|\sigma'\|_\infty^2(|\mathcal{X}|+1)^2}{m}((a_{0,1})^2 + (a_1)^2)}\Big].$$

The right-hand-side is finite if $\frac{8p^2 n\|\sigma'\|_\infty^2(|\mathcal{X}|+1)^2}{m} - \frac{1}{2} < 0$, that is, $m > 16p^2 n\|\sigma'\|_\infty^2(|\mathcal{X}| + 1)^2$, and then it is decreasing with respect to m. By putting all together, (7) is proved.

A.2 Proof of Theorem 1

It is enough to prove that both of $\{A^{(m)}(t)\}_{m=1}^\infty$ and $\{B^{(m)}(t)\}_{m=1}^\infty$ are tight. For this, from [3, Chapter I, Section 4, Theorem 4.3], it is sufficient to show that (i) $\sup_m \mathbf{E}\big[|A_0^{(m)}(t)| + |B_0^{(m)}(t)|\big] < \infty$ and (ii) there exist $\gamma, \alpha > 0$ such that

$$\sup_m \sup_{\substack{s,u\in[0,1]: \\ s\neq u}} \left(\frac{\mathbf{E}\big[|A_s^{(m)}(t) - A_u^{(m)}(t)|^\gamma\big]}{|s-u|^{1+\alpha}} + \frac{\mathbf{E}\big[|B_s^{(m)}(t) - B_u^{(m)}(t)|^\gamma\big]}{|s-u|^{1+\alpha}} \right) < \infty.$$

(i) is clear since $A_0^{(m)}(t) = B_0^{(m)}(t) = 0$. Thus we show only (ii). We will only show the one for $A^{(m)}(t)$. Since $A^{(m)}(t)$ is a piecewise linear interpolation of values on $\{s_k = \frac{k}{m}\}_{k=0}^m$, it suffices to show that for some $\gamma, \alpha > 0$, it holds that

$$\sup_m \sup_{\substack{1\le k,j\le m:\\ k\neq j}} \frac{\mathbf{E}\big[|A_{s_k}^{(m)}(t) - A_{s_j}^{(m)}(t)|^\gamma\big]}{|s_k - s_j|^{1+\alpha}} < \infty. \tag{8}$$

Let $k, j \in \{1, 2, \dots, m\}$ be arbitrary. Without loss of generality, we assume that $j < k$. Then we have

$$|A_{s_k}^{(m)}(t) - A_{s_j}^{(m)}(t)| \le \frac{1}{\sqrt{m}}\Big|\sum_{l=j+1}^k \big(a_l(t) - \mathbf{E}[a_l(t)]\big)\Big| + \frac{1}{\sqrt{m}}\Big|\sum_{l=j+1}^k \mathbf{E}[a_l(t)]\Big|. \tag{9}$$

We shall make estimates for two terms on the right-hand-side.

Lemma 2. *With $G_l(t)$ defined in (6), we have*

$$\Big|\sum_{l=j+1}^k (a_l(t) - \mathbf{E}[a_l(t)])\Big| \le \Big|\sum_{l=j+1}^k a_l(0)\Big| + \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}} \sum_{l=j+1}^k \big(\sqrt{L(\theta(0))}\, G_l(t) + \mathbf{E}[\sqrt{L(\theta(0))}\, G_l(t)]\big).$$

Proof. Since $\mathbf{E}[a_l(0)] = 0$, we have $a_l(t) - \mathbf{E}[a_l(t)] = \int_0^t \dot{a}_l(u)\mathrm{d}u - \int_0^t \mathbf{E}[\dot{a}_l(u)]\mathrm{d}u + a_l(0)$. By summing up this over $l = j+1, j+2, \dots, k$ and by using (1) and (2),

$$\begin{aligned}\Big|\sum_{l=j+1}^k (a_l(t) - \mathbf{E}[a_l(t)])\Big| \le \Big|\sum_{l=j+1}^k a_l(0)\Big| &+ \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\sqrt{L(\theta(0))} \sum_{l=j+1}^k \int_0^t |b_l(u)|\mathrm{d}u \\ &+ \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\mathbf{E}\big[\sqrt{L(\theta(0))} \sum_{l=j+1}^k \int_0^t |b_l(u)|\mathrm{d}u\big].\end{aligned}$$

Finally, by applying Proposition 2, we get the conclusion.

Lemma 3. *We have* $\Big|\sum_{l=j+1}^k \mathbf{E}[a_l(t)]\Big| \le \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\mathbf{E}\big[\sqrt{L(\theta(0))} \sum_{l=j+1}^k G_l(t)\big]$.

Proof. By (1), $\mathbf{E}[a_l(t)] = \int_0^t \mathbf{E}[-\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i(\theta(u)) - y_i\big)\sigma'\big(a_l(u)x_i + a_{0,l}(u)\big)x_i \frac{b_l(u)}{\sqrt{m}}]\mathrm{d}u$. By taking the sum over $l = j+1, j+2, \dots, k$, we have

$$\Big|\sum_{l=j+1}^k \mathbf{E}[a_l(t)]\Big| \le \frac{\|\sigma'\|_\infty|\mathcal{X}|}{\sqrt{m}}\mathbf{E}\Big[\sqrt{L(\theta(0))} \sum_{l=j+1}^k \int_0^t |b_l(u)|\mathrm{d}u\Big].$$

Then by using Proposition 2, we reach the conclusion.

Turning back to Eq. (9), we apply Lemma 2 and Lemma 3 to get

$$|A_{s_k}^{(m)}(t) - A_{s_j}^{(m)}(t)| \le \frac{1}{\sqrt{m}}\Big|\sum_{l=j+1}^k a_l(0)\Big| + \frac{\|\sigma'\|_\infty|\mathcal{X}|}{m} \sum_{l=j+1}^k \big(2H_l(t) + \mathbf{E}[H_l(t)]\big),$$

where $H_l(t) = \sqrt{L(\theta(0))}\, G_l(t)$. By an easy estimate: $(x+y)^4 \le 2^4(x^4+y^4)$,

$$\begin{aligned}&\big(A^{(m)}_{s_k}(t) - A^{(m)}_{s_j}(t)\big)^4\\ &\le \frac{2^4}{m^2}\left(\sum_{l=j+1}^{k} a_l(0)\right)^4 + 2^4\|\sigma'\|_\infty^4|\mathcal{X}|^4\left(\frac{k-j}{m}\right)^4\left(\frac{1}{k-j}\sum_{l=j+1}^{k}\big(2H_l(t)+\mathbf{E}[H_l(t)]\big)\right)^4.\end{aligned}$$

Therefore $\mathbf{E}[\big(A^{(m)}_{s_k}(t) - A^{(m)}_{s_j}(t)\big)^4] = \frac{2^4}{m^2} I + 2^4\|\sigma'\|_\infty^4|\mathcal{X}|^4(s_k - s_j)^4 I\!I$. Here,

$$I := \mathbf{E}[\Big(\sum_{l=j+1}^{k} a_l(0)\Big)^4], \quad I\!I := \mathbf{E}[\Big(\frac{1}{k-j}\sum_{l=j+1}^{k}\big(2H_l(t)+\mathbf{E}[H_l(t)]\big)\Big)^4].$$

First, we shall focus on $I\!I$. By Jensen's inequality,

$$I\!I \le \frac{1}{k-j}\sum_{l=j+1}^{k}\mathbf{E}[\big(2H_l(t)+\mathbf{E}[H_l(t)]\big)^4] = \mathbf{E}[\big(2H_1(t)+\mathbf{E}[H_1(t)]\big)^4].$$

On the other hand, for I, since $a_1(0), a_2(0), \ldots, a_m(0)$ are independent and identically distributed, and each of them is distributed in $\mathrm{N}(0,1)$, we have $I = 3(k-j)^2$. Hence

$$\begin{aligned}&\mathbf{E}[\big(A^{(m)}_{s_k}(t) - A^{(m)}_{s_j}(t)\big)^4]\\ &\le 2^4\cdot 3(s_k - s_j)^2 + 2^4\|\sigma'\|_\infty^4|\mathcal{X}|^4(s_k-s_j)^4\mathbf{E}[\big(2H_1(t)+\mathbf{E}[H_1(t)]\big)^4].\end{aligned}$$

Finally, by noting Proposition 3, we see that (8) holds for $\gamma = 4$ and $\alpha = 1$.

A.3 Proof of Theorem 2

By the law of large numbers, we see that $\frac{1}{m}\sum_{j=1}^{m}\big(b_j(0)\big)^2 \to \mathbf{E}[\big(b_j(0)\big)^2] = 1$ as $m \to \infty$. Then it suffices to show that

$$\mathbf{E}[\Big|\frac{1}{m}\sum_{j=1}^{m}\big(b_j(t) - \mathbf{E}[b_j(t)]\big)^2 - \frac{1}{m}\sum_{j=1}^{m}\big(b_j(0)\big)^2\Big|] \to 0.$$

Since $b_j(t) - \mathbf{E}[b_j(t)] = b_j(0) + \int_0^t \big(\dot{b}_j(u) - \mathbf{E}[\dot{b}_j(u)]\big)\mathrm{d}u$, we have $(b_j(t) - \mathbf{E}[b_j(t)])^2 - (b_j(0))^2 = \big(\int_0^t \big(\dot{b}_j(u) - \mathbf{E}[\dot{b}_j(u)]\big)\mathrm{d}u\big)^2 + 2b_j(0)\int_0^t(\dot{b}_j(u) - \mathbf{E}[\dot{b}_j(u)])\mathrm{d}u$. Thus we have

$$\begin{aligned}&\Big|\frac{1}{m}\sum_{j=1}^{m}\big(b_j(t) - \mathbf{E}[b_j(t)]\big)^2 - \frac{1}{m}\sum_{j=1}^{m}\big(b_j(0)\big)^2\Big|\\ &\le \frac{1}{m}\sum_{j=1}^{m}\Big|\left(\int_0^t\big(\dot{b}_j(u) - \mathbf{E}[\dot{b}_j(u)]\big)\mathrm{d}u\right)^2 + 2b_j(0)\int_0^t\big(\dot{b}_j(u) - \mathbf{E}[\dot{b}_j(u)]\big)\mathrm{d}u\Big|.\end{aligned}$$

By taking the expectation, we get

$$\mathbf{E}\Big[\Big|\frac{1}{m}\sum_{j=1}^{m}\big(b_j(t)-\mathbf{E}[b_j(t)]\big)^2-\frac{1}{m}\sum_{j=1}^{m}\big(b_j(0)\big)^2\Big|\Big]$$
$$\le \frac{1}{m}\sum_{j=1}^{m}\left\{\mathbf{E}\Big[\Big(\int_0^t\big(|\dot b_j(u)|+\mathbf{E}\big[|\dot b_j(u)|\big]\big)\mathrm{d}u\Big)^2\Big]+2\mathbf{E}\Big[|b_j(0)|\int_0^t\big(|\dot b_j(u)|+\mathbf{E}\big[|\dot b_j(u)|\big]\big)\mathrm{d}u\Big]\right\}.$$

For the term $\int_0^t|\dot b_j(u)|\mathrm{d}u$ appeared above, we know by Proposition 2 that

$$\int_0^t|\dot b_j(u)|\mathrm{d}u\le\sqrt{\frac{L(\theta(0))}{m}}\left\{\|\sigma'\|_\infty(|\mathcal{X}|+1)G_j(t)+\sigma(0)t\right\}=:\frac{M_j(t)}{\sqrt{m}},$$

where note that $M_j(t)$ depends on the width m. Thus, $\int_0^t\big(|\dot b_j(u)|+\mathbf{E}\big[|\dot b_j(u)|\big]\big)\mathrm{d}u\le\frac{M_j(t)+\mathbf{E}[M_j(t)]}{\sqrt{m}}$. By Proposition 3, we have $\limsup_{m\to\infty}\mathbf{E}\big[\big(M_1(t)+\mathbf{E}[M_1(t)]\big)^2\big]<\infty$ and $\limsup_{m\to\infty}\mathbf{E}\big[|b_1(0)|\big(M_1(t)+\mathbf{E}[M_1(t)]\big)\big]<\infty$. Hence as $m\to\infty$,

$$\mathbf{E}\Big[\Big|\frac{1}{m}\sum_{j=1}^{m}\big(b_j(t)-\mathbf{E}[b_j(t)]\big)^2-\frac{1}{m}\sum_{j=1}^{m}\big(b_j(0)\big)^2\Big|\Big]$$
$$\le\frac{1}{m}\sum_{j=1}^{m}\left\{\frac{\mathbf{E}\big[\big(M_j(t)+\mathbf{E}[M_j(t)]\big)^2\big]}{m}+2\frac{\mathbf{E}\big[|b_j(0)|\big(M_j(t)+\mathbf{E}[M_j(t)]\big)\big]}{\sqrt{m}}\right\}$$
$$=\frac{\mathbf{E}\big[\big(M_1(t)+\mathbf{E}[M_1(t)]\big)^2\big]}{m}+2\frac{\mathbf{E}\big[|b_1(0)|\big(M_1(t)+\mathbf{E}[M_1(t)]\big)\big]}{\sqrt{m}}\to 0.$$

References

1. Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, (NeurIPS 2019). Curran Associates, Inc (2018)
2. Goldberg, P., Williams, C., Bishop, C.: Regression with input-dependent noise: a Gaussian process treatment. In: Advances in Neural Information Processing Systems, vol. 10, NIPS 1997. MIT Press (1998)
3. Ikeda, N., Watanabe, S.: Stochastic Differential Equations and Diffusion Processes, Second edn. North-Holland Mathematical Library, 24. North-Holland Publishing Co., Amsterdam; Kodansha Ltd, Tokyo, p. xvi+555 (1989). ISBN: 0-444-87378-3
4. Jacot, A., Gabriel. F., Hongler. C.: Neural tangent kernel: convergence and generalization in neural networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, pp. 8571–8580. Curran Associates, Inc (2018)
5. Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., Sohl-Dickstein, J.: Deep neural networks as Gaussian processes. In: International Conference on Learning Representations, (ICLR 2018) (2018)

6. Lee, J., et al.: Wide neural networks of any depth evolve as linear models under gradient descent. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, (NeurIPS 2019), Curran Associates, Inc (2019)
7. Neal, R.M.: Priors for infinite networks. In: Bayesian Learning for Neural Networks, pp. 29–53. Springer, New York (1996). https://doi.org/10.1007/978-1-4612-0745-0_2
8. Sonoda, S., Murata, N.: Neural network with unbounded activation functions is universal approximator. Appl. Comput. Harmonic Anal. **43**(2), 233–268 (2017)
9. Suzuki, T.: Generalization bound of globally optimal non-convex neural network training: transportation map estimation by infinite dimensional Langevin dynamics. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.), Advances in Neural Information Processing Systems, vol. 33, (NeurIPS 2020), pp. 19224–19237. Curran Associates, Inc (2020)
10. Suzuki, T., Akiyama, S.: Benefit of deep learning with non-convex noisy gradient descent: provable excess risk bound and superiority to kernel methods. To appear in International Conference on Learning Representations, 2021 (ICLR 2021) (2021)
11. Murata, N.: An integral representation of functions using three-layered networks and their approximation bounds. Neural Netw. **9**(6), 947–956 (1996)
12. Williams, C.: Computing with infinite networks. In: Mozer, M.C., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9, (NIPS 1996), MIT Press (1997)