Generalization and learning error for non linear perceptron

M.Shcherbina[*], B.Tirozzi[†]

Abstract

A rigouros derivation of the asymptotic behaviour of learning and prediction error for the non linear perceptron is presented. The saddle-point method is used for evaluating these quantities.

1 Introduction

Neural networks (NN) have shown to be very useful in many tasks of data analysis. Almost any problem of modelling a set of data $\{(\mathbf{x}^{(\mu)}, y^{(\mu)})\}_{\mu=1}^{M}$ has been successfully solved by back propagation neural networks as well as predicting new data. For example, if we have the stochastic process $x(t)$, we can consider as an input vector $\mathbf{x} = x(t-1), x(t-2), \dots, x(t-n)$ and as an output $y = x(t)$ - the value of the process at the time t (here the time t plays the role of μ). This kind of NN is widely used and applied in many fields like meteorology (see, e.g.[1]), geology [2], economy [3], etc.

The back propagation for two layers perceptron is an algorithm inspired to the adaptation process of the brain for solving particular tasks. It has been noticed since a long time that the synaptic weights of the brain change during the growth of humans. Each group of synaptic weights of neurons changes in such a way that an elementary task, e.g. recognition of an object situated at a certain angle in the plain of observation, is solved. The use of sigmoid functions as input-output relationship and the terminology of neurons, synaptic weights are coming from the analogy with real neurones. For NN used in solving the data analysis problem (called also artificial neural networks ANN) an architecture is also introduced. There is a first layer of n input neurons, i.e. suppose that our data are n-dimensional vectors $\mathbf{x}^{(\mu)} = (x_1^{(\mu)}, \dots, x_n^{(\mu)})$ and to each neuron i is associated an input data $x_i^{(\mu)}$, then the output neuron n is connected through a synaptic weights w_j with the input neuron j and it receives as a total synaptic input the sum $\sum w_i x_i^{(\mu)} = (\mathbf{x}^{(\mu)}, \mathbf{w})$, the output of this neuron is

$$z^{(\mu)} = \sigma((\mathbf{x}^{(\mu)}, \mathbf{w})),$$

where $\sigma(x)$ is a sigmoid function similar to the usual input-output function of real neurones and is of the form

$$\sigma(x) = \frac{1}{1 + e^{-\lambda x}}.$$

This simple model of neural networks is the perceptron or one layer back propagation. The learning process is some evolution of the synaptic weights which minimize the learning error over a sequence of input-output pairs $\{(\mathbf{x}^{(\mu)}, y^{(\mu)})\}$. According to these definitions $y^{(\mu)}$ is the desired output and $z^{(\mu)}$ is the output of the network. Thus it is necessary to apply a minimizing algorithm to the error $H(\mathbf{w}, \overline{\mathbf{x}}, \overline{y})$

$$H(\mathbf{w}, \overline{\mathbf{x}}, \overline{y}) = \sum_{\mu=1}^{M} (z^{(\mu)} - y^{(\mu)})^2. \tag{1.1}$$

[*] Institute for Low Temperature Physics,Ukr. Ac. Sci., 47 Lenin ave., Kharkov, Ukraine

[†] Department of Physics of Rome University "La Sapienza", 5, p-za A.Moro, Rome, Italy

Usually the minimization is done by means of the steepest descent or Monte-Carlo method. The minimum of $H(\mathbf{w}, \overline{\mathbf{x}}, \overline{y})$ is called the learning error. Once minimization of $H(\mathbf{w}, \overline{\mathbf{x}}, \overline{y})$ is achieved on the $(\overline{\mathbf{x}}, \overline{y})$ which is called the learning set, the important question arises about the error which the ANN makes on a new input data $\mathbf{x}^{(M+1)}$ presented to the network i.e. how one can estimate the probability that the output z of the network differs from the real expected output more than a certain given ε or also directly estimate ε. Here ε is defined as the generalization error. This question has been treated in many different approaches. Vapnik and Chervonenkis [4] give an estimate of the probability that the difference among the learning error and generalization error is greater than a given constant δ. They have obtained an exponential decay if the number of pattern is larger than a constant called the Vapnik - Chervonenkis dimension, But this constant is not possible to compute in general situation.

Another known approach is the one of Amari [5] who estimates the generalization error for the perceptron and finds that it is inversely proportional to the number of patterns M. The proof of Amari is not complete because there are some assumptions which can be verified only qualitatively. Another original approach is the one done by Feng [6] who has got the same bound as Amari, using extreme value theory for the perceptron of the type

$$z = \mathrm{sign}((\mathbf{w}, \mathbf{x})).$$

There are many other approaches based on bayesian formulation of probability [7]. Some interesting estimate is also obtained by introducing the Gibbs measure, associated to the Hamiltonian (1.1). One particular variant of this problem has been found by Solla and Levin [7], using statistical mechanics estimates. They assume, that there is a perceptron transformation which generates the data

$$z^{(\mu)} = \sigma((\mathbf{w}^0, \mathbf{x}^{(\mu)})) + t\eta^{(\mu)}.$$

This is what is called the "teacher" rule. The vector $\mathbf{w}^0$ is the one which generates the output $y^{(\mu)}$, when the input $\mathbf{x}^{(\mu)}$ is presented to the network and $\eta^{(\mu)}$ is the noise which is in the data and is represented by a set of independent identically distributed Gaussian variables with variance 1. Suppose now that we want to find $\mathbf{w}$ which best approximates $\mathbf{w}^0$ for the collection of input-output $\{(\mathbf{x}^{(\mu)}, y^{(\mu)})\}$. The best approximating $\mathbf{w}$ is called the "student rule". The learning error is

$$\mathcal{L} = \frac{1}{M} E\{\langle \sum_{\mu=1}^{M} (\sigma((\mathbf{w}, \mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}^0, \mathbf{x}^{(\mu)})) + t\eta^{(\mu)})^2 \rangle\},$$

where the average is taken with respect to the Gibbs measure, generated by H (the symbol $\langle \ldots \rangle$) and also with respect to the distribution of $\mathbf{x}^{(\mu)}$ and $\eta^{(\mu)}$. In the work [7] the expression of $\mathcal{L}$ and of the generalization error $\mathcal{G}$ has been obtained in the case of linear perceptron $\sigma(x) = x$.

In the present paper we derive rigorously the asymptotic expression for the learning and generalization errors when $M \to \infty$ in the case of nonlinear perceptron, extending the result of Solla and Levin[7], without using the assumptions made in their work.

The paper is organized as follows. In Section 2 we describe the model and the results, while in Sections 2 and 3 there are the proofs of the main and auxiliary results respectively.

2 Model and results

Consider a bounded function $\sigma(\lambda)$ $(\lambda \in \mathbf{R})$ which varies from in $(-1, 1)$ and satisfies the conditions

$$\sigma'(\lambda) > 0, \quad \sigma(\pm\infty) = \pm 1, \quad |\sigma''(\lambda)| \leq \mathrm{const}\,, \quad 1 - |\sigma(\lambda)| \leq \frac{\mathrm{const}}{|\lambda|}, \; (|\lambda| \to \infty). \qquad (2.1)$$

Let $\{\mathbf{x}^{(\mu)}\}_{\mu=1}^{M+1}$ be n-dimensional independent identically distributed random vectors, whose components $\mathbf{x}_i^{(\mu)}$ are also independent for different i and

$$E\{\mathbf{x}^{(\mu)}\} = 0, \quad E\{(\mathbf{x}^{(\mu)}, \mathbf{x}^{(\mu)})\} = n, \quad E\{|\mathbf{x}^{(\mu)}|^4\} \leq \mathrm{const}\,. \qquad (2.2)$$

Besides, we assume that the distribution of $\mathbf{x}_i^{(\mu)}$ is continuous at the point $\mathbf{x}_i^{(\mu)} = 0$. Here and below $(...,...)$ means a usual scalar product in $\mathbf{R}^n$, and $|...|$ is a norm, corresponding to this scalar product. We denote also by $E\{\ldots\}$ the averages over the distribution of all random parameters of the problem.

Let $\{\eta^{(\mu)}\}_{\mu=1}^{M+1}$ be independent random variables, such that

$$E\{\eta^{(\mu)}\} = 0, \quad E\{(\eta^{(\mu)})^2\} = 1, \quad E\{e^{\lambda|\eta^{(\mu)}|}\} \le \text{const} \tag{2.3}$$

with some positive λ. We consider the Hamiltonian of the form

$$H_M(\overline{\mathbf{w}}; \overline{\mathbf{x}}, t) \equiv \frac{1}{2}\sum_{\mu=1}^{M} (\sigma((\mathbf{w}, \mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}^0, \mathbf{x}^{(\mu)})) + t\eta^{(\mu)})^2, \tag{2.4}$$

where $\mathbf{w} \in \mathbf{R}^n$, $\mathbf{w}^0 \in \mathbf{R}^n$ - is some fixed vector and we denote $\overline{\mathbf{x}} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mathbf{M})})$ and $\overline{\eta} = (\eta^{(1)}, \ldots, \eta^{(M)})$. Consider the corresponding partition function

$$Z_M(\beta, \overline{\mathbf{x}}, t) = \int \rho(\mathbf{w}) d\mathbf{w} \exp\{-\beta H(\mathbf{w}; \overline{\mathbf{x}}, \overline{\eta})\}, \tag{2.5}$$

where we denote $d\mathbf{w} \equiv \prod_{i=1}^n dw_i$ and the density function $\rho(\mathbf{w})$ supposed to be nonnegative and

$$\int \rho(\mathbf{w}) d\mathbf{w} \le \text{const}.$$

We assume also, that $\rho(\mathbf{w}^0) \neq 0$ and $\rho(\mathbf{w})$ has two bounded derivatives at the point $\mathbf{w} = \mathbf{w}^0$.

We are interested in the behaviour of the learning error

$$\mathcal{L}_M(t) = \frac{1}{M}\langle H(\mathbf{w}, \overline{\mathbf{x}}, t)\rangle, \tag{2.6}$$

where the symbol $\langle \ldots \rangle$ means the Gibbs averaging with respect to the Hamiltonian H_M:

$$\langle \ldots \rangle \equiv Z_M^{-1} \int \rho(\mathbf{w}) d\mathbf{w} (\ldots) \exp\{-\beta H(\mathbf{w}, \overline{\mathbf{x}}, t)\}.$$

The other subject of our interest is the generalization error which is given by ([5, 7])

$$\mathcal{G}_M(t) = -\ln \frac{Z_{M+1}}{Z_M}. \tag{2.7}$$

Theorem 1 *Under above conditions*

$$\begin{aligned} E\{\mathcal{L}_M(0)\} &= \frac{n}{2\beta M} + o(M^{-1}), \quad E\{\mathcal{G}_M(0)\} = \frac{n\beta}{2M} + o(M^{-1}), \\ E\{\mathcal{L}_M(t)\} &= \frac{t^2}{2} + O(M^{-1}), \quad E\{\mathcal{G}_M(t)\} = \frac{\beta t^2}{2} + O(M^{-1}) \end{aligned} \tag{2.8}$$

3 Proof of Main Results

We start from proof of Theorem 1 for $t = 0$. Let us define the constants:

$$\begin{aligned} d^2 &\equiv \min_{|w|=1} E\{(\mathbf{w}, \mathbf{x}^{(1)})^2 \theta(2n - |\mathbf{x}^{(1)}|)\} > 0, \\ K &\equiv \min_{|w|=1} E\{\theta(|(\mathbf{w}, \mathbf{x}^{(1)})| - \frac{d}{2})\theta(2n - |\mathbf{x}^{(1)}|)\} > 0. \end{aligned} \tag{3.1}$$

Remark. Let us note, that $d \neq 0$, because for any $\mathbf{w} : |\mathbf{w}| = 1$ $E\{(\mathbf{w}, \mathbf{x}^{(1)})^2\} = n$ and $\mathrm{Prob}\,\{(\mathbf{w}, \mathbf{x}^{(1)})^2 = 0\} = 0$. Thus, $\mathrm{Prob}\,\{(\mathbf{w}, \mathbf{x}^{(1)})^2 \leq 2n,\ (\mathbf{w}, \mathbf{x}^{(1)})^2 \neq 0\} \neq 0$. Similarly, from definition of d it follows that $K \neq 0$.

Take then $\varepsilon = \frac{d}{8n}$ and choose the finite set of points $\{\mathbf{w}^{(i)}\}$, $(i = 1, \dots N)$ which have unit norm $|\mathbf{w}^{(i)}| = 1$ and

$$\mathcal{S}(0,1) \subset \cup \mathcal{B}(\mathbf{w}^{(i)}, \varepsilon),$$

where $\mathcal{S}(\mathbf{w}, r)$ is the n-dimensional sphere of radius r, with center in $\mathbf{w}$, and $\mathcal{B}(\mathbf{w}, \varepsilon)$ is the n-dimensional ball of radius ε, centered in $\mathbf{w}$.

Consider the inequalities

$$\frac{1}{M}\sum_{\mu=1}^{M} \theta(|(\mathbf{w}^{(i)}, \mathbf{x}^{(\mu)})| - \frac{d}{2})\theta(2n - |\mathbf{x}^{(\mu)}|) \geq \frac{1}{2}K, \quad (i = 1, \dots, N) \tag{3.2}$$

Remark. Denoting $\xi_i^{\mu} = \theta(|(\mathbf{w}^{(i)}, \mathbf{x}^{(\mu)})| - \frac{d}{2})\theta(2n - |\mathbf{x}^{(\mu)}|)$, we have for any i the set of independent bounded random variables ξ_i^{μ} whose mean values are larger than K (see (3.1)). Therefore from the Chebyshev inequality we get that the probability to have (3.2) is larger than $(1 - e^{-M\,\mathrm{const}})$.

We put conditions (3.2) in the set of points $\mathbf{w}^{(i)}$ in order to obtain that some weaker condition holds uniformly in $\mathbf{w}$.

Proposition 1 *If inequalities (3.2) are true, then for any* $\mathbf{w} : |\mathbf{w}| = 1$

$$\frac{1}{M}\sum_{\mu=1}^{M} \theta(|(\mathbf{w}, \mathbf{x}^{(\mu)})| - \frac{d}{4})\theta(2n - |\mathbf{x}^{(\mu)}|) \geq \frac{1}{2}K. \tag{3.3}$$

Now let us study matrices which will appear in our considerations.

$$\begin{aligned}
\tilde{A}_{i,j} &\equiv \frac{1}{M}\sum_{\mu=1}^{M} |\sigma'((\mathbf{w}^0, \mathbf{x}^{\mu}))|^2 x_i^{(\mu)} x_j^{(\mu)}, \quad A_{i,j} \equiv E\{\tilde{A}_{i,j}\} = E\{|\sigma'((\mathbf{w}^0, \mathbf{x}^{(1)}))|^2 x_i^{(1)} x_j^{(1)}\}, \\
\tilde{X}_{i,j} &\equiv \frac{1}{M}\sum_{\mu=1}^{M} \theta(2n - |\mathbf{x}^{(\mu)}|) x_i^{(\mu)} x_j^{(\mu)}, \quad X_{i,j} \equiv E\{\tilde{X}_{i,j}\} = E\{\theta(2n - |\mathbf{x}^{(1)}|) x_i^{(1)} x_j^{(1)}\}.
\end{aligned} \tag{3.4}$$

It is easy to see, that the matrices $\mathbf{A} = \{A_{ij}\}$ and $\mathbf{X} = \{X_{ij}\}$ defined in such a way are positive

$$\mathbf{A} \geq \delta_1 \mathbf{I}, \quad \mathbf{X} \geq \delta_2 \mathbf{I}$$

for any distribution of $\mathbf{x}^{(\mu)}$, satisfying (2.2). This can be derived from the relations

$$\begin{aligned}
(\mathbf{A}\mathbf{w}, \mathbf{w}) &\equiv \sum_{i,j=1}^{n} A_{i,j} w_i w_j = E\{|\sigma'((\mathbf{w}^0, \mathbf{x}^{(1)}))|^2 (\mathbf{x}^{(1)}, \mathbf{w})^2\} > 0, \\
(\mathbf{X}\mathbf{w}, \mathbf{w}) &\equiv \sum_{i,j=1}^{n} X_{i,j} w_i w_j = E\{\theta(2n - |\mathbf{x}^{(1)}|)(\mathbf{x}^{(1)}, \mathbf{w})^2\} > 0,
\end{aligned}$$

but the values of the constants δ_1 and δ_2 depend on the distribution of $\mathbf{x}^{(\mu)}$.

Proposition 2 *Denote*

$$\begin{aligned}
\lambda_1 &\equiv -\log \min_{|\mathbf{w}|=1} E\{\exp\{-|\sigma'((\mathbf{w}^0, \mathbf{x}^{(1)}))|^2 (\mathbf{x}^{(1)}, \mathbf{w})^2\}\} > 0, \\
\lambda_2 &\equiv -\log \min_{|\mathbf{w}|=1} E\{\exp\{-(\mathbf{x}^{(1)}, \mathbf{w})^2\}\} > 0.
\end{aligned} \tag{3.5}$$

Then the inequalities

$$\tilde{\mathbf{A}} \geq \frac{\lambda_1}{4}\mathbf{I}, \quad \mathbf{X} \geq \frac{\lambda_2}{4}\mathbf{I}, \tag{3.6}$$

$$\frac{1}{M}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|) \leq \frac{1}{2}. \tag{3.7}$$

hold with probability more than $(1 - e^{-M\,const})$.

Now we are ready to find the asymptotic expression of $\mathcal{L}_M$. Let us take

$$\begin{gathered} L \equiv \frac{16n|\mathbf{w}^0|}{d}, \quad C_1 \equiv \min_{|\lambda|\leq 2n(L+|\mathbf{w}^0|)} \sigma'(\lambda), \\ C_2 \equiv \min\{|\sigma(4n|\mathbf{w}^0|) - \sigma(2n|\mathbf{w}^0|)|; \quad |\sigma(-4n|\mathbf{w}^0|) - \sigma(-2n|\mathbf{w}^0|)|\} \end{gathered} \tag{3.8}$$

and divide integrals in (2.6) in three parts:

$$\begin{gathered} (\int_{|\mathbf{w}-\mathbf{w}^0|<M^{-2/5}} + \int_{M^{-2/5}\leq|\mathbf{w}-\mathbf{w}^0|\leq L} + \int_{|\mathbf{w}-\mathbf{w}^0|\geq L})H(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})\exp\{-\beta H(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})\}\rho(\mathbf{w})d\mathbf{w} \\ = I_1 + I_2 + I_3. \end{gathered} \tag{3.9}$$

Let us estimate I_2 and I_3, assuming that inequalities (3.2), (3.6) and (3.7) hold. If $M^{-2/5} \leq |\mathbf{w}-\mathbf{w}^0| \leq L$ and $|\mathbf{x}^{(\mu)}| \leq 2n$, then $|(\mathbf{w},\mathbf{x}^{(\mu)})| \leq 2nL$ and

$$\begin{aligned} H(\mathbf{w};\overline{\mathbf{x}},0) &\geq \frac{1}{2}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|)(\sigma((\mathbf{w},\mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}^0,\mathbf{x}^{(\mu)})))^2 \\ &\geq \frac{C_1^2}{2}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|)((\mathbf{w},\mathbf{x}^{(\mu)}) - (\mathbf{w}^0,\mathbf{x}^{(\mu)}))^2 \\ &= \frac{C_1^2 M}{2}(\tilde{\mathbf{X}}(\mathbf{w}-\mathbf{w}^0),\mathbf{w}-\mathbf{w}^0) \geq \frac{C_1^2 M\lambda_2}{8}|\mathbf{w}-\mathbf{w}^0|^2 \geq \frac{C_1^2\lambda_2}{8}M^{1/5}. \end{aligned} \tag{3.10}$$

Thus, since evidently $|H(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})| \leq 2M$, we get

$$|I_2| \leq e^{-\,\mathrm{const}\,M^{2/3}}. \tag{3.11}$$

Consider now the domain $|\mathbf{w}| \geq L$. Let us note, that if $|\mathbf{x}^{(\mu)}| \leq 2n$, then $|(\mathbf{w}^0,\mathbf{x}^{(\mu)})| \leq 2n|\mathbf{w}^0|$, and if $|(\mathbf{w},\mathbf{x}^{(\mu)})| \geq \frac{dL}{4} = 4n|\mathbf{w}^0|$, then $|(\sigma((\mathbf{w},\mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}^0,\mathbf{x}^{(\mu)})))| \geq C_2$. Therefore

$$\begin{aligned} H(\mathbf{w};\overline{\mathbf{x}},0) &\geq \frac{1}{2}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|)\theta(|(\mathbf{w},\mathbf{x}^{(\mu)})| - \frac{dL}{4})(\sigma((\mathbf{w},\mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}^0,\mathbf{x}^{(\mu)})))^2 \\ &\geq \frac{C_2^2}{2}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|)\theta(|(\mathbf{w},\mathbf{x}^{(\mu)})| - \frac{dL}{4}) \\ &\geq \frac{C_2^2}{2}\sum_{\mu=1}^{M}\theta(2n - |\mathbf{x}^{(\mu)}|)\theta(|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})| - \frac{d}{4}) \geq \frac{MC_2^2K}{4}, \end{aligned} \tag{3.12}$$

where the last inequality is due to Proposition 1. Thus, similarly to (3.11),

$$|I_3| \leq e^{-\,\mathrm{const}\,M}. \tag{3.13}$$

Now let us calculate the first integral. Since

$$\sigma((\mathbf{w},\mathbf{x}^{(\mu)}))-\sigma((\mathbf{w}^0,\mathbf{x}^{(\mu)}))=\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu}))(\mathbf{w}-\mathbf{w}^0,\mathbf{x}^{(\mu)})$$
$$+\frac{\sigma''((\mathbf{w}^0,\mathbf{x}^{(\mu)}))}{2}(\mathbf{w}-\mathbf{w}^0,\mathbf{x}^{(\mu)})^2++\frac{\sigma'''((\zeta^{(\mu)}(\mathbf{w}),\mathbf{x}^{(\mu)}))}{6}(\mathbf{w}-\mathbf{w}^0,\mathbf{x}^{(\mu)})^3$$

with $\zeta^{(\mu)}(\mathbf{w})$ situated between $\mathbf{w}0$ and $\mathbf{w}$, we have

$$H_M(\mathbf{w};\overline{\mathbf{x}},0)=\frac{M}{2}(\tilde{\mathbf{A}}(\mathbf{w}-\mathbf{w}^0),\mathbf{w}-\mathbf{w}^0)+M\sum_{i,j,k=1}^{n}\tilde{D}^{(M)}_{i,j,k}(w_i-w_i^0)(w_j-w_j^0)(w_k-w_k^0)+Mf^{(M)}(\mathbf{w}),$$

where

$$\tilde{D}^{(M)}_{i,j,k}\equiv\frac{1}{M}\sum_{\mu=1}^{M}\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu}))\sigma''((\mathbf{w}^0,\mathbf{x}^{(\mu)}))x_i^{(\mu)}x_j^{(\mu)}x_k^{(\mu)}$$
$$f^{(M)}(\mathbf{w})\equiv\frac{1}{M}\sum_{\mu=1}^{M}(\frac{1}{4}(\sigma''((\mathbf{w}^0,\mathbf{x}^{(\mu)})))^2+\frac{1}{6}\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu}))\sigma'''((\zeta^{(\mu)}(\mathbf{w}),\mathbf{x}^{(\mu)})))(\mathbf{w}-\mathbf{w}^0,\mathbf{x}^{(\mu)})^4.$$

Performing the change of variables $\tilde{\mathbf{w}}=\sqrt{M}(\mathbf{w}-\mathbf{w}^0)$, we get then

$$H_M(\tilde{\mathbf{w}};\overline{\mathbf{x}},\overline{\eta})=\frac{1}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})+\frac{1}{\sqrt{M}}\sum_{i,j,k=1}^{n}\tilde{D}^{(M)}_{i,j,k,l}\tilde{w}_i\tilde{w}_j\tilde{w}_k+\frac{1}{M}f^{(M)}(\tilde{\mathbf{w}}),\tag{3.14}$$

Besides,

$$\rho(\mathbf{w}^0+M^{-1/2}\tilde{\mathbf{w}})=\rho(\mathbf{w}^0)(1+\frac{1}{M^{1/2}}(\mathbf{r},\tilde{\mathbf{w}})+\frac{1}{2M}(\mathbf{S}(\zeta(\mathbf{w}))\tilde{\mathbf{w}},\tilde{\mathbf{w}}))),\tag{3.15}$$

where $\mathbf{r}=\frac{\nabla\rho(\mathbf{w}^0)}{\rho(\mathbf{w}^0)}$ and $\mathbf{S}$ is the matrix of the second derivatives of the function $\rho(\mathbf{w})$ divided by $\rho(\mathbf{w}^0)$, and the point $\zeta(\mathbf{w})$ is situated between $\mathbf{w}^0$ and $\mathbf{w}$.

Then, for $|\tilde{\mathbf{w}}|=\sqrt{M}|\mathbf{w}-\mathbf{w}^0|\le M^{1/10}$ under the conditions (3.6) and using the simple relation $e^x=1+x+O(x^2)$, we obtain

$$\begin{aligned}I_1=\int_{|\tilde{\mathbf{w}}|\le M^{1/10}}&\rho(\mathbf{w}^0)[1+\frac{1}{M^{1/2}}(\mathbf{r},\tilde{\mathbf{w}})+\frac{1}{2M}(\mathbf{S}\tilde{\mathbf{w}},\tilde{\mathbf{w}})]\cdot\\ &[\frac{1}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})+\frac{1}{\sqrt{M}}\sum_{i,j,k=1}^{n}\tilde{D}^{(M)}_{i,j,k,l}\tilde{w}_i\tilde{w}_j\tilde{w}_k++\frac{1}{M}f^{(M)}(\tilde{\mathbf{w}})]\cdot\\ \exp\{-\frac{\beta}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}&[1-\frac{\beta}{\sqrt{M}}\sum_{i,j,k=1}^{n}\tilde{D}^{(3)}_{i,j,k,l}\tilde{w}_i\tilde{w}_j\tilde{w}_k-\frac{\beta}{M}f^{(M)}(\tilde{\mathbf{w}})+O(\frac{|\mathbf{w}|^4}{M})].\end{aligned}\tag{3.16}$$

Using the bounds, valid for any matrix $\mathbf{B}$, satisfying inequality $\mathbf{B}\ge\tilde{\delta}\mathbf{I}$

$$\begin{aligned}\int_{|\tilde{\mathbf{w}}|\ge M^{1/10}}\exp\{-(\mathbf{B}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}d\mathbf{w}&\le e^{-\frac{\tilde{\delta}}{2}M^{1/5}}\\ \int_{|\tilde{\mathbf{w}}|\ge M^{1/3}}|\tilde{\mathbf{w}}|^s\exp\{-(\mathbf{B}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}d\mathbf{w}&\le\text{const}\,e^{-\frac{\tilde{\delta}}{2}M^{1/5}},\quad(s=2,...15),\end{aligned}\tag{3.17}$$

we can extend the integration in the r.h.s. of (3.16) to the whole $\mathbf{R}^n$. Thus, if inequalities (3.2) and (3.6) hold, then by the rules of Gaussian integration we get

$$I_1=\frac{\rho(\mathbf{w}^0)}{2\beta M}\frac{\det^{-1/2}\tilde{\mathbf{A}}}{(2\pi\beta)^{n/2}}\sum_{i,j=1}^{n}\tilde{\mathbf{A}}_{i,j}(\tilde{\mathbf{A}}^{-1})_{i,j}(1+O(\frac{1}{M}))=\frac{n\rho(\mathbf{w}^0)}{2\beta M}\frac{\det^{-1/2}\tilde{\mathbf{A}}}{(2\pi\beta)^{n/2}}(1+O(\frac{1}{M})).\tag{3.18}$$

Calculating by the same way Z_M, we get

$$Z_M = \rho(\mathbf{w}^0)\frac{\det^{-1/2}\tilde{\mathbf{A}}}{(2\pi\beta)^{n/2}}(1+O(\frac{1}{M})) + O(e^{-\operatorname{const} M^{2/3}}) + O(e^{-\operatorname{const} M}).$$

Combining with (3.18), we obtain

$$\mathcal{L}_{\mathcal{M}} = \frac{n}{2\beta M} + O(e^{-\operatorname{const} M^{2/3}}) + O(e^{-\operatorname{const} M}). \tag{3.19}$$

Now denote by Ω the set of $\mathbf{x}^{(\mu)}$, for which inequalities (3.2) and (3.6) hold. It is easy to see, that

$$\mathrm{Prob}\{\overline{\Omega}\} \equiv E\{1-\chi_\Omega(\overline{\mathbf{x}})\} \le e^{-\operatorname{const} M}, \tag{3.20}$$

where $\chi_\Omega(\overline{\mathbf{x}})$ is the indicator function of the set Ω. Thus, since for all $\overline{\mathbf{x}}$ $|\mathcal{L}_M| \le \max_{\mathbf{w}} H_M(\mathbf{w};\overline{\mathbf{x}},\overline{\eta}) \le 2M$ we have

$$|E\{\mathcal{L}_{\mathcal{M}}(1-\chi_\Omega(\overline{\mathbf{x}}))\}| \le ME\{(1-\chi_\Omega(\overline{\mathbf{x}}))\} \le e^{-\operatorname{const} M}$$

$$E\{\mathcal{L}_{\mathcal{M}}\} = E\{\mathcal{L}_{\mathcal{M}}\chi_\Omega(\overline{\mathbf{x}})\} + E\{\mathcal{L}_{\mathcal{M}}(1-\chi_\Omega(\overline{\mathbf{x}}))\} = \frac{n}{2\beta M} + O(e^{-\operatorname{const} M^{2/3}}) + O(e^{-\operatorname{const} M}). \tag{3.21}$$

To find $\mathcal{G}_M$ we again divide all integrals into three parts and estimate I_2' and I_3' by the same way as in (3.10)-(3.13). To calculate I_1', we again perform the change of variables $\tilde{\mathbf{w}} = \sqrt{M}(\mathbf{w}-\mathbf{w}^0)$ and then obtain

$$\begin{gathered}H_{M+1}(\tilde{\mathbf{w}};\overline{\mathbf{x}},0) = \frac{1}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}}) + \frac{1}{2M}(\tilde{\mathbf{A}}^{(M+1)}\tilde{\mathbf{w}},\tilde{\mathbf{w}}) \\ +\frac{1}{\sqrt{M}}\sum_{i,j,k=1}^{n}\tilde{D}^{(M)}_{i,j,k,l}\tilde{w}_i\tilde{w}_j\tilde{w}_k + \frac{1}{M}f^{(M)}(\tilde{\mathbf{w}}) + O(\frac{|\tilde{\mathbf{w}}|^5}{M^{3/2}}),\end{gathered} \tag{3.22}$$

where

$$\tilde{A}^{(M+1)}_{i,j} \equiv |\sigma'((\mathbf{w}^0,\mathbf{x}^{(M+1)}))|^2 x_i^{(M+1)}x_j^{(M+1)}.$$

Then, similarly to (3.16)-(3.18) we get

$$\begin{gathered}Z_{M+1} = \int_{|\tilde{\mathbf{w}}|\le M^{1/3}} \rho(\mathbf{w}^0)\exp\{-\frac{\beta}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}[1+\frac{1}{M^{1/2}}(\mathbf{r},\tilde{\mathbf{w}}) + \frac{1}{2M}(\mathbf{S}\tilde{\mathbf{w}},\tilde{\mathbf{w}})] \\ (1-\frac{\beta}{2M}(\tilde{\mathbf{A}}^{(M+1)}\tilde{\mathbf{w}},\tilde{\mathbf{w}}) - \frac{\beta}{\sqrt{M}}\sum_{i,j,k=1}^{n}\tilde{D}^{(M)}_{i,j,k,l}\tilde{w}_i\tilde{w}_j\tilde{w}_k - \frac{\beta}{M}f^{(M)}(\tilde{\mathbf{w}}) + \frac{1}{M^{3/2}}O(|\tilde{\mathbf{w}}|^5))d\mathbf{w}.\end{gathered} \tag{3.23}$$

Using again estimates (3.17) we can extend integration here to the whole space and, denoting

$$\begin{gathered}\tilde{f}^{(M)} \equiv \frac{\int \exp\{-\frac{\beta}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}f^{(M)}(\tilde{\mathbf{w}})d\tilde{\mathbf{w}}}{\det^{1/2}\tilde{\mathbf{A}}(2\pi\beta)^{n/2}}, \\ \tilde{S} \equiv \frac{\int \exp\{-\frac{\beta}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})\}(\mathbf{S}\tilde{\mathbf{w}},\tilde{\mathbf{w}})d\mathbf{w}}{\det^{1/2}\tilde{\mathbf{A}}(2\pi\beta)^{n/2}},\end{gathered}$$

we get from (3.23)

$$Z_{M+1} = \rho(\mathbf{w}^0)\frac{\det^{-1/2}\tilde{\mathbf{A}}}{(2\pi\beta)^{n/2}}(1-\frac{\beta}{2M}\sum_{i,j=1}^{n}\tilde{A}^{(M+1)}_{i,j}(\tilde{A}^{-1})_{i,j} - \frac{\beta}{M}\tilde{f}^{(M)} + \frac{\tilde{S}}{2M} + O(\frac{1}{M^{3/2}})) \tag{3.24}$$

and

$$Z_M = \rho(\mathbf{w}^0)\frac{\det^{-1/2}\tilde{\mathbf{A}}}{(2\pi\beta)^{n/2}}(1-\frac{\beta}{M}\tilde{f}^{(M)} + \frac{\tilde{S}}{2M} + O(\frac{1}{M^{3/2}})). \tag{3.25}$$

So we derive that if inequalities (3.2), (3.6) are true

$$\mathcal{G}_M(0) = \frac{\beta}{2M} \sum_{i,j=1}^{n} \tilde{A}_{i,j}^{(M+1)} (\tilde{A}^{-1})_{i,j} + o(\frac{1}{M}). \tag{3.26}$$

Now, let us note, that since $\tilde{A}_{i,j} \to A_{i,j}$ for almost all $\overline{\mathbf{x}}$, we have under condition (3.6), that $(\tilde{A}^{-1})_{i,j} \to (A^{-1})_{i,j}$. Therefore, for these $\overline{\mathbf{x}}$

$$\mathcal{G}_M(0) = \frac{\beta}{2M} \sum_{i,j=1}^{n} \tilde{A}_{i,j}^{(M+1)} (A^{-1})_{i,j} + o(\frac{1}{M}). \tag{3.27}$$

Now, we use again the fact that inequalities (3.2) and (3.6) are true with probability more then $1 - e^{-\operatorname{const} M}$. Thus, due to the bound $|\mathcal{G}_M| \leq |\ln \exp\{-2\beta\}| \leq \operatorname{const}$, valid for all $\overline{\mathbf{x}}$, we can, similarly to (3.20)-(3.21), get from (3.24)

$$\begin{aligned} \mathcal{G}_M(0) &= \frac{\beta}{2M} \sum_{i,j=1}^{n} E\{\tilde{A}_{i,j}^{(M+1)}\} (A^{-1})_{i,j} + o(\frac{1}{M}) \\ &= \frac{\beta}{2M} \sum_{i,j=1}^{n} A_{i,j} (A^{-1})_{i,j} + o(\frac{1}{M}) = \frac{\beta n}{2M} + o(\frac{1}{M}). \end{aligned} \tag{3.28}$$

To study the case $t \neq 0$ we have to add some other conditions to (3.2) and (3.6) . Let $L^*(M)$ be the minimal positive number such that

$$1 - |\sigma(L^*(M))| \leq \frac{1}{M}. \tag{3.29}$$

It follows from the condition (2.1), that $L^*(M) \leq \operatorname{const} M$. We choose also some M-independent number $\tilde{\varepsilon} > 0$ such that

$$2 \max_{|\mathbf{w}|=1} E\{\theta(2\tilde{\varepsilon} - |(\mathbf{x}^{(1)}, \mathbf{w})|)\theta(\sqrt{M} - |\mathbf{x}^{(1)}|)\} E\{|\eta^{(1)}|\} \leq \frac{KC_2^2}{10t}, \tag{3.30}$$

where the constant K is defined in (3.1) and C_2 is defined in (3.8). Such a choice is always possible because the distribution of $\mathbf{x}^{(1)}$ has no singularity at the point $\mathbf{x}^{(1)} = 0$. Consider the system of N_1 points ($N_1 \leq \operatorname{const} M^{2n}$) $\{\mathbf{w}^{(*i)}\}_{i=1}^{N_1}$, such that $|\mathbf{w}^{(*i)}| \leq L^* \tilde{\varepsilon}$ and

$$\mathcal{B}(0, \frac{L^*}{\tilde{\varepsilon}}) \subset \cup_{i=1}^{N_1} \mathcal{B}(\mathbf{w}^{(*i)}, M^{-2}).$$

We denote

$$R(\mathbf{w}; \overline{\mathbf{x}}, \overline{\eta}) \equiv \sum_{\mu=1}^{M} (\sigma((\mathbf{w}^0, \mathbf{x}^{(\mu)})) - \sigma((\mathbf{w}, \mathbf{x}^{(\mu)}))) \eta^{(\mu)} \tag{3.31}$$

and assume that the following inequalities are true

$$\begin{aligned} &\sum_{\mu=1}^{M} \theta(|\mathbf{x}^{(\mu)}| - \sqrt{M}) |\eta^{(\mu)}| \leq M^{1/6}, \quad \frac{1}{M} \sum_{\mu=1}^{M} |\eta^{(\mu)}| \leq M^{1/8}, \\ &\frac{2}{M} \sum_{\mu=1}^{M} \theta(2\tilde{\varepsilon} - |(\mathbf{x}^{(\mu)}, \mathbf{w}^{(*i)})|) \theta(\sqrt{M} - |\mathbf{x}^{(\mu)}|) |\eta^{(\mu)}| \leq \frac{KC_2^2}{9t}, \\ &\frac{1}{\sqrt{M}} |R(\mathbf{w}^{(*i)}; \overline{\mathbf{x}}, \overline{\eta})| \leq M^{1/8}, \quad (i = 1, \ldots, N_1). \end{aligned} \tag{3.32}$$

Proposition 3 *The probability that all inequalities (3.32) hold is more than* $1 - e^{-M^{1/6} \mathit{const}}$.

Proposition 4 *If inequalities (3.2), (3.6), (3.7) and (3.32) are fulfilled, then*

$$\inf_{|\mathbf{w}-\mathbf{w}^0|\ge M^{-2/5}}[H(\mathbf{w};\overline{\mathbf{x}},0)+tR(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})]\ge \mathit{const}M^{1/5}. \tag{3.33}$$

By using this proposition, it is easy to obtain that similarly to the case $t=0$, for our purposes it is enough to study only the integral inside the domain $|\mathbf{w}-\mathbf{w}^0|\le M^{-2/5}$. In this domain, using again the variables $\tilde{\mathbf{w}}=\sqrt{M}(\mathbf{w}-\mathbf{w}^0)$, we get

$$H_M(\tilde{\mathbf{w}};\overline{\mathbf{x}},0)=Y+\frac{1}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})+(\mathbf{b},\tilde{\mathbf{w}})+f_M(\tilde{\mathbf{w}},\overline{\eta}), \tag{3.34}$$

$$\begin{aligned}H_{M+1}(\tilde{\mathbf{w}};\overline{\mathbf{x}},\overline{\eta})&=Y+\frac{1}{2}(\tilde{\mathbf{A}}\tilde{\mathbf{w}},\tilde{\mathbf{w}})+(\mathbf{b},\tilde{\mathbf{w}})+f^{(M)}(\tilde{\mathbf{w}},\overline{\eta})+\frac{t^2(\eta^{(M+1)})^2}{2}\\&+\frac{1}{2M}(\tilde{\mathbf{A}}^{(M+1)}\tilde{\mathbf{w}},\tilde{\mathbf{w}})+\frac{t\eta^{(M+1)}}{\sqrt{M}}\sigma'((\mathbf{w}^0,\mathbf{x}^{(M+1)}))(\tilde{\mathbf{w}},\mathbf{x}^{(M+1)})+\frac{1}{M}g^{(M+1)}(\tilde{\mathbf{w}},\mathbf{x}^{(M+1)}).\end{aligned} \tag{3.35}$$

Here

$$Y=\frac{t^2}{2}\sum_{\mu=1}^{M}(\eta^{(\mu)})^2,\quad \mathbf{b}=\frac{t}{\sqrt{M}}\sum_{\mu=1}^{M}\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu}))\eta^{(\mu)}\mathbf{x}^{(\mu)}. \tag{3.36}$$

The explicit form of the functions $f^{(M)}(\tilde{\mathbf{w}},\overline{\eta})$ and $g^{(M+1)}(\tilde{\mathbf{w}},\mathbf{x}^{(M+1)})$ is not important for us because, if you have the fraction of the form

$$F(M)=\frac{1+f(M)+\frac{A}{M}}{1+f(M)}$$

and we know that $f(M)\to 0$, as $M\to\infty$, then we can write

$$F(M)=1+\frac{A}{M}\frac{1}{1+f(M)}=1+\frac{A}{M}+O(\frac{1}{M}f(M))=1+\frac{A}{M}+o(\frac{1}{M}).$$

That is why we need not to know the exact form of $f(M)$. In fact, we have used already this trick, calculating $\mathcal{G}(0)$ in formulae (3.24)- (3.26), but there all the calculations were shown explicetely.

Therefore we get that, if inequalities (3.2), (3.6) and (3.32) hold, then

$$\mathcal{L}_M(t)=\frac{1}{M}Y+\frac{1}{2M}((\tilde{\mathbf{A}})^{-1}\mathbf{b},\mathbf{b})+O(M^{-1}) \tag{3.37}$$

$$\mathcal{G}_M(t)=-\ln(\exp\{-\frac{\beta t^2(\eta^{(M+1)})^2}{2}\}(1-\frac{1}{2\beta M}(\tilde{\mathbf{A}}^{(M+1)}\tilde{\mathbf{A}}^{-1}\mathbf{b},\tilde{\mathbf{A}}^{-1}\mathbf{b})+O(M^{-1})). \tag{3.38}$$

Taking into account, that according to Proposition 3 inequalities (3.32) hold with probability larger than $1-\exp\{-\operatorname{const}M^{1/4}\}$, and inequalities (3.2), (3.6) and (3.7) hold with probability larger than $1-\exp\{-\operatorname{const}M\}$, similarly to (3.20)-(3.21) we obtain the second line of (2.8). Theorem 1 is proven.

4 Auxiliary results

Proof of Proposition 1

Due to the choice of $\{\mathbf{w}^{(i)}\}$ for any $\mathbf{w}$ one can find a $\mathbf{w}^{(i)}$ such that $|\mathbf{w}-\mathbf{w}^{(i)}|\le\varepsilon$. Thus, if $|\mathbf{x}^{(\mu)}|\le 2n$ and $|(\mathbf{w}^{(i)},\mathbf{x}^{(\mu)})|\ge\frac{d}{2}$, then

$$|(\mathbf{w},\mathbf{x}^{(\mu)})|\ge|(\mathbf{w}^{(i)},\mathbf{x}^{(\mu)})|-|\mathbf{x}^{(\mu)}||\mathbf{w}-\mathbf{w}^{(i)}|\ge\frac{d}{2}-2n\varepsilon=\frac{d}{4}$$

Therefore

$$\theta(|(\mathbf{w},\mathbf{x}^{(\mu)})|-\frac{d}{4})\ge\theta(|(\mathbf{w}^{(i)},\mathbf{x}^{(\mu)})|-\frac{d}{2})$$

and so

$$\sum_{\mu=1}^{M}\theta(|(\mathbf{w},\mathbf{x}^{(\mu)})|-\frac{d}{4})\theta(2n-|\mathbf{x}^{(\mu)}|)\geq\sum_{\mu=1}^{M}\theta(|(\mathbf{w}^{(i)},\mathbf{x}^{(\mu)})|-\frac{d}{2})\theta(2n-|\mathbf{x}^{(\mu)}|)\geq\frac{KM}{2}$$

Proposition 1 is proven.

Proof of Proposition 2

The estimate for the probability of inequality (3.7) follows from the standard form of the Chebyshev inequality for the set of bounded independent random variables $\theta(2n-|\mathbf{x}^{(\mu)}|)$, if their mean values are larger than $\frac{1}{2}$. But since

$$E\{\theta(|\mathbf{x}^{(\mu)}|-2n)\}\leq E\{\theta(|\mathbf{x}^{(\mu)}|-2n)\frac{|\mathbf{x}^{(\mu)}|^2}{(2n)^2}\}\leq E\{\frac{|\mathbf{x}^{(\mu)}|^2}{(2n)^2}\}=\frac{1}{4n},$$

we have that

$$E\{\theta(2n-|\mathbf{x}^{(\mu)}|)\}=1-E\{\theta(|\mathbf{x}^{(\mu)}|-2n)\}\geq 1-\frac{1}{4n}\geq\frac{3}{4}.$$

Thus, we get easily that (3.7) holds with probability more than $(1-e^{-M\,\mathrm{const}})$.

To prove (3.7) let us prove first that for any $\mathbf{w}:|\mathbf{w}|=1$

$$\mathrm{Prob}\,\{(\tilde{\mathbf{A}}\mathbf{w},\mathbf{w})\leq\frac{\lambda_1}{2}\}\leq e^{-M\frac{\lambda_1}{2}}. \tag{4.1}$$

From the Chebyshev inequality we get

$$\begin{aligned}\mathrm{Prob}\,\{(\tilde{\mathbf{A}}\mathbf{w},\mathbf{w})\leq\frac{\lambda_1}{2}\}&\leq E\{\exp\{M\frac{\lambda_1}{2}-M(\tilde{\mathbf{A}}\mathbf{w},\mathbf{w})\}\}\\ =e^{-M\lambda_1/2}&\prod_{\mu+1}^{M}E\{\exp\{-|\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu}))|^2(\mathbf{x}^{(\mu)},\mathbf{w})^2\}\}\}\\ =e^{-M\lambda_1/2}&(E\{\exp\{-|\sigma'((\mathbf{w}^0,\mathbf{x}^{(1)}))|^2(\mathbf{x}^{(1)},\mathbf{w})^2\}\})^M\\ &\leq e^{-M\lambda_1/2}e^{-M\lambda_1}=e^{-M\lambda_1/2}.\end{aligned} \tag{4.2}$$

Now denote $C\equiv\max_{\lambda}|\sigma'(\lambda)|$ and take the finite M-independent set of the points $\mathbf{w}^{(1i)}$ such that

$$\mathcal{S}(0,1)\subset\cup\mathcal{B}(\mathbf{w}^{(1i)},\frac{\lambda_1}{16C^2n^2}).$$

Suppose that inequalities (3.7) and (4.1) for all points $\mathbf{w}^{(1i)}$ are valid. Since for any $\mathbf{w}:|\mathbf{w}|=1$ there exists the point $\mathbf{w}^{(1i)}$ such that $|\mathbf{w}-\mathbf{w}^{(1i)}|\leq\frac{\lambda_1}{16Cn^2}$ we get

$$\begin{aligned}(\tilde{\mathbf{A}}\mathbf{w},\mathbf{w})&=(\tilde{\mathbf{A}}\mathbf{w}^{(1i)},\mathbf{w}^{(1i)})+((\tilde{\mathbf{A}}\mathbf{w},\mathbf{w})-(\tilde{\mathbf{A}}\mathbf{w}^{(1i)},\mathbf{w}^{(1i)}))\\ &=(\tilde{\mathbf{A}}\mathbf{w}^{(1i)},\mathbf{w}^{(1i)})-\frac{1}{M}\sum_{\mu=1}^{M}(\sigma'((\mathbf{w}^0,\mathbf{x}^{\mu})))^2(\mathbf{w}-\mathbf{w}^{(1i)},\mathbf{x}^{(\mu)})(\mathbf{w}+\mathbf{w}^{(1i)},\mathbf{x}^{(\mu)})\\ &\geq\frac{\lambda_1}{2}-\frac{1}{M}\sum_{\mu=1}^{M}2C^2|\mathbf{w}-\mathbf{w}^{(1i)}|\cdot|\mathbf{x}^{(\mu)}|^2\geq\frac{\lambda_1}{2}-\frac{2C\lambda_1}{16MCn^2}\sum_{\mu=1}^{M}(2n)^2\theta(2n-|\mathbf{x}^{(\mu)}|)\\ &\geq\frac{\lambda_1}{2}-\frac{\lambda_1}{2M}\sum_{\mu=1}^{M}\theta(2n-|\mathbf{x}^{(\mu)}|)\geq\frac{\lambda_1}{4}.\end{aligned}$$

The proof for the matrix $\tilde{\mathbf{X}}$ is the same.

Proof of Proposition 3

Taking $\lambda > 0$ so that condition (2.3) is fulfilled and using the Chebyshev inequality, we get that

$$\begin{aligned}\mathrm{Prob}\,\{\sum_{\mu=1}^{M}\theta(|\mathbf{x}^{(\mu)}|-\sqrt{M})|\eta^{(\mu)}|\geq\sqrt{M}\} &\leq e^{-\lambda M^{1/6}}E\{\prod_{\mu=1}^{M}\exp\{\lambda\theta(|\mathbf{x}^{(\mu)}|-\sqrt{M})|\eta^{(\mu)}|\}\}\\ &= e^{-\lambda M^{1/6}}(1-p_M+p_M E\{e^{\lambda|\eta^{(1)}|}\})^M \leq \mathrm{const}\, e^{-\lambda M^{1/6}}.\end{aligned}$$

Here $p_M \equiv \mathrm{Prob}\,\{|\mathbf{x}^{(\mu)}| \geq \sqrt{M}\}$ and $p_M \leq nM^{-1}$ because of condition (2.2). The estimates of the probabilities for the rest of inequalities in (3.32) are similar.

Proof of Proposition 4

If $|\mathbf{w}| \leq \frac{L^*}{\tilde{\varepsilon}}$, then we choose the point $\mathbf{w}^{(*i)}$ such that $|\mathbf{w}-\mathbf{w}^{(*i)}| \leq \frac{1}{M^2}$. Then

$$\begin{aligned}|(R(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})-R(\mathbf{w}^{(*i)};\overline{\mathbf{x}},\overline{\eta}))| &\leq 2\sum_{\mu=1}^{M}\theta(|\mathbf{x}^{(\mu)}|-\sqrt{M})|\eta^{(\mu)}|+C\sum_{\mu=1}^{M}\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\mathbf{x}^{(\mu)}||\mathbf{w}^{(*i)}-\mathbf{w}||\eta^{(\mu)}|\\ &\leq \sum_{\mu=1}^{M}\theta(|\mathbf{x}^{(\mu)}|-\sqrt{M})|\eta^{(\mu)}|+\frac{C\sqrt{M}}{M^2}\sum_{\mu=1}^{M}|\eta^{(\mu)}| \leq 2M^{1/6}+O(1),\end{aligned}$$

where C is the upper bound for the first derivative of the function σ, and we have used the first line of (3.32). Then, using the last line of (3.32), we get for $|\mathbf{w}| \leq \frac{L^*}{\tilde{\varepsilon}}$

$$|R(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})| \leq \mathrm{const}\, M^{1/6}. \tag{4.3}$$

Since $M^{1/6} << M^{1/5}$ we derive the statement of Proposition 4 from this inequality and estimate (3.10).

If $|\mathbf{w}| > \frac{L^*}{\tilde{\varepsilon}}$, then we consider

$$\begin{aligned}|R(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})-R(\frac{L^*}{\tilde{\varepsilon}}\cdot\frac{\mathbf{w}}{|\mathbf{w}|};\ \overline{\mathbf{x}},\overline{\eta})| \leq 2\sum_{\mu=1}^{M}\theta(|\mathbf{x}^{(\mu)}|-\sqrt{M})|\eta^{(\mu)}|\\ +\sum_{\mu=1}^{M}(\sigma((\mathbf{w},\mathbf{x}^{(\mu)}))-\sigma(\frac{L^*}{\tilde{\varepsilon}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})))\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}| \equiv 2\Sigma_1+\Sigma_2.\end{aligned} \tag{4.4}$$

Because of (3.32) $2|\Sigma_1| \leq 2M^{1/6}$. Let us estimate Σ_2.

$$\begin{aligned}|\Sigma_2| \leq \sum_{\mu=1}^{M}|(\sigma((\mathbf{w},\mathbf{x}^{(\mu)}))-\sigma(\frac{L^*}{\tilde{\varepsilon}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)}))|\theta(|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})|-\tilde{\varepsilon})\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}|\\ +2\sum_{\mu=1}^{M}|\theta(\tilde{\varepsilon}-|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})|)\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}|\\ \leq \sum_{\mu=1}^{M}|\sigma((\mathbf{w}^0,\mathbf{x}^{(\mu)}))-\sigma(\frac{L^*}{\tilde{\varepsilon}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})|\theta(|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})|-\tilde{\varepsilon})|\eta^{(\mu)}|\\ +2\sum_{\mu=1}^{M}\theta(\tilde{\varepsilon}-|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})|)\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}| \equiv \Sigma_3+2\Sigma_4.\end{aligned} \tag{4.5}$$

But, if $|(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)})| \geq \tilde{\varepsilon}$, then

$$(\mathbf{w},\mathbf{x}^{(\mu)}) \geq L^*,\ \ \frac{L^*}{\tilde{\varepsilon}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)}) \geq L^*, \quad \mathit{or} \quad (\mathbf{w},\mathbf{x}^{(\mu)}) \leq -L^*,\ \ \frac{L^*}{\tilde{\varepsilon}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)}) \leq -L^*.$$

Then in both cases

$$|\sigma((\mathbf{w},\mathbf{x}^{(\mu)}))-\sigma(\frac{L^*}{\tilde{\mathbf{w}}}\cdot(\frac{\mathbf{w}}{|\mathbf{w}|},\mathbf{x}^{(\mu)}))|\leq\frac{1}{M}$$

because of the definition of L^* (see (3.29)). Therefore

$$\Sigma_3\leq\frac{1}{M}\sum_{\mu=1}^{M}|\eta^{(\mu)}|\leq M^{1/8}.$$

Let us take any point $\mathbf{w}$ such that $|\mathbf{w}|=1$. Then there exists the poit $\mathbf{w}^{(*i)}$ such that $|\mathbf{w}-\mathbf{w}^{(*i)}|\leq M^{-2}$. Then, if $|\mathbf{x}^{(\mu)}|\leq\sqrt{M}$, one can conclude that $|(\mathbf{w}^{(*i)},\mathbf{x}^{(\mu)})-(\mathbf{w},\mathbf{x}^{(\mu)})|\leq M^{-3/2}$ and therefore for M large enough ($M^{3/2}>\tilde{\varepsilon}^{-1}$) we have $|(\mathbf{w},\mathbf{x}^{(\mu)})|\geq|(\mathbf{w}^{(*i)},\mathbf{x}^{(\mu)})|-\tilde{\varepsilon}$. Thus,

$$\theta(\tilde{\varepsilon}-|(\mathbf{w},\mathbf{x}^{(\mu)})|)\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}|\leq\theta(2\tilde{\varepsilon}-|(\mathbf{w}^{(*i)},\mathbf{x}^{(\mu)})|)\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}|.$$

Summing this inequalities with respect to μ and using inequality (3.32), we get the estimate

$$\sup_{|\mathbf{w}|=1}2\sum_{\mu=1}^{M}\theta(\tilde{\varepsilon}-|(\mathbf{w},\mathbf{x}^{(\mu)})|)\theta(\sqrt{M}-|\mathbf{x}^{(\mu)}|)|\eta^{(\mu)}|\leq\frac{MKC_2^2}{9t}.$$

Therefore

$$2\Sigma_4\leq\frac{MKC_2^2}{9t}. \tag{4.6}$$

From (4.4)-(4.6) we get, that

$$|R(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})-R(\frac{L^*}{\tilde{\varepsilon}}\cdot\frac{\mathbf{w}}{|\mathbf{w}|};\ \overline{\mathbf{x}},\overline{\eta})|\leq\frac{MC_2^2K}{8t}+\text{const}\,M^{1/8}.$$

Now, using inequality (4.3) proved above, we have

$$|R(\frac{L^*}{\tilde{\varepsilon}}\cdot\frac{\mathbf{w}}{|\mathbf{w}|};\ \overline{\mathbf{x}},\overline{\eta})|\leq\text{const}\,M^{1/6},$$

Thus, we get that for $|\mathbf{w}|>\frac{L^*}{\tilde{\varepsilon}}$

$$|R(\mathbf{w};\overline{\mathbf{x}},\overline{\eta})|\leq\frac{MC_2^2K}{8t}.$$

But since from (3.12) we have that $H(w,0)\geq\frac{MC_2^2K}{4}$, we obtain for $|\mathbf{w}|>\frac{L^*}{\tilde{\varepsilon}}$ that

$$H(w,0)+tR(w)\geq\frac{MC_2^2K}{4}-|t||R(w)|\geq\frac{MC_2^2K}{4}-\frac{MC_2^2K}{8}=\frac{MC_2^2K}{8}.$$

Proposition 4 is proven.

References

[1] C.-P. Tsai, T.-L. Lee: Back propagation neural network in tidal- level forecasting. Journal of waterway, port, coastal and ocean engineering, July-August 1999, 195-199.

[2] G.Cimino, G.Del Duce, L.K.Kadonaga, G.Rotundo, A.Sisani, G.Stabile, B.Tirozzi, M.Whiticar: Time series analysis of geological data. Chemical Geology, **161**, 253-270, 1999.

[3] G.Rotundo, B.Tirozzi, M.Valente Neural networks for financial forecast. In: Proceedings of 6th European Symposium on Artificial Neural Networks. 351-356, Bruges, Belgium, 1998.

[4] V.N.Vapnik and A.Ya.Chervonenkis: On the uniform convergence of relative frequencies of events to their probabilities. Theor. Prob. and Appl. **16**, N2, 264-280, 1971.

[5] S.-I. Amari: A universal theorem of learning Curves. Neural Networks **6**, 161-166, 1993.

[6] J.Feng: Generalization error of the simple perceptron. Journal of Physics A: Mathematical and General **31**, N17, 4037-4048, 1998

[7] S.Sola, E.Levin : Learning in linear neural network: The validity of the annealed approximation. Phys.Rev.A, **46**, N4, 2124-2130