Asymptotic properties of one-layer artificial neural networks with sparse connectivity

Christian Hirsch

Department of Mathematics, Ny Munkegade 118, 8000 Aarhus C, Denmark

Matthias Neumann*, Volker Schmidt

Institute of Stochastics, Ulm University, Helmholtzstraße 18, 89069 Ulm, Germany

Abstract

A law of large numbers for the empirical distribution of parameters of a one-layer artificial neural networks with sparse connectivity is derived for a simultaneously increasing number of both, neurons and training iterations of the stochastic gradient descent.

Keywords: artificial neural network, law of large numbers, random network, sparse connectivity, stochastic gradient descent, weak convergence

2020 MSC: 60D05, 60G55, 68T07

1. Introduction

Artificial neural networks (ANNs) provide powerful tools for a data-driven gain of knowledge. The simplest architecture of an ANN is given by a *single-layer perceptron* (SLP), i.e., a feed-forward network with one layer and fully connected neurons. For *deep learning*, which relies on ANNs with multiple layers, fully connected layers are still indispensable [1]. However, connections between neurons in biological neural networks are typically sparse [2]. This inspired the development of ANNs with sparse connectivity between neurons, which exhibit – in terms of accuracy – the same quality as their fully connected counterparts [3]. In the present paper, we provide a theoretical analysis of SLPs with sparse connectivity, which are trained via stochastic gradient descent (SGD) [1]. By extending the methods considered in [4], we derive a law of large numbers

*corresponding author: matthias.neumann@uni-ulm.de

(LLN) for the empirical distribution of parameters for the asymptotic regime, where both, the number of neurons and training iterations of the SGD are simultaneously increasing. We consider a model with random sparsity [5], which is – in contrast to the adaptive approach considered in [3] – pre-defined before training [6]. Connections between input data and the different neurons in the hidden layer are removed independently. The considered model particularly covers the Erdős-Rényi graph, which serves as the initial state for the adaptive connectivity model in [3]. Section 2 defines our ANN model and gives the main results. Subsequently, in Section 3, the main results are illustrated by means of a simulation study. The rest of the paper is dedicated to the proofs, where we follow the basic idea of [4] and consider the development of the empirical distribution of the ANN-parameters as an element in an appropriately chosen Skorokhod space. Then, weak convergence of these objects in the asymptotic regime mentioned above is obtained by building on a blueprint that has already been successfully implemented in a variety of contexts such as those considered in [7, 8, 9]. More precisely, in our case, we show tightness of the sequence under consideration in the Skorokhod space (Section 5), uniqueness of the limit (Section 6) and identify the limit (Section 7).

2. Definitions and main results

As in [4], we investigate asymptotic properties of an SLP consisting of an input layer with $d \geqslant 1$ nodes and one hidden layer of $N \geqslant 1$ nodes. More precisely, let $x \in \mathbb{R}^d$ be the input vector, and $c^1, \dots, c^N \in \mathbb{R}$, $w^1, \dots, w^N \in \mathbb{R}^d$ be the weights of the SLP for the output and hidden layer, respectively. Denoting by $\boldsymbol{\theta} = (c^1, \dots, c^N, w^1, \dots, w^N) \in \mathbb{R}^{(1+d)N}$ the weight vector, the SLP $g(x, \boldsymbol{\theta})$ with parameter θ is defined by

$$g(x, \boldsymbol{\theta}) = \frac{1}{N} \sum_{i \leqslant N} c^i \sigma(x^\top w^i), \tag{1}$$

where we assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is a twice differentiable bounded function with bounded derivatives.

Formalizing the setup of [3], we modify the above SLP such that for $1 \leqslant i \leqslant N$, the ith node in the hidden layer is influenced only by a certain subset $\xi^i \subseteq \{1, \dots, d\}$ of the coordinates of the input vector. Thus, for each $1 \leq i \leq N$, we put those coordinates of w^i equal to 0 that do not belong to ξ^i. Depending on the application context, it may make sense to select ξ^i only from a subset of *admissible prunings* $\mathcal{C} \subset \{A : A \subseteq \{1, \dots, d\}\}$, which is fixed henceforth. An essential example corresponds to the setting, where the $\{\xi^i\}_{i \geqslant 1}$ are realizations of independent and identically distributed (iid) configurations $\{\Xi^i\}_{i \geqslant 1}$. For instance, in the simulation study described in Section 3, we consider *Erdős-Rényi pruning* with parameter $0 < p \leqslant 1$, where $\mathbb{P}(\Xi^1 = \xi) = p^{\#\xi}(1-p)^{d-\#\xi}$.

Now, let $\{(X_k, Y_k)\}_{k \geqslant 1}$ be a random sequence of iid training data, where for each $k \geq 1$, the random vector (X_k, Y_k) is a copy of a random vector $(X, Y) : \Omega \to \mathbb{R}^{d+1}$. Then, we train the SLP through SGD with respect to the squared-error loss function $(x, y) \mapsto (y - g(x, \theta))^2$ and learning rate $\alpha_N = \alpha/N$ for some $\alpha > 0$. More precisely, we initialize the network with random weights $\boldsymbol{\theta}_0$ and then iteratively update them via

$$\begin{aligned} c_{k+1}^i &= c_k^i + \frac{1}{N} g(X_k, Y_k, \boldsymbol{\theta}_k)\, \sigma(X_k^\top w_k^i), \\ w_{k+1}^i &= w_k^i + \frac{1}{N} g(X_k, Y_k, \boldsymbol{\theta}_k)\, c_k^i\, \sigma'(X_k^\top w_k^i) X_k(\xi^i), \end{aligned} \tag{2}$$

where $g(X_k, Y_k, \boldsymbol{\theta}_k) = \alpha(Y_k - g(X_k, \boldsymbol{\theta}_k))$ and $X_k(\xi^i)$ denotes the modification of X_k with entries of X_k outside ξ^i set to 0.

The main result of the present paper describes the evolution of the parameter $\boldsymbol{\theta}$ if the number of SGD iterations is of order N. Our key innovation to the analysis in comparison to [4] is that due to the recursion given in (2), where weights corresponding to different ξ^i evolve differently. Hence, when understanding the evolution over time, these groups of weights need to be separated. As a result, we obtain a quenched LLN.

The main idea to arrive at the quenched LLN is to choose a tailormade state space that allows for a smooth extension of the argument used in [4]. More precisely, let $S_\xi = \mathbb{R}^{1+d}$ be a separate copy of $\mathbb{R}^{1+d}$ for each $\xi \in \mathcal{C}$, and let $S = \bigsqcup_{\xi \in \mathcal{C}} S_\xi$, be the disjoint union of these copies. In this set-up the ith weight vector θ^i is considered to be embedded inside $S_{\xi^i} \subseteq S$. Moreover, a function $f : S \to \mathbb{R}$ corresponds to a collection of functions $f = \{f_\xi\}_{\xi \in \mathcal{C}}$ defined on each S_ξ. For each $\xi \in \mathcal{C}$, a probability measure μ on S defines a probability measure on S_ξ via $\mu_\xi(\cdot) = \mu(\cdot)/\mu(S_\xi)$.

In this interpretation, we let $\nu_k^N = \frac{1}{N} \sum_{i \leqslant N} \delta_{\theta_k^i}$ denote the empirical measure of the weights after $k \geqslant 1$ iterations. In particular, ν_k^N is a random element in the space $\mathcal{M}(S)$ of probability measures on S. We interpret $g(X_k, \nu_k^N) = \langle g(X_k, \cdot), \nu_k^N \rangle = \int_S g(X_k, \theta) \nu_k^N(\mathrm{d}\theta)$ as the integration of the function $g(X_k, \cdot) \colon \mathbb{R}^{1+d} \to \mathbb{R}$, $(c, w) \mapsto c\sigma(X_k^\top w)$ with respect to ν_k^N. A similar remark holds for $g(X_k, Y_k, \nu_k^N)$. Then, we show that as $N \to \infty$, the time-rescaled measure $\mu_t^N = \nu_{\lfloor Nt \rfloor}^N$ converges to the solution of an evolution equation described in (4) below. We think of μ^N as a random element in the Skorokhod space $D([0,T], \mathcal{M}(S))$. We fix $p_\xi > 0$, $\xi \in \mathcal{C}$ with $\sum_{\xi \in \mathcal{C}} p_\xi = 1$ and assume that

(E) $\lim_{N\to\infty} \frac{1}{N} \#\{i \leqslant N : \xi^i = \xi\} = p_\xi$ (ergodicity condition),

(M) the random sequences of the initial parameters $\{c_0^i\}_{i \leqslant N}$ and $\{w_0^i\}_{i\leqslant N}$ are both iid, independent of each other, and satisfy $\mathbb{E}[\exp(q|c_0^i|) + |w_0^i|^4] < \infty$ for some $q > 0$. Moreover, $\mathbb{E}[|X|^6 + Y^6] < \infty$ (moment condition).

Theorem 1 (Quenched LLN)**.** *Under the conditions* **(E)** *and* **(M)***, the limit trajectory* $\bar\mu_\cdot = \lim_{N\to\infty} \mu_\cdot^N$ *exists and decomposes as*

$$\bar\mu_t = \sum_{\xi \in \mathcal{C}} p_\xi \bar\mu_{t,\xi}. \tag{3}$$

Moreover, for each $f \in C_b^2(S)$*, the trajectory* $\{\bar\mu_t\}_{t \leqslant T}$ *satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle f, \bar\mu_t \rangle = \langle A(\cdot; \bar\mu_t)\nabla f, \bar\mu_t \rangle \tag{4}$$

with $A(\theta; \bar\mu_t) = \big(A_{\mathsf{c}}(\theta;\bar\mu_t), A_{\mathsf{w}}(\theta;\bar\mu_t)\big)$*, where*

$$\begin{aligned} A_{\mathsf{c}}(\theta;\bar\mu_t) &= \mathbb{E}\big[g(X,Y,\bar\mu_t)\sigma(X^\top w)\big], \\ A_{\mathsf{w}}(\theta;\bar\mu_t) &= \mathbb{E}\big[g(X,Y,\bar\mu_t) c\sigma'(X^\top w) X(\xi)\big] \qquad \text{if } \theta \in S_\xi \subseteq S. \end{aligned}$$

We now rewrite $A(\theta;\bar\mu_t)$ to express it in the shape encountered in [10]. More precisely, setting $\widetilde g(x,\theta) = c\sigma(x^\top w)$, $x \in \mathbb{R}^d$, $\theta = (c,w) \in S$, we have $A(\theta;\bar\mu_t) = \nabla V(\theta;\bar\mu_t)$, where $V(\theta;\bar\mu_t) = \mathbb{E}\big[g(X,Y,\bar\mu_t)\widetilde g(X,\theta)\big]$. Rewriting the evolution equation (4) in the gradient form already indicates that the long-time limit of $\bar\mu_t$ can be seen as a solution of a suitable minimization. We now elaborate on this interpretation in further detail. First, for a measure $\bar\mu$, we define the loss functional $\mathcal{E}[\bar\mu] = \frac{1}{2}\mathbb{E}\big[g(X,Y,\bar\mu)^2\big]$. Then, an explicit computation shows that if $\bar\mu_t$ solves the evolution equation (4), then

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}[\bar\mu_t] = \mathbb{E}\big[g(X,Y,\bar\mu_t)\frac{\mathrm{d}}{\mathrm{d}t}g(X,Y,\bar\mu_t)\big] \quad = -\mathbb{E}\big[g(X,Y,\bar\mu_t)\big\langle \nabla V(\cdot;\bar\mu_t)\nabla \widetilde g(X,\cdot), \bar\mu_t\big\rangle\big].$$

As this expression equals $= -\big\langle |\nabla V(\cdot;\bar\mu_t)|^2, \bar\mu_t\big\rangle$, the loss $\mathcal{E}$ decays along solution curves.

While this argument proves that the loss decreases monotonely during training, it does not yet imply the convergence to a minimum as $t \to \infty$. Concerning a standard, non-sparsified network, [10, Proposition 3.4] gives sufficient conditions for the convergence of the loss function $\mathcal{E}$ to a global minimum as $t \to \infty$. However, this result does not readily apply in our setting since it assumes that the weights w are taken from a compact manifold. This framework is incompatible with our update equations (2), where we rely on the vector space structure of $\mathbb{R}^d$. Hence, an extension of the global convergence property to our setting would require to refine the argumentation in [10, Proposition 3.4] in such a way that it no longer relies on the compactness of the weight space.

Besides [4, 10], also [11] relies on interacting particle systems to study SLPs with a large number of neurons. The long-time asymptotics of the weight evolution is described in [11, Theorem 6], which shows that the convergence to the long-time limit occurs at exponential speed. Since [11] does not need to assume that the parameter space is a compact manifold, it is substantially closer to our setting, and we expect that an extension of the arguments to sparse networks is feasible. A technical nuisance of [11, Theorem 6] is that it requires positivity of the smallest eigenvalue associated with the Hessian of V, which could be difficult to verify in specific examples. However, in contrast to [10], there is no characterization of the limit as the minimum of a suitable loss function. A possible perspective in this direction is offered by [12, Theorem 3.5], which shows that if there is convergence, then the limiting weight distribution minimizes the loss function. However, the challenge when applying this result in the present setting is that the established weak convergence is not enough since the methodology from [12] requires convergence with respect to the 2-Wasserstein distance. Moreover, as discussed in [12, Section D.3], to ensure regularity properties of specific limiting functions, further assumptions on the distribution of (X, Y) are needed, which may be difficult to verify in specific examples.

3. Simulation study

In this section, we perform simulation studies to illustrate the results of Theorem 1, see Section 3.1, and to investigate the influence of sparsity on the goodness-of-fit when using the ANN from (1) for a certain regression problem in Section 3.2.

3.1. Empirical distribution of parameters in the asymptotic regime

In order to illustrate the law of large numbers stated in Theorem 1, we approximate the function $f : [0,1]^2 \to \mathbb{R}$ defined by $f(s,t) = \sin(st)\sqrt{\log(1+t)}+\cos(t^2)$, $(s,t) \in [0,1^2]$ by the SLP $g(x,\boldsymbol{\theta})$ after Erdős-Rényi pruning with parameter $p = 1/2$ as defined in Section 2. The activation function is chosen to be $\sigma : \mathbb{R} \to \mathbb{R}$ with $\sigma(t) = (1-\exp(-t))/(2+2\exp(-t))$. Training is performed via SGD as given in (2) with learning rate $\alpha = 100$ and the number of iterations is chosen to be KN, where we put $K = 1,000$. As training data, we consider collections of random vectors $(X_1, f(X_1)), \dots, (X_{KN}, f(X_{KN}))$, where $X_1, \dots, X_{KN}$ are independent and uniformly distributed on the unit square, i.e., $X_i \sim U([0,1]^2)$ for each $i \in \{1, \dots, KN\}$. The initial parameter configuration is chosen at random, where each of the sequences $c_0^1, \dots, c_0^N$ and $w_0^1, \dots, w_0^N$ is iid with $c_0^1 \sim U(-10,10), w_0^1 \sim U([-10,10]^2)$.

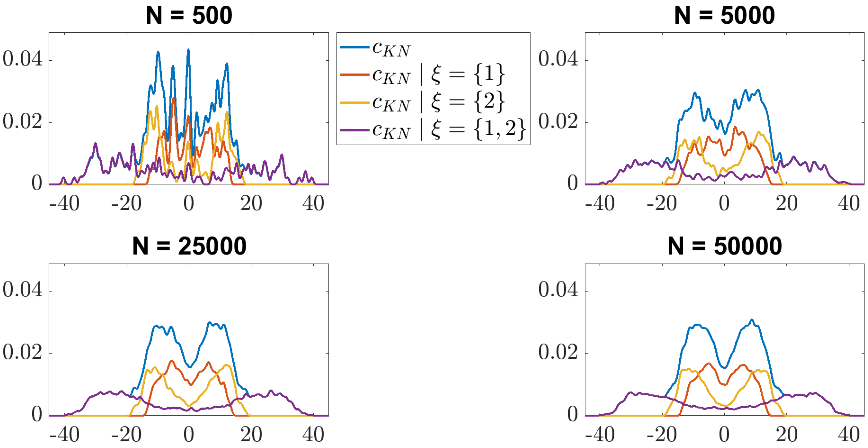

Figure 1: Probability density function of the real-valued parameter c_{KN} after KN iterations of the SGD for $N \in \{500, 5000, 25000, 50000\}$, obtained via kernel density estimation. The distribution is a mixture of the conditional distributions c_{KN} given that $\Xi = \xi$, where ξ is a subset of $\{1,2\}$, which are also shown in each of the four plots. Note that conditional probability density functions are scaled such that their sum gives the (unconditional) probability density functions of c_{KN}.

For $N \in \{500, 5000, 25000, 50000\}$, Figure 1 shows the empirical distribution of the sample $c_{KN}^1, \dots, c_{KN}^N$ in terms of probability density functions which are obtained by kernel density estimation with a Gaussian kernel and a fixed bandwidth of 0.5. Moreover, the empirical distributions conditioned on the realization of Ξ (defining the network topology as described in Section 2) are shown. This illustrates clearly that the distribution of c_{KN} is a mixture of the distributions conditioned on realizations of Ξ. Figure 1 shows the convergence of the distribution of c_{KN}. Only minor changes in the distribution can be observed between $N = 25000$ and $N = 50000$. Additionally, the empirical bivariate distributions of the samples $(c^1_{KN}, (w^1_{KN})_1), \ldots, (c^N_{KN}, (w^N_{KN})_1)$, $(c^1_{KN}, (w^1_{KN})_2), \ldots, (c^N_{KN}, (w^N_{KN})_2)$, and $((w^1_{KN})_1), (w^1_{KN})_2), \ldots, ((w^1_{KN})_1), (w^N_{KN})_2)$ are provided as supplementary material.

3.2. Influence of sparsity on the test error

Next, we present a simulation study to support the practical relevance of the new model. More precisely, we investigate the influence of the thinning parameter p in the Erdős-Rényi pruning on the goodness-of-fit, namely on the test error. For this purpose, we consider the Boston housing data[1], see [13], which serves as a benchmark data set for regression [14]. The data set consists of the median values for owner-occupied homes in 506 census tracts in Boston together with 13-dimensional covariate vectors. First, we randomly split the data into 405 training samples and 101 test samples. Then, we train the ANN with predefined sparsity from (1) by SGD, where we consider different numbers of hidden layers, i.e., $N \in \{100, 300, 500, 800, 2000, 5000\}$. For each choice of N, the model is trained for $p \in \{0.5, 0.6, \ldots, 1.0\}$. The learning rate $\alpha = 100$ and the number of iterations $(2.5 \cdot 10^{11})$ are kept fix.

Figure 2 shows the influence of N and p on the MSE of test data. The results shown here are averaged over 15 realizations of both, the SGD and the random sparsity. We observe that while the test error decreases with increasing N, thinning up to 50% of the connections does not deteriorate the prediction quality substantially. In particular, even a simple and non-adaptive concept of thinning for connection between neurons allows to reduce the number of parameters in ANNs without negatively affecting the test error.

[1] The Boston housing data is, e.g., contained in the package MASS of the statistical software R, see https://CRAN.R-project.org/package=MASS.

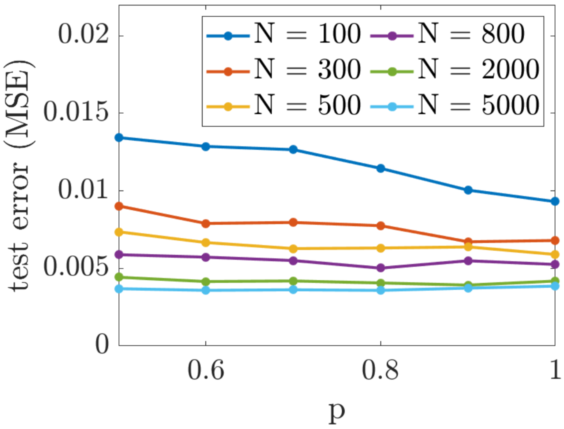

Figure 2: Mean squared error of the prediction by trained ANNs applied to the test data for different constellations of the number of hidden layers and the thinning parameter p.

4. Outline of proof

As in [4], we pursue the well-established three-step procedure for weak convergence towards a limiting process, which has been implemented in [7, 8]. We now state the three steps in detail and observe that they indeed imply the asserted Theorem 1. The proofs of the three results are deferred to Sections 5, 6 and 7.

Proposition 2 (Tightness). *Under conditions* **(E)** *and* **(M)** *the sequence* $\{\mathcal{L}(\mu^N)\}_{N\geqslant 1}$ *of distributions of the measures* μ^N *is tight.*

Proposition 3 (Uniqueness). *For a given initial value and given* $\{p_\xi\}_{\xi\in\mathcal{C}}$ *with* $\sum_{\xi\in\mathcal{C}} p_\xi = 1$*, Equation* (4) *has at most one solution* $\bar\mu_t$ *with* $\bar\mu_t(S_\xi) = p_\xi$*.*

Proposition 4 (Limit identification). *Under condition* **(E)***, any weak accumulation point of* $\{\mathcal{L}(\mu^N)\}_{N\geqslant 1}$ *satisfies Equation* (4)*.*

To make the presentation self-contained, we formally conclude the proof of Theorem 1.

Proof of Theorem 1. First, under condition **(E)**, $\mu_t^N(S_\xi) = \frac{1}{N}\#\{i \leqslant N : \xi^i = \xi\}$ converges to p_ξ, thereby yielding the decomposition (3). Next, by Proposition 2, any subsequence of $\{\mathcal{L}(\mu_\cdot^N)\}_{N\geqslant 1}$ has a weakly convergent subsequence. By Propositions 3 and 4, any such subsequence converges weakly to the unique solution of (4). Hence, also the entire sequence $\{\mathcal{L}(\mu_\cdot^N)\}_{N\geqslant 1}$ converges in distribution to that solution. □

5. Tightness

In this section, we show tightness of the sequence $\{\mathcal{L}(\mu^N)\}_{N\geqslant 1}$ in the Skorokhod space $D([0,T],\mathcal{M}(S))$. To that end, we rely on the established method, which involves compact containment and regularity, see Theorem 4.5 in [15]. In particular, the following assertions are true.

Proposition 5 (Compact containment)**.** *Let $\varepsilon > 0$.Then, for some compact $K \subseteq S$,*

$$\sup_{N\geqslant 1}\sup_{t\leqslant T}\mathbb{P}(\mu_t^N \notin K) \leqslant \varepsilon.$$

For regularity, we rely on Aldous' celebrated criterion, see Lemma 16.12 in [16].

Proposition 6 (Aldous' criterion)**.** *Let $f \in C_b^2(S_\xi)$. Then,*

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\tau}\mathbb{P}\big(\sup_{u\leqslant\delta}|\langle f,\mu^N_{\tau+u}\rangle - \langle f,\mu^N_\tau\rangle| \geqslant \varepsilon\big) = 0, \tag{5}$$

where, τ is taken from the family of all stopping times that are bounded by T.

Note that in order to verify (5) it is sufficient to show that

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\sup_{\sigma,\tau}\mathbb{E}\left[|\langle f,\mu^N_\sigma\rangle - \langle f,\mu^N_\tau\rangle| \wedge 1\right] = 0,$$

where τ and σ are taken from the family of stopping times fulfilling $\sigma \leq \tau \leq \sigma+\delta \leqslant T$. The proof of Proposition 5 is analogous to that of Lemma 2.2 in [4]. Hence, we focus on Proposition 6, where the arguments that we present differ from those used in [4]. The reason is that we found it difficult to extend the conditional expectation bounds from [4, Lemma 2.3] to our setting of sparsified networks. Hence, we give an alternative proof.

First, we bound the increments of the parameters during SGD. To that end, we rewrite (2) succinctly as

$$\theta^i_{k+1} - \theta^i_k = \frac{1}{N}B^N_k(\theta^i_k), \tag{6}$$

where $B^N_k(\theta) = \big(B_{k,\mathsf{c}}(\theta), B_{k,\mathsf{w}}(\theta)\big)$ with $B^N_{k,\mathsf{c}}(\theta) = g(X_k,Y_k,\nu^N_k)\sigma(w^\top X_k)$ and

$$B^N_{k,\mathsf{w}}(\theta) = g(X_k,Y_k,\nu^N_k)c\sigma'(w^\top X_k)X_k(\xi) \text{ if } \theta \in S_\xi.$$

To prove Proposition 6, we first discuss an auxiliary result. Instead of directly bounding the parameters as in Lemma 2.1 of [4], we found it more convenient to concentrate on the increments. As a preliminary step, we also rely on a related property for independent random variables, which we state and prove here to make the presentation self-contained.

Lemma 7 (Regularity for independent random variables). *Let $\{Z_k\}_{k\geqslant 1}$ be a family of iid non-negative random variables with finite second moment. Then, as δ tends to 0,*

$$\limsup_{N\to\infty}\frac{1}{N}\mathbb{E}\Big[\max_{k\leqslant N}\sum_{k\leqslant \ell\leqslant k+\delta N} Z_\ell\Big]\in O(\delta).$$

Proof. Note that if $k\leqslant N$ and $m\geqslant 1$ are such that $m\delta N\leqslant k\leqslant (m+1)\delta N$, then $k+\delta N\leqslant m\delta N+2\delta N$. Hence, we expand the expression under the expectation as

$$\max_{k\leqslant N}\sum_{\ell=k}^{k+\delta N} Z_\ell\leqslant \max_{m\leqslant 1/\delta}\sum_{\ell=m\delta N}^{m\delta N+2\delta N} Z_\ell = 2\delta N\mathbb{E}Z_1+\sqrt{N}\max_{m\leqslant 1/\delta}\sum_{\ell=m\delta N}^{m\delta N+2\delta N}\frac{Z_\ell-\mathbb{E}Z_\ell}{\sqrt{N}}.$$

Since $\sqrt{N}\in o(N)$, it suffices to show that the second moment of the above sum is bounded for each m. Now, leveraging independence, we get that

$$\mathsf{Var}\Big(\sum_{\ell=m\delta N}^{m\delta N+2\delta N}\frac{Z_\ell-\mathbb{E}Z_\ell}{\sqrt{N}}\Big)=2\delta\mathsf{Var}Z_1<\infty.$$ □

Lemma 8 (Boundedness of increments). *Assume condition* (**M**). *Then, as δ tends to* 0*,*

$$\limsup_{N\to\infty}\frac{1}{N}\mathbb{E}\Big[\max_{k\leqslant NT}\sum_{k\leqslant \ell\leqslant k+\delta N}\langle |B_\ell^N(\cdot)|^2,\nu_\ell^N\rangle\Big]\in O(\delta).$$

The proof of Lemma 8 is mainly based on arguments of Lemma 2.1 in [9] and is provided in the supplementary material. Finally, we prove Proposition 6.

Proof of Proposition 6. To ease notation, we omit henceforth the $\lfloor\cdot\rfloor$-symbols and write Ns instead of $\lfloor Ns\rfloor$. In particular, we write $\mu_t^N=\nu_{Nt}^N$. Then, by Taylor expansion, we find intermediate values $\{\bar\theta_k^i\}_{i\geqslant 1}\subseteq S$ such that

$$\begin{aligned}|\langle f,\nu_{N(\tau+u)}^N\rangle-\langle f,\nu_{N\tau}^N\rangle| &\leqslant \frac{1}{N}\sum_{N\tau\leqslant\ell\leqslant N(\tau+u)}\Big|\langle B_\ell^N(\cdot)\nabla f,\nu_\ell^N\rangle\Big|\\ &\quad+\frac{1}{2N^2}\sum_{i\leqslant N}\sum_{N\tau\leqslant\ell\leqslant N(\tau+u)}\Big|B_\ell^N(\theta_\ell^i)\nabla^2 f(\bar\theta_\ell^i)B_\ell^N(\theta_\ell^i)^\top\Big|.\end{aligned}$$

By assumption, all first- and second-order partial derivatives of f are uniformly bounded, which means that there exist $C_1,C_2>0$ such that

$$\begin{aligned}|\langle f,\nu_{N(\tau+u)}^N\rangle-\langle f,\nu_{N\tau}^N\rangle| &\leqslant \frac{C_1}{N}\sum_{N\tau\leqslant\ell\leqslant N(\tau+u)}\big(\langle |B_\ell^N(\cdot)|,\nu_\ell^N\rangle+\langle |B_\ell^N(\cdot)|^2,\nu_\ell^N\rangle\big)\\ &\leqslant \frac{C_2}{N}\sum_{N\tau\leqslant\ell\leqslant N(\tau+u)}\big(1+\langle |B_\ell^N(\cdot)|^2,\nu_\ell^N\rangle\big).\end{aligned}$$

Hence, applying Lemma 8 concludes the proof. □

6. Uniqueness

In this section, we show that Equation (4) admits a unique solution. For this we rely on a Picard-type argument for the ODE on S of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta_t = A(\theta_t; \mu_t), \tag{7}$$

with a generic $\mu_\cdot \in D([0,T], \mathcal{M}(S))$. Writing $D_T = D([0,T], S)$, this system gives rise to an operator $H : \mathcal{M}(D_T) \to \mathcal{M}(D_T)$ as follows. First, if $\mu \in \mathcal{M}(D_T)$ describes the distribution of a random path, then we let $\mu_0 \in \mathcal{M}(S)$ denote the distribution of the initial point. Now, we define $H(\mu)$ to be the distribution of the solution $\{\theta_t\}_{t \leqslant T}$ to (7) with initial value distributed according to μ_0.

The key observation is that H has a unique fixed point if restricted to a smaller space. To introduce this space rigorously, we first put $C_T = C([0,T], S)$ and $M_T = \mathcal{M}(C_T)$. Next, proceeding as in [4, p.742], for $\mu, \mu' \in M_T$ let the *coupling set* $P(\mu, \mu')$ denote the family of all probability measures on $C_T \times C_T$ coinciding with μ and μ' when projecting on the first and second marginal, respectively. Then,

$$d_{\mathsf{W},T}(\mu, \mu') = \inf_{\nu \in P(\mu,\mu')} \Big(\int 1 \wedge \sup_{s \leqslant T} |u_s - v_s|_4^4 \nu(\mathrm{d}(u_\cdot, v_\cdot)) \Big)^{1/4} \tag{8}$$

defines the 4-*Wasserstein* distance, where $|\cdot|_4$ is the ℓ^4-distance in S. We write $N_T \subseteq \mathcal{M}(C([0,T], S))$ for the subspace of all $\mu \in M_T$ such that $\int \sup_{s \leqslant T} |u_s|_4^4 \mu(\mathrm{d}u_\cdot) < \infty$. By [4, p.743], N_T becomes a Banach space with respect to $d_{\mathsf{W},T}$.

Lemma 9 (Regularity of solutions)**.** *Let $\mu \in \mathcal{M}(D_T)$ and let condition* **(M)** *be fulfilled. Then, $H(\mu) \in N_T$.*

Lemma 10 (Fixed point)**.** *If T is sufficiently small, then the restriction of H to the space N_T admits a unique fixed point.*

The proofs of Lemmas 10 and 9 are similar to those of Lemmas 4.1 and 4.3 in [9], respectively, and thus provided as supplementary material. Finally, we conclude the proof of Proposition 3.

Proof of Proposition 3. We may choose T to be small enough, so that Lemma 10 applies. First, as in Section 4 of [4], general results on Markov processes from [17] yield that solutions to (4) correspond uniquely to solutions of (7) by taking $\bar{\mu}_t$ to be the law of $\{\theta_t\}_{t\leqslant T}$. In particular, the law of $\{\theta_t\}_{t\leqslant T}$ is a fixed point of H and therefore contained in N_T by Lemma 9. Hence, the uniqueness result from Lemma 10 concludes the proof. □

7. Limit identification

Last not least, we prove Proposition 4. That is, any limit point of the processes $\{\mu_t^N\}_t$ satisfies Equation (4). Fix $f \in C_b^2(S)$. The central task is to quantify the error of $\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle$ in comparison to (4).

Lemma 11 (Deviation from evolution equation)**.** *Let $t \leqslant T$ and $f \in C_b^2(S)$. Then,*

$$\Delta_t(\mu_\cdot^N) = \langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle - \int_0^t \langle A(\cdot;\mu_s^N)\nabla f, \mu_s^N\rangle \mathrm{d}s$$

converges to 0 in probability as $N \to \infty$.

We elucidate how to derive Proposition 4 from Lemma 11, whose proof will be provided in the supplementary material.

Proof of Proposition 4. Let $\mu_\cdot$ be a process distributed according to a weak accumulation point $\mathbb{Q}$ of $\{\mu_\cdot^N\}_{N\geqslant 1}$. It suffices to prove that $\Delta_\cdot(\mu_\cdot) \equiv 0$ as a stochastic process, since then $\mathbb{Q}$ is concentrated on the unique solution of Equation (4). To that end, we verify that $\mathbb{E}_{\mathbb{Q}}[\Delta_t(\mu_\cdot)G(\mu_\cdot)] = 0$, for every $t > 0$ and bounded function $G : \mathcal{M}(D_T) \to [0,\infty)$ that is measurable with respect to $\{\mu_s\}_{s\leqslant t}$. Since measurability is considered via the product σ-algebra, it suffices to fix arbitrary $s_1 < \cdots < s_p \leqslant t$ and $g_1,\dots,g_p \in C_b(\mathbb{R}^{1+d})$, and then show that $\mathbb{E}_{\mathbb{Q}}\Delta'(\mu_\cdot) = 0$, where $\Delta'(\mu_\cdot) = \Delta_t(\mu_\cdot)\langle g_{s_1}, \mu_{s_1}\rangle \cdots \langle g_{s_p}, \mu_{s_p}\rangle$. Now, since $\Delta'(\mu_\cdot)$ is bounded and continuous in $\mu_\cdot$, and $\mathbb{Q}$ is a weak accumulation point of a subsequence $\{\mathcal{L}(\mu^{N_j})\}_{j\geqslant 1}$, we leverage Lemma 11 to deduce that $\mathbb{E}_{\mathbb{Q}}|\Delta'(\mu_\cdot)| \leqslant \limsup_{j\to\infty} \mathbb{E}|\Delta'(\mu^{N_j})| \leqslant \max_i(|g_i|_\infty) \lim_{j\to\infty} \mathbb{E}|\Delta_t(\mu^{N_j})| = 0$, as asserted. □

Acknowledgements. We thank both anonymous referees for their careful review. Their suggestions helped to substantially improve the quality of the manuscript. In particular, one of the referees motivated the detailed discussion of the convergence properties of SGD for increasing training time. CH acknowledges financial support of the CogniGron research center and the Ubbo Emmius Funds (University of Groningen). This work contributes to the research performed at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe). The work of MN was funded by the German Research Foundation (DFG) under Project ID 390874152 (POLiS Cluster of Excellence, EXC 2154).

References

[1] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge (MA), 2016.

[2] L. Pessoa, Understanding brain networks and brain organization, Phys. Life Rev. 11 (2014) 400–435.

[3] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, A. Liotta, Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science, Nat. Commun. 9 (2018) 2383.

[4] J. Sirignano, K. Spiliopoulos, Mean field analysis of neural networks: A law of large numbers, SIAM J. Appl. Math. 80 (2020) 725–752.

[5] S. Kaviani, I. Sohn, Influence of random topology in artificial neural networks: A survey, ICT Express 6 (2020) 145–150.

[6] S. Dey, K.-W. Huang, P. A. Beerel, K. M. Chugg, Pre-defined sparse neural networks with hardware acceleration, IEEE Trans. Emerg. Sel. Topics Circuits Syst. 9 (2019) 332–345.

[7] C. da Costa, B. F. P. da Costa, M. Jara, Reaction-diffusion models: From particle systems to SDE's, Stochastic Process. Appl. 129 (2019) 4411–4430.

[8] T. Bodineau, I. Gallagher, L. Saint-Raymond, S. Simonella, Fluctuation theory in the Boltzmann-Grad limit, J. Stat. Phys. 180 (2020) 873–895.

[9] J. Sirignano, K. Spiliopoulos, Mean field analysis of neural networks: A central limit theorem, Stochastic Process. Appl. 130 (2020) 1820–1852.

[10] G. M. Rotskoff, E. Vanden-Eijnden, Trainability and accuracy of neural networks: An interacting particle system approach, Commun. Pure Appl. Anal. (2022).

[11] S. Mei, A. Montanari, P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. USA 115 (33) (2018) E7665–E7671.

[12] L. Chizat, F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, Adv. Neural Inf. Process. Syst. 31 (2018).

[13] D. Harrison Jr, D. L. Rubinfeld, Hedonic housing prices and the demand for clean air, Journal of Environmental Economics and Management 5 (1978) 81–102.

[14] B. Clarke, E. Fokoué, H. H. Zhang, Principles and Theory for Data Mining and Machine Learning, Springer, New York, 2009.

[15] J. Jacod, A. N. Shiryaev, Limit Theorems for Stochastic Processes, 2nd Edition, Springer, Berlin, 2003.

[16] O. Kallenberg, Foundations of Modern Probability, 2nd Edition, Springer, New York, 2002.

[17] V. N. Kolokoltsov, Nonlinear Markov Processes and Kinetic Equations, Cambridge University Press, Cambridge, 2010.

[18] S. N. Ethier, T. G. Kurtz, Markov Processes, J. Wiley & Sons, New York, 1986.

Supplementary material

As supplementary material, we provide additional proofs as well as extended results of the simulation study performed in Section 3.1.

A. Proofs

Proof of Lemma 8. We deal with the c^i_k- and w^i_k- increments separately. First, according to (6) and the boundedness of σ, there are constants $C_1, C_1' > 0$ such that for every $k \leqslant \ell \leqslant k'$,

$$|c^i_{\ell+1}| \leqslant |c^i_\ell| + \frac{C_1}{N}|Y_\ell| + \frac{C_1}{N}|g(X_\ell, \nu_\ell)| \leqslant |c^i_\ell| + \frac{C_1}{N}|Y_\ell| + \frac{C_1'}{N^2}\sum_{j \leqslant N}|c^j_\ell|. \tag{9}$$

Hence, writing $\overline{Y}_N = \frac{1}{N}\sum_{\ell \leqslant N}|Y_\ell|$, we argue as in [4, p.734, line -1] to show that

$$|c^i_\ell| \leqslant C_2\Big(|c^i_0| + \frac{1}{N}\sum_{j \leqslant N}|c^j_0| + \overline{Y}_N\Big), \tag{10}$$

for some $C_2 > 0$. In particular, we can find a suitable $C_3 > 0$ such that $\langle B^N_{\ell,\mathsf{c}}(\cdot)^2, \nu^N_\ell\rangle \leqslant C_3\big(Y_\ell^2 + \overline{Y}^2_N + \overline{C}^2_N\big)$, where $\overline{C}^2_N = N^{-1}\sum_{j \leqslant N}(c^j_0)^2$. Thus, Lemma 7 yields the claim for the c^i_k-increments. Similarly, for the w^i_k-increments, the bound (10) yields suitable constants $C_4, C_4' > 0$ such that

$$\big|B^N_{\ell,\mathsf{w}}(\theta^i_\ell)\big| \leqslant C_4\big(|Y_\ell| + \frac{1}{N}\sum_{j \leqslant N}|c^j_\ell|\big)|X_\ell||c^i_\ell| \leqslant C_4'\big(|Y_\ell| + \frac{1}{N}\sum_{j \leqslant N}|c^j_0| + \overline{Y}_N\big)|X_\ell||c^i_\ell|.$$

In particular, applying (10) and using $abc \leqslant (a^3 + b^3 + c^3)/3$ for $a, b, c > 0$, we get that

$$\langle |B^N_{\ell,\mathsf{w}}(\cdot)|^2, \nu^N_\ell\rangle \leqslant C_5\big(Y_\ell^2 + \overline{C}^2_N + \overline{Y}^2_N\big)\frac{|X_\ell|^2}{N}\sum_{i \leqslant N}|c^i_\ell|^2 \leqslant C_5'\big(Y_\ell^6 + |X_\ell|^6 + \overline{C}^6_N + \overline{Y}^6_N\big)$$

for suitable $C_5, C_5' > 0$. Therefore,

$$\sum_{k \leqslant \ell \leqslant k'}\langle |B^N_{\ell,\mathsf{w}}(\cdot)|^2, \nu^N_\ell\rangle \leqslant C_5'\sum_{k \leqslant \ell \leqslant k'}(Y_\ell^6 + |X_\ell|^6) + (k' - k)\overline{C}^6_N + (k' - k)\overline{Y}^6_N,$$

so that an application of Lemma 7 concludes the proof. □

Proof of Lemma 9. First, analogously to Lemma 4.1 in [4], there exists a constant $C > 0$ such that $\mathbb{E}[(c_t - c_s)^4] \leqslant C(t-s)^4$ and $\mathbb{E}[|w_t - w_s|^4] \leqslant C(\mathbb{E}[|c_0|^4] + 1)(t-s)^4$. These bounds imply that the processes $\{c_t\}_{t \geq 0}$ and $\{w_t\}_{t \geq 0}$ have continuous versions according to the Kolmogorov-Chentsov criterion, see Theorem 3.23 in [16]. Moreover, they also imply that the solution curves have bounded fourth moments, so that indeed $H(\mu_\cdot) \in N_T$. □

Proof of 10. Having set up the distance notion in (8), we now show that H is a contraction with respect to $d_{\mathsf{W},T}$. First, the evolution equation for c_t does not change at all through our pruning, so that we can import the estimates from Lemma 4.3 in [4] to conclude that

$$|c_t^{(1)} - c_t^{(2)}| \leqslant C \int_0^t (|w_s^{(1)} - w_s^{(2)}| + d_{\mathsf{W},s}(\mu^{(1)}, \mu^{(2)}))\mathrm{d}s$$

for a suitable $C > 0$. Next, we decompose $w_t^{(1)} - w_t^{(2)}$ as

$$\begin{aligned} w_t^{(1)} - w_t^{(2)} &= \int_0^t \mathbb{E}\Big[X(\xi)(g(X,Y,\mu_s^{(1)}) - g(X,Y,\mu_s^{(2)}))c_s^{(1)}\sigma'(w_s^{(1)} \cdot X)\Big]\mathrm{d}s \\ &\quad + \int_0^t \mathbb{E}\Big[X(\xi)g(X,Y,\mu_s^{(2)})(c_s^{(1)}\sigma'(w_s^{(1)} \cdot X) - c_s^{(2)}\sigma'(w_s^{(2)} \cdot X))\Big]\mathrm{d}s. \end{aligned}$$

The only difference to the corresponding expression in Lemma 4.3 of [4] is that we now see $X(\xi)$ instead of X. However, in the ensuing estimates X only appears through its length $|X|$. Since $|X(\xi)| \leqslant |X|$, the arguments extend to the novel setting. Note that Lemma 4.3 in [4] requires that $\mathbb{E}\exp(q|c_0^i|) < \infty$ for some $q > 0$. More precisely, this expression appears after an application of Grönwall's Lemma, see, e.g., Appendix 5 in [18]. □

It remains to prove Lemma 11.

Proof of Lemma 11. By relying on a Taylor expansion as in the proof of Proposition 6, we see that

$$\Big|\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle - \frac{1}{N}\sum_{k \leqslant Nt}\langle B_k(\cdot;\nu_k^N)\nabla f, \nu_k^N\rangle\Big| \leqslant \frac{C_1}{N^2}\sum_{k \leqslant Nt}\langle |B_\ell^N(\cdot)|^2, \nu_\ell^N\rangle$$

for some constant $C_1 > 0$. Now, as in Proposition 6, we deduce that the expression $N^{-2}\,\mathbb{E}[\sum_{k\leqslant Nt}\langle |B_\ell^N(\cdot)|^2, \nu_\ell^N\rangle]$ tends to 0 as $N \to \infty$. Thus, it suffices to show that

$$\begin{aligned} M(t) &= \frac{1}{N}\sum_{k \leqslant Nt}\langle B_k(\cdot;\nu_k^N)\nabla f, \nu_k^N\rangle - \int_0^t \langle A(\cdot;\mu_s^N)\nabla f, \mu_s^N\rangle \mathrm{d}s \\ &= \frac{1}{N}\sum_{k \leqslant Nt}\langle (B_k(\cdot;\nu_k^N) - A(\cdot;\nu_k^N))\nabla f, \nu_k^N\rangle - \int_{\lfloor Nt\rfloor/N} \langle A(\cdot;\nu_{\lfloor Ns\rfloor}^N)\nabla f, \nu_{\lfloor Ns\rfloor}^N\rangle \mathrm{d}s \end{aligned}$$

tends to 0 in probability. We even show that it tends to 0 in L^1. Since

$$\mathbb{E}|\langle A(\cdot;\nu_{\lfloor Ns\rfloor}^N))\nabla f, \nu_{\lfloor Ns\rfloor}^N\rangle| \leqslant C_2\mathbb{E}\langle \mathbb{E}[|Y| + |g(X,\cdot)|], \nu_{\lfloor Ns\rfloor}^N\rangle \leqslant C_3 + \frac{C_3}{N}\sum_{i \leqslant N}\mathbb{E}|c^i_{\lfloor Ns\rfloor}|$$

for some constants $C_2, C_3 > 0$, we obtain by (10) that the integral term of $M(t)$ tends to 0 in L^1. Moreover, we show that the sum appearing in the expression of $M(t)$ tends to 0 in L^2. By setting $M_k = N^{-1}\left\langle (B_k(\cdot;\nu_k^N) - A(\cdot;\nu_k^N))\nabla f, \nu_k^N \right\rangle$, we observe that since the training data $\{(X_k, Y_k)\}_{k\geqslant 1}$ is iid, the sequence $\{M_k\}_{k\geq 1}$ defines a martingale difference sequence. Therefore, the cross-terms in the expansion of the square disappear, i.e., $\sum_{k<k'\leqslant Nt} \mathbb{E}[M_k M_{k'}] = 0$ and thus $\mathbb{E}\big(\sum_{k\leqslant Nt} M_k\big)^2 = \sum_{k\leqslant Nt} \mathbb{E}M_k^2$. Noting that each summand $\mathbb{E}M_k^2$ is of order $1/N$ concludes the proof. □

B. Additional results obtained by simulation studies

We provide further results related to the simulation study presented in Section 3. We show the empirical bivariate distributions of the samples $(c^1_{KN}, (w^1_{KN})_1), \ldots, (c^N_{KN}, (w^N_{KN})_1)$, $(c^1_{KN}, (w^1_{KN})_2), \ldots, (c^N_{KN}, (w^N_{KN})_2)$, and $((w^1_{KN})_1), (w^1_{KN})_2), \ldots, ((w^1_{KN})_1), (w^N_{KN})_2)$ for $N = 500, N = 5000, N = 25000, N = 50000$ in Figures 3, 4, 5, respectively. The distributions are shown in terms of probability density functions estimated by kernel density estimation. For this purpose, a bivariate Gaussian kernel with a bandwidth of 0.2 is used. The values of the estimated probability density functions are represented by a heat map on the log-scale. Domains which do not belong to the support of the estimated probability density function are represented in white. Figures 3, 4, 5 nicely show how the considered empirical bivariate distributions approach the limit distribution with increasing values of N.

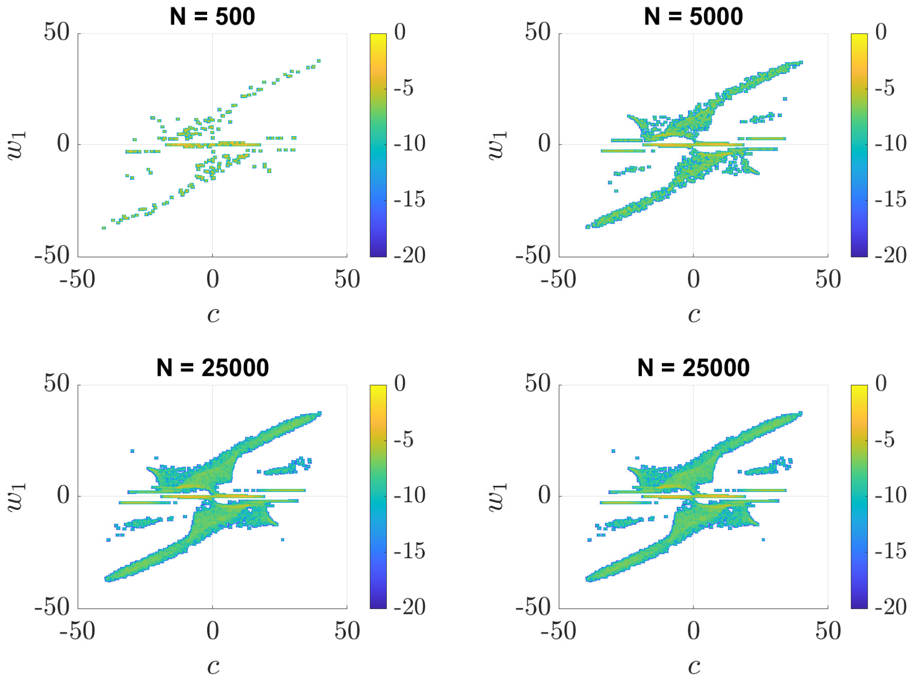

Figure 3: Probability density function of the parameter vector $(c_{KN}, (w_{KN})_1)$ after KN iterations of the SGD for $N \in \{500, 5000, 25000, 50000\}$, obtained via kernel density estimation. The values of the density are represented on the log-scale, where domains which do not belong to the support of the estimated distribution are represented in white.

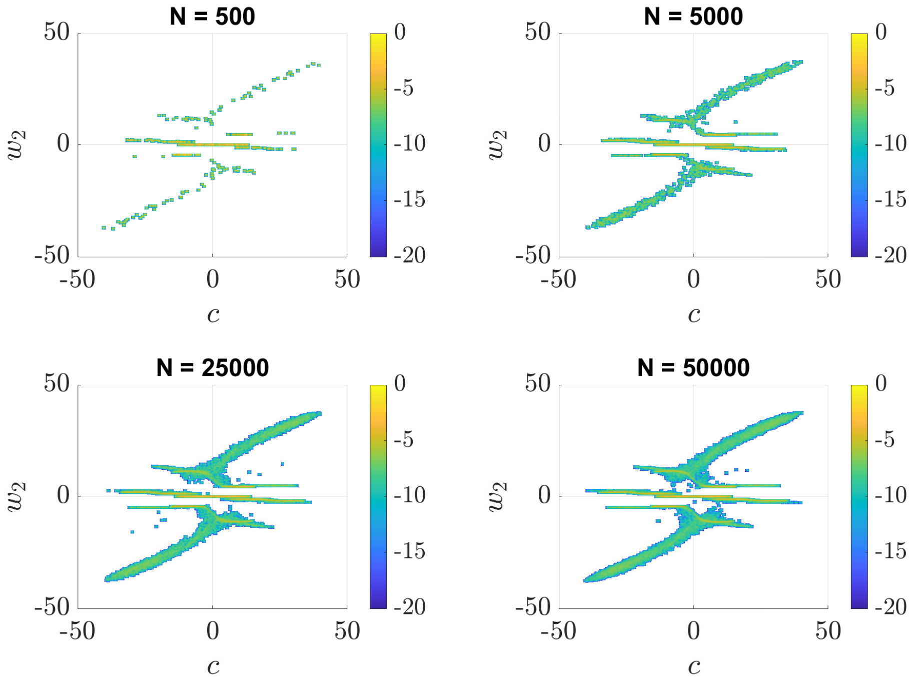

Figure 4: Probability density function of the parameter vector $(c_{KN}, (w_{KN})_2)$ after KN iterations of the SGD for $N \in \{500, 5000, 25000, 50000\}$, obtained via kernel density estimation. The values of the density are represented on the log-scale, where domains which do not belong to the support of the estimated distribution are represented in white.

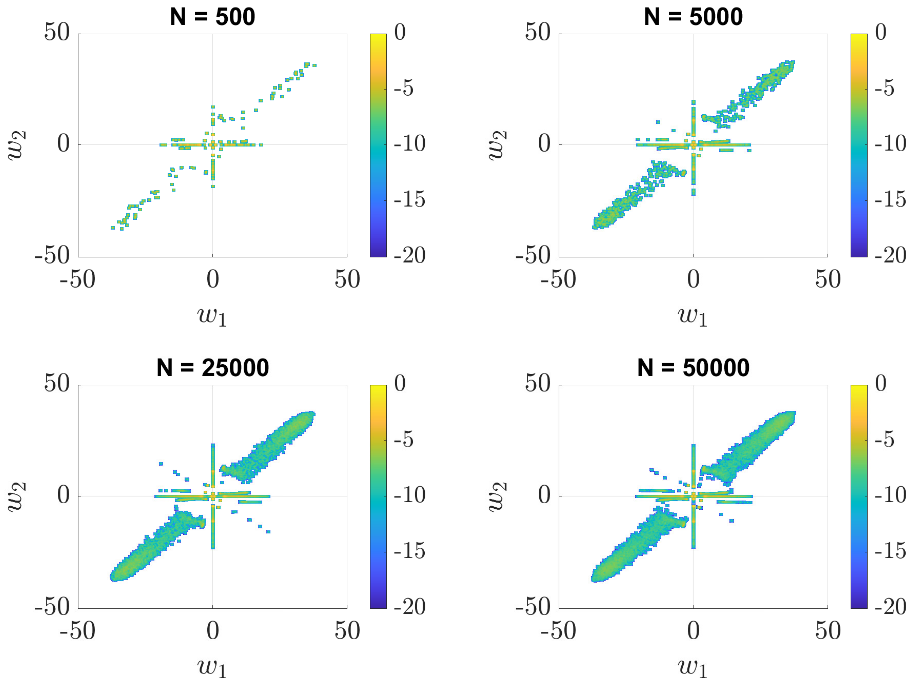

Figure 5: Probability density function of the parameter vector $((w_{KN})_1, (w_{KN})_2)$ after KN iterations of the SGD for $N \in \{500, 5000, 25000, 50000\}$, obtained via kernel density estimation. The values of the density are represented on the log-scale, where domains which do not belong to the support of the estimated distribution are represented in white.