Theories of Deep Learning Project: A Law of Large Numbers for Shallow Neural Networks

Thomas Tendron

January 2021

Abstract

The purpose of this note is to present the results in the recent paper *Mean Field Analysis Of Neural Networks: A Law of Large Numbers* of J. Sirignano and K. Spiliopoulos ([1]). The authors prove the convergence in probability of the empirical measure of the weights of a Convolutional Neural Network (CNN) with a single hidden layer, as the width of the hidden layer and the number of Stochastic Gradient Descent (SGD) iterations go to infinity, to the unique solution of a measure evolution equation. The shallow CNN architectures considered are mainly suitable for regression tasks, as the ouput is a single real number. To illustrate the theory, we experiment with a regression problem.

1. Introduction

We consider a fully connected one hidden layer CNN with output $g_\theta^N(x)$ defined by

$$g_\theta^N(x) = \frac{1}{N}\sum_{i=1}^N c^i \sigma(w^i \cdot x), \quad c^i \in \mathbb{R}, \quad w^i \in \mathbb{R}^d, \tag{1.1}$$

where $x \in \mathbb{R}^d$ is the input data, $\theta = (c^1, \cdots, c^N, w^1, \cdots, w^N) \in \mathbb{R}^{(1+d)N}$ denotes the set of weights to be learned, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function. Since $g_\theta^N(x) \in \mathbb{R}$, we are in the setting of a regression task. We therefore consider the Mean-Squared Error (MSE) loss function

$$L^N(\theta) = \frac{1}{2}\mathbb{E}_{(X,Y)\sim\pi}[(Y - g_\theta^N(X))^2], \tag{1.2}$$

for some $\pi \in \text{Prob}(\mathcal{X} \times \mathcal{Y})$, where $\mathcal{X} \times \mathcal{Y}$ denotes the state space of the data-label pair (X, Y).

SGD will be used to learn the parameters θ, that is, if $(x_k, y_k) \sim \pi$ and $L_k^N(\theta) \coloneqq (y_k - g_\theta^N(x_k))^2$, then we set

$$c_{k+1}^i = c_k^i - \alpha\partial_{c_k^i} L_k^N(\theta_k) = c_k^i + \frac{\alpha}{N}(y_k - g_{\theta_k}^N(x_k))\sigma(w_k^i \cdot x_k), \tag{1.3}$$

$$w_{k+1}^{i,j} = w_k^{i,j} - \alpha\partial_{w_k^{i,j}} L_k^N(\theta_k) = w_k^{i,j} + \frac{\alpha}{N}(y_k - g_{\theta_k}^N(x_k))c_k^i\sigma'(w_k^i \cdot x_k)x_k^j, \quad j \in \{1, \cdots, d\}, \tag{1.4}$$

where α is the learning rate. We will study the assymptotic behavior of the CNN via the empirical measure of its weights

$$\nu_k^N(dc, dw) = \frac{1}{N}\sum_{i=1}^N \delta_{(c_k^i, w_k^i)}(dc, dw). \tag{1.5}$$

To take into account the SGD iterations, we consider the rescaled version $\mu_t^N \coloneqq \nu_{\lfloor Nt \rfloor}^N$, for $t \in [0, T]$. It is easy to see that the floor function in the rescaling ensures that the paths $t \mapsto \mu_t^N$ are càdlàg in the topology of weak convergence, so that $\mu_\cdot^N$ is a random element of the Skorokhod space $D_E([0, T])$, where $T > 0$ and $E \coloneqq \text{Prob}(\mathbb{R}^{1+d})$. Throughout the note, we work on an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ rich enough to define all our random variables. We next introduce the assumptions the authors work under, and briefly discuss their implications.

Assumptions 1.1. *(i) The nonlinear activation function σ belongs to $C_b^2(\mathbb{R})$.*

(ii) The data samples (x_k, y_k) are iid with law π, and $\mathbb{E}[\|x_k\|^4] + \mathbb{E}[|y_k|^4] < \infty$.

(iii) The initial parameters (c_0^i, w_0^i) are iid with law $\bar{\mu}_0$, the c_0^i have exponential tails, and $\mathbb{E}[\|w_0^i\|^4] < \infty$.

We note that Assumption 1.1.(i) includes the widely used sigmoid function. Moreover, we have $\mu_0^N \overset{d}{\to} \bar{\mu}_0$ as $N \to \infty$. Indeed, if $f \in C_b(\mathbb{R}^{1+d})$, then Assumption 1.1.$(iii)$ and the strong law of large numbers imply that

$$\int_{\mathbb{R}^{1+d}} f(c,w) d\mu_0^N(c,w) = \frac{1}{N}\sum_{i=1}^N f(c_0^i, w_0^i) \overset{N\to\infty}{\to} \int_{\mathbb{R}^{1+d}} f(c,w) d\bar{\mu}_0(c,w).$$

Finally, we observe that the SGD iterations (1.3) and (1.4) show that (c_k^i, w_k^i) is a functional of (c_0^i, w_0^i) and $(x_j, y_j)_{j=1}^k$. Thus, the iid-ness Assumptions 1.1.(ii) and 1.1.(iii) imply the equality in distribution $(c_k^{\rho(i)}, w_k^{\rho(i)})_{i=1}^N \overset{d}{=} (c_k^i, w_k^i)_{i=1}^N$ for any permutation ρ of the indices $\{1, \cdots, N\}$, that is $(c_k^i, w_k^i)_{i=1}^N$ is exchangeable.

2. Main Results and some Remarks

The main result is the convergence in probability of $\mu_\cdot^N$ to a nonlinear measure evolution equation, which, as we will see below, generalizes some continuous optimization problems known in optimal transportation theory as gradient flows.

Given a measurable function f and a probability measure μ, we define $\langle f, \mu\rangle = \int f d\mu$.

Theorem 2.1. *The scaled empirical measure $\mu_\cdot^N$ converges in probability to a deterministic measure $\bar{\mu}_\cdot$. The path $t \mapsto \bar{\mu}_t$ is càdlàg, and $\bar{\mu}_\cdot$ is the unique solution of the measure evolution equation*

$$\langle f, \bar{\mu}_t\rangle = \langle f, \bar{\mu}_0\rangle + \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c'\sigma(w'\cdot x), \bar{\mu}_s\rangle)(\langle \nabla(c\sigma(w\cdot x))\cdot\nabla f), \bar{\mu}_s\rangle \pi(dx,dy)\right) ds, \quad (2.1)$$

where $\nabla f = \nabla_\theta f = (\partial_c f, \nabla_w f)$.

If the limit $\bar{\mu}_0$ of the law of the initial data μ_0^N as a density, the equation simplifies to a known problem.

Corollary 2.2. *If $\bar{\mu}_0$ has a density p_0 and there exists a unique solution to the nonlinear partial differential equation (PDE)*

$$\begin{aligned} \partial_t p(t,\theta) &= -\alpha \operatorname{div}_\theta(p(t,\theta)\nabla_\theta v(\theta, p(t,\cdot))), \quad p(0,\theta) = p_0(\theta), \quad t > 0, \quad \theta \in \mathbb{R}^{1+d}, \qquad (2.2)\\ v(\theta, p(t,\cdot)) &= \int_{\mathcal{X}\times\mathcal{Y}} \left(\left(y - \int_{\mathbb{R}^{1+d}} c'\sigma(w'\cdot x) p(t,c',w') dc'dw' \right) c\sigma(w\cdot x)\right) \pi(dx,dy) \end{aligned}$$

such that $p(t,c,w)$ as $|c|, \|w\| \to \infty$, then we have that the solution $\bar{\mu}_\cdot$ to (2.1) is absolutely continuous with respect to dcdw with

$$\bar{\mu}_t(dc,dx) = p(t,c,w)dcdw.$$

By Corollary 2.2, we have $\langle c\sigma(w\cdot x), \mu_t^N\rangle \overset{N\to\infty}{\to} \langle c\sigma(w\cdot x), p(t,c,w)dcdw\rangle$. Therefore, by the dominated convergence theorem, the loss function $L^N(\theta)$ of our shallow CNN, which was defined in (1.2), satisfies

$$\begin{aligned} L^N(\theta_{\lfloor Nt\rfloor}) &= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} (y - \langle c\sigma(w\cdot x), \mu_t^N\rangle)^2 \pi(dx,dy)\\ &\overset{N\to\infty}{\to} \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} (y - \langle c\sigma(w\cdot x), p(t,c,w)dcdw\rangle)^2 \pi(dx,dy) =: \bar{L}(p(t,\cdot)) \end{aligned}$$

The PDE (2.2) is known as a continuity equation, and its solution as a gradient flow for the objective functional $\bar{L}$. As seen in Theorem 11.1.3 in the book [2], gradient flows such as the solution of (2.2) coincide with curves of maximal slope. This means that as t grows, we expect the solution of (2.2) to converge towards a critical point of the functional $\bar{L}$. Hence, Theorem 2.1 and Corollary 2.2 show that in the limit of large width of the hidden layer and large number of SGD iterations, the distribution of the weights of our shallow CNN is described by an infinite-dimensional optimization problem, with the continuous analogue of the MSE as the objective functional.

Finally, Theorem 2.1 and the exchangeability of $(c_k^i, w_k^i)_{i=1}^N$ that we obtained from Assumptions 1.1.(ii) and 1.1.(iii) allow us to apply the Sznitman-Tanaka Theorem to obtain propagation of Chaos.

Theorem 2.3. *For any $t \geq 0$, let ρ_t^N denote the joint law of $(c^i_{\lfloor Nt \rfloor}, w^i_{\lfloor Nt \rfloor})_{i=1}^N$. Then, the sequence of probability measures $(\rho_t^N)_{N\geq 1}$ is $\bar{\mu}_t$-chaotic, that is for each $k \in \mathbb{N}$, we have asymptotic independence of the first k components:*

$$\lim_{N\to\infty} \Big\langle \prod_{i=1}^k f_i(x^i), \rho_t^N(dx^1, \cdots, dx^N) \Big\rangle = \prod_{i=1}^k \langle f_i, \bar{\mu}_t \rangle, \quad \forall f \in C_b^2(\mathbb{R}^{1+d}).$$

3. Sketches of the Proofs

The techniques used are reminiscent of the proofs often found in the study of interacting particle systems arising from statistical physics and population genetics models. The bulk of the paper is Theorem 2.1 and its proof. It will easily imply Corollary 2.2 and Theorem 2.3.

3.1. Theorem 2.1

It suffices to prove that $(\mu_\cdot^N)_{N\geq 1}$ is relatively compact in $D_E([0,T])$, and that the limit point of every convergent subsequence of $(\mu_\cdot^N)_{N\geq 1}$ is the unique solution to the measure evolution equation (2.1). The most important part of the proof of relative compactness is Lemma 2.1. in the paper of interest [1], which shows, using the structure of the SGD algorithm (i.e. equations (1.3) and (1.4)) along with the fourth moment and exponential tails in Assumptions 1.1.(i) and 1.1.(i), that for all $k \leq TN$ and uniformly in $i \in \mathbb{N}$, there exists $C \in (0,\infty)$ such that

$$\mathbb{E}[|c_k^i| + \|w_k^i\|] < C. \tag{3.1}$$

Using Section 13 of Billingsley's book on weak convergence [3], this provides enough control to show that the law of $(\mu_\cdot^N)_{N\geq 1}$ is tight in the sense that for any $\epsilon > 0$, there exists a compact subset $\mathcal{K}$ of $D_E([0,T])$ such that

$$\sup_{N\geq 1} \mathbb{P}(\mu_\cdot^N \notin \mathcal{K}) < \epsilon.$$

By Prokhorov's Theorem, this implies that $(\mu_\cdot^N)_{N\geq 1}$ is relatively compact in $D_E([0,T])$.

Since the measure evolution equation (2.1) can be seen to have at most a single solution using the usual Picard iteration and Banach Fixed Point Theorem technique, it only remains to show that any convergent subsequence satifies (2.1).

Fix any test function $f \in C_b(\mathbb{R}^{1+d})$. By Taylor's Theorem, (1.3)-(1.4), and (3.1), we can write

$$\begin{aligned}\langle f, \nu_{k+1}^N \rangle - \langle f, \nu_k^N \rangle = &\frac{1}{N^2} \sum_{i=1}^N \partial_c f(c_k^i, w_k^i) \alpha(y_k - g_{\theta_k}^N(x_k)) \sigma(w_k^i \cdot x_k) \\ &+ \frac{1}{N^2} \sum_{i=1}^N \alpha(y_k - g_{\theta_k}^N(x_k)) c_k^i \sigma'(w_k^i \cdot x_k) \nabla_w f(c_k^i, w_k^i) \cdot x_k + O_{\mathbb{P}}(N^{-2}).\end{aligned}$$

In order to split the right-hand side into drift and martingale parts, we introduce

$$\begin{aligned}D_k^{1,N} &:= \frac{1}{N} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \nu_k^N \rangle) \langle \sigma(w\cdot x) \partial_c f, \nu_k^N \rangle \pi(dx, dy) \\ D_k^{2,N} &:= \frac{1}{N} \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \nu_k^N \rangle) \langle c\sigma'(w\cdot x) x \cdot \nabla_w f, \nu_k^N \rangle \pi(dx, dy),\end{aligned}$$

and

$$M_k^{1,N} \coloneqq \frac{1}{N}\alpha(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle)\langle \sigma(w\cdot x_k)\partial_c f, \nu_k^N\rangle - D_k^{1,N}$$
$$M_k^{2,N} \coloneqq \frac{1}{N}\alpha(y_k - \langle c\sigma(w\cdot x_k), \nu_k^N\rangle)\langle c\sigma'(w\cdot x_k)x\cdot\nabla_w f, \nu_k^N\rangle - D_k^{2,N}.$$

It follows that

$$\langle f, \nu_{k+1}^N\rangle - \langle f, \nu_k^N\rangle = D_k^{1,N} + D_k^{2,N} + M_k^{1,N} + M_k^{2,N} + O_{\mathbb{P}}(N^{-2}). \tag{3.2}$$

Consequently, after summing (3.2) from $k=0$ to $\lfloor Nt\rfloor - 1$ and taking N large, we obtain by a Riemann sum approximation

$$\begin{aligned}\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle &= \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle)\langle \sigma(w\cdot x)\partial_c f, \mu_s^N\rangle \pi(dx,dy)\right) ds \\ &+ \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle)\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N\rangle \pi(dx,dy)\right) ds \\ &+ \sum_{k=0}^{\lfloor Nt\rfloor}(M_k^{1,N} + M_k^{2,N}) + O_{\mathbb{P}}(N^{-1}).\end{aligned} \tag{3.3}$$

Moreover, it is not hard to show that the last sum converges to 0 in L^2. (see Lemma 3.1. in [1]). Finally, let $(\eta^N)_N$ denote any subsequence of the sequence of laws of $(\mu_\cdot^N)$. By the relative compactness obtained above, there exists a subsequence $(\eta^{N_k})_k$ converging weakly to an element of $\mathrm{Prob}(D_E([0,T]))$. For each $0 \le s_1 < \cdots < s_p \le t \le T$, $f \in C_b^2(\mathbb{R}^{1+d})$ and $g_1, \cdots, g_p \in C_b(\mathbb{R}^{1+d})$, we define a functional $F : D_E([0,T]) \to [0,\infty)$ by

$$\begin{aligned}F(\mu) = \Bigg|\Bigg(\langle f, \mu_t^N\rangle - \langle f, \mu_0^N\rangle - \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle)\langle \sigma(w\cdot x)\partial_c f, \mu_s^N\rangle \pi(dx,dy)\right) ds \\ - \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle c\sigma(w\cdot x), \mu_s^N\rangle)\langle c\sigma'(w\cdot x)x\cdot\nabla_w f, \mu_s^N\rangle \pi(dx,dy)\right) ds\Bigg) \prod_{i=1}^p \langle g_i, \mu_{s_i}^N\rangle\Bigg|\end{aligned}$$

By (3.3), we have

$$\mathbb{E}_{\eta^N}[F(\mu)] = \mathbb{E}[F(\mu^N)] \le C\left(\frac{1}{\sqrt{N}} + \frac{1}{N}\right) \overset{N\to\infty}{\to} 0,$$

so that $E_\eta[F(\mu)] = 0$. By arbitrariness of the test functions and times, and uniqueness of the solution of (2.1), we infer $\eta = \delta_{\bar\mu}$.

3.2. Corollary 2.2 and Theorem 2.3

Corollary 2.2 follows from Theorem 2.1 by integration by parts. As mentioned in Section 2, Theorem 2.3 is an immediate consequence of the Sznitman-Tanaka Theorem on propagation of chaos, which we state below.

Theorem 3.1. (Theorem 3.2 in [4]) *Let (S,d) be a separable metric space. Let ρ be a law on S, and for each $n \in \mathbb{N}$, let ρ_n be law on S^n that is invariant under the action of permutation group on n elements. For $s_i \in S$, $i = 1, \cdots, n$, we define the map*

$$\epsilon_n((s_1, \cdots, s_n)) \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{s_i}.$$

Then $(\rho_n)_n$ is ρ-chaotic if and only if $\rho_n \circ \epsilon_n^{-1} \overset{n\to\infty}{\to} \delta_\rho$ in $\mathrm{Prob}(\mathrm{Prob}(S))$.

In our application, $S = \mathbb{R}^{1+d}$, $\rho = \bar\mu$ by Theorem 2.1 and $\rho_n = \mathbb{P}\circ((c^i_{\lfloor nt\rfloor}, w^i_{\lfloor nt\rfloor})_{i=1}^n)^{-1}$, which is invariant under the permutations on n elements by exchangeability of $(c^i_{\lfloor nt\rfloor}, w^i_{\lfloor nt\rfloor})_{i=1}^n$.

4. Numerical Experiment: Application of the Theory to a Regression Problem

As an illustration of the theory, we consider a regression task associated to a data set first used in the paper [5]. The problem is to infer a real number representing the miles per gallon (mpg) of a car given some of its other characteristics. The data set has 398 entries. For each car in the data set, we have access to eight attributes, five of which are real numbers, and three of which are categorical but changed to real numbers for the training. We use TensorFlow ([6]) to train the model on a fully connected CNN with a single hidden layer with sigmoid activation function, and a single unit in the output layer. The objective function is the MSE. We have used no biases in order to match the theoretical framework (1.1). We have trained the model four times with $N = 10,000, 50,000, 100,000, 250,000$, and a corresponding number of epochs equal to $N/10$. We note that the the out put of our trained network is of the form

$$Ng_\theta^N(x) = \sum_{i=1}^{N} c^i \sigma(w^i \cdot x), \tag{4.1}$$

and that the learning rate used in the SGD algorithm is $\frac{1}{N}$. Using Tensoboard, we built histograms of the weights of the hidden and output layers every ten epoch with the aim of seeing a similar shape arise from the different sessions, especially for the largest values of N.

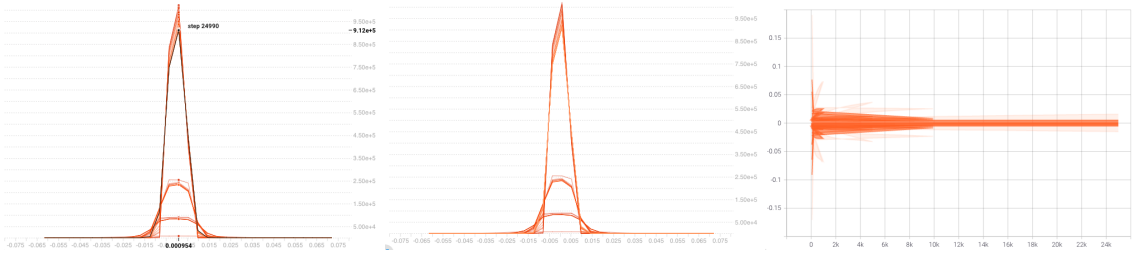

Figure 1: These three plots concern the weights of the hidden layer. In the two leftmost histograms, we observe four clusters of graphs corresponding to the four training sessions with different N. The clusters appear with increasing height as N increases since the histograms are not rescaled (see (4.1)). In the leftmost graph, one of the latest epochs is highlighted. We observe that after rescaling, the plots for different values of N have similar shapes, as predicted by Theorem 2.1. The rightmost plot shows curves representing the evolution of the percentiles [1, 0.93, 0.84, 0.69, 0.50, 0.31, 0.16, 0.07, 0] as the number of epochs grows. The curves are almost parallel for any N, and again up to rescaling the different training sessions are visibly close. This is also a manifestation of Theorem 2.1 and its corollaries.

We also include the histograms and percentile plot for the weights of the output layer; the interpretation is the same as in figure 1.

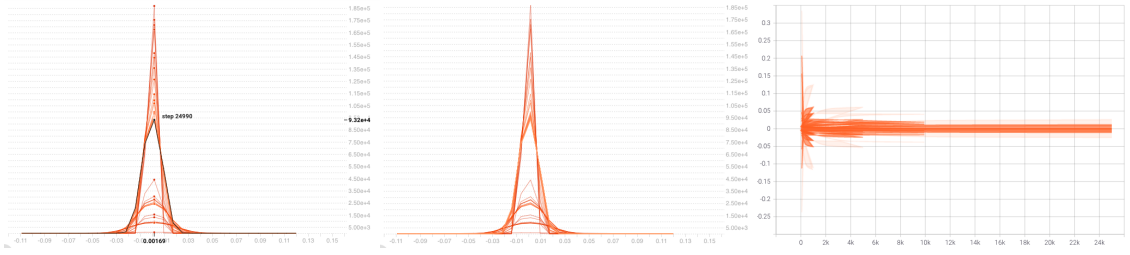

Figure 2: Histograms and percentile plot for the weights of the output layer

Finally, we include the python code.

```
import matplotlib.pyplot as plt
```

```
import numpy as np
import pandas as pd
import seaborn as sns
import datetime

np.set_printoptions(precision=3, suppress=True)

import tensorflow as tf

from tensorflow import keras
from tensorflow.keras import layers
from tensorflow.keras.layers.experimental import preprocessing
import time
from tensorflow.keras.callbacks import TensorBoard

widthHiddenLayer = 250000; # parameter for the width of the hidden layer, and the n

NAME = "RegrPb-{}".format(int(time.time()))
tensorboard_callback = TensorBoard(log_dir='logs/RegrPb', histogram_freq=10, write_i

url = 'http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.d
column_names = ['MPG', 'Cylinders', 'Displacement', 'Horsepower', 'Weight',
                'Acceleration', 'Model Year', 'Origin']

raw_dataset = pd.read_csv(url, names=column_names,
                          na_values='?', comment='\t',
                          sep=' ', skipinitialspace=True)

dataset = raw_dataset.copy()
dataset.tail()

dataset.isna().sum()

dataset = dataset.dropna()

dataset['Origin'] = dataset['Origin'].map({1: 'USA', 2: 'Europe', 3: 'Japan'})

dataset = pd.get_dummies(dataset, prefix='', prefix_sep='')
dataset.tail()

train_dataset = dataset.sample(frac=0.8, random_state=0)
test_dataset = dataset.drop(train_dataset.index)

train_dataset.describe().transpose()

train_features = train_dataset.copy()
test_features = test_dataset.copy()

train_labels = train_features.pop('MPG')
test_labels = test_features.pop('MPG')

train_dataset.describe().transpose()[['mean', 'std']]

normalizer = preprocessing.Normalization()

normalizer.adapt(np.array(train_features))
```

```
first = np.array(train_features[:1])

def build_and_compile_model(norm, N):
  model = keras.Sequential([
      norm,
      layers.Dense(N, use_bias=False, activation='sigmoid'),
      layers.Dense(1, use_bias=False)
  ])

  model.compile(loss='mean_squared_error',
                optimizer=tf.keras.optimizers.SGD(1/N))

  return model

dnn_model = build_and_compile_model(normalizer, widthHiddenLayer)
dnn_model.summary()

history = dnn_model.fit(
    train_features, train_labels,
    validation_split=0.2, epochs=widthHiddenLayer//10, callbacks=[tensorboard_callba

test_results = {}

test_results['dnn_model'] = dnn_model.evaluate(test_features, test_labels)
```

References

[1] Justin A. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.*, 80:725–752, 2020. 1, 3, 4

[2] N. Gigli L. Ambrosio and G. Savare. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2008. 3

[3] B Billingsley. *Convergence of probability measures.* New York, Wiley, 1968. 3

[4] A.D. Gottlieb. *Markov transitions and the propagation of chaos.* PhD thesis, University of California, Berkeley, ProQuest LLC, Ann Arbor, MI., 1998. 4

[5] J. Quinlan. Combining instance-based and model-based learning. In *ICML*, 1993. 5

[6] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org. 5