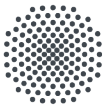
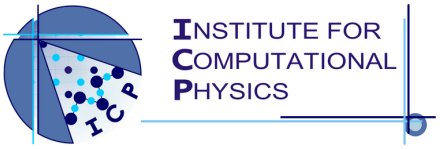


University of Stuttgart
Germany

Investigating the Evolution of Eisher Information for Neural Network Dynamics

Marc Sauter

ICP



University of stuttgart

A thesis presented for the degree of

B.Sc.

2023

Algebra is like sheet music. The important thing isn't can you read music, its can you hear it.

— Cristopher Nolan

Acknowledgements

I would like to express ...

Declaration

I, Marc Sauter, declare that ...

Signature	*Date*

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Keywords— Keyword1 - Keyword2 - Keyword3

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI . . . . . Artificial Intelligence

ML . . . . . Machine Learning

ReLU . . . . . Rectified Linear Unit

NN . . . . . Neural Network

GD . . . . . Gradient Descent

SGD . . . . . Stochastic Gradient Descent

NTK . . . . . Neural Tangent Kernel

FI . . . . . Fisher Information

1 | Machine Learning Basics

1.1 Introduction to machine learning

Over 70 years ago in October of 1950, at a time when computers weighed several tons, could only perform a few thousand operations per second and the pinnacle of machine intelligence were analogous robots that could follow light sources [1], Alan Turing published a paper in the journal of Nature discussing the question "Can machines think?" [2]. In this paper, Turing tries to tackle the question by proposing a game he calls the "imitation game". This game puts a human, whom we will refer to as Alice, in a room where she can communicate via written messages with two different entities, one of which being a human called Bob, the other being a machine. Alice's goal is to determine from this simple communication alone which of the two entities is the human. The goal of both the machine and Bob is to convince Alice that they are the human.

The largest execution of such an experiment to date took place in early 2023 in the form of an online chat portal where players had two minutes to talk to either another human or an Artificial Intelligence (AI) without knowing the type of their interlocutor. After more than two million participants had played the game for a total of more than 15 million conversations, only $68\,\%$ of the attempted classifications were correct guesses.

All of the advanced AI-bots used in this experiment were achieved using machine learning methods. The Oxford Learning Dictionary defines machine learning as "type of artificial intelligence in which computers use large amounts of data to learn how to do tasks rather than being programmed to do them" [3]. The theoretical foundation of such will be explained in the following sections.

Section 1.2 discusses the workings of neurons and how they work together to form

neural networks. Section 1.3 covers how neural networks are trained by discussing how to structure datasets, define loss functions and optimize network behavior.

1.2 Neural networks

1.2.1 Neurons

The term "**neural network**" (NN) is a reference to the workings of nervous systems of humans [4]. These systems consist of a net of neurons, biological cells that are intricately connected to other neurons through structures called synapses [5]. These connections carry electric pulses between the different neurons that can excite them when they exceed certain thresholds. These thresholds vary from neuron to neuron and change over time. When this excitation occurs, new pulses in turn propagate from the excited neuron outwards to possibly excite other neurons. This interplay between excitation and transmission through the network of neurons may create what we perceive as thinking.

Mathematical analyses of these systems in attempts to eventually understand and replicate this thinking process have been done as early as the 1940's [6]. The artificial neural networks used in machine learning today are mathematical concepts that replicate the transmission of excitation between neurons. To examine how this is achieved using mathematical functions and values instead of biological cells and electric pulses, let's look at how the artificial neural networks are built.

The neurons, which are the building blocks of the artificial neural networks, are mathematical entities that take a fixed number of scalar values as inputs and convert them into a single output value. For a more visual explanation let's take a look at Fig. 1.1, which illustrates the example of a neuron with 2 inputs. The big circle in the middle represents the neuron itself. It takes the activation values a_i as input, multiplies them with their corresponding weights ω_i and sums them

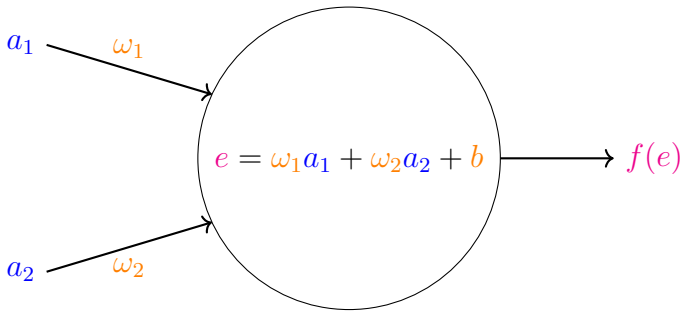


Figure 1.1: This figure aids the explanation of the operating principle of neurons in neural networks. The parameters ω_i are denoted in orange, the input activations a_i in blue and the activation e with its corresponding activation function f in magenta.

up to obtain the e, adds a bias value b and then applies the activation function to obtain the resulting output value. The input activations a_i correspond to the strength of electronic pulses in the nervous system. No pulse in the biological system would be represented by an activation of zero in the mathematical model. The weights ω_i are a representation of how important single input values are for the activation of the neuron. In the biological counterpart this might correspond to how thick or conductive the connections between the nerve cells are. Finally, the combination of bias b and activation function determines how large the sum of the input-weight-pairs has to be to activate the neuron and how the resulting value for the activation of the neuron changes for higher input activations. For example, a simple output activation function would be the ReLU function (Rectified Linear Unit). This function is defined as

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{1.2.1}$$

When applying a ReLU activation function, the neuron is activated as soon as the sum of the input-weight-pairs is larger than $-b$ and the value of the activation

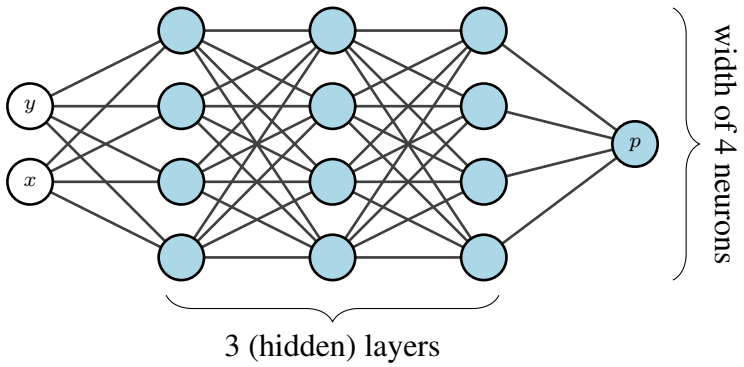


Figure 1.2: This figure shows an example of a neural network. It consists of multiple neurons connected to each other. This network takes in 2 input values and returns one output value. It consists of 3 hidden layers, each having a width of 4 neurons.

increases linearly with e.

1.2.2 Neural networks

A neural network can be built from these neurons by connecting the outputs of neurons to the inputs of others. An example of such a network is illustrated in Fig. 1.2. This neural network consists of three layers of four neurons each, takes two values as input and outputs one value. For example, it could be used as an approximator of whether a point on a 2D-grid is inside or outside of a given region. The input values would be the x and y coordinates of the point and the output value could represent the predicted probability that the point is inside the desired region. How to find parameters that make the network correctly classify a region will be explained in Section 1.3.

This network serves as an illustrative example, showing what a neural network can look like. For real-world applications, there are various kinds of networks used to learn different tasks [7]. For this thesis we will only talk about "fully connected" or "dense" neural networks. This means that every neuron in the first layer

will receive every possible input value, and every neuron in later layers will receive the output of every neuron in the previous layer as an input. All neurons are equipped with the same activation function. The weights and biases vary throughout. How the output of the network gets handled may still differ through different use cases.

When working on classification tasks, it can be beneficial to feed the output of the neural network into a softmax before analyzing it. This function converts the scalar outputs of the network, which can be any number from $\mathbb{R}$, into values corresponding to probability predictions for the different classes. For a set of outputs z_i, the standard softmax function is defined as

$$\sigma_i(z) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}. \tag{1.2.2}$$

The resulting values lie between 0 and 1 and add up to a total of 1. A larger value of an output node results in a higher corresponding probability. For further details on the softmax function, see [8].

The structure of a neural network is generally referred to as the "architecture" of the network, with the hidden layers or simply layers referring to the columns of neurons in between the input values and output neurons and the amount of neurons per layer referred to as the "width" of the network. This is also denoted in Fig. 1.2.

1.2.3 Mathematical view

The previous explanations have been very visual and step by step to make the topic more accessible. However, these concepts can be broken down to rather short mathematical expressions.

To start off, we can define the inputs as $a_i^{(0)}$, $i = 1, \cdots n$. The weights of the first hidden layer can be denoted by $\omega_i^{(1)}$, $i = 1, \cdots n$. Further, we define $\omega_{i,j}^{(k)}$ as

the weight that connects the i-th neuron in the k-th layer to the j-th neuron in the $(k-1)$-th layer. The maximum values of i and j depend on the widths of the respective layers, k can reach values between 1 and the amount of hidden layers plus 1. The bias of the i-th neuron in the k-th layer is denoted as $b_i^{(k)}$. Using this notation we can write out the output of the i-th neuron in the $k+1$-th layer as

$$a_i^{(k+1)} = f\left(\sum_j \left(\omega_{i,j}^{(k+1)} a_j^{(k)}\right) + b_i^{(k+1)}\right). \tag{1.2.3}$$

To actually calculate this value, the activations $a_j^{(k)}$ have to be recursively replaced with their full calculation until one arrives at the input values of the network.
For simplicity reasons we will refer to the weights $\omega_{i,j}^{(k)}$ and biases $b_i^{(k)}$ together as "**parameters**" of the neural network. We will denote these collected in one ordered set as $\theta = \{\theta_i\}_{i=1}^N$ if N is the total number of weights and biases added together. The mapping of the parameters onto θ can be arbitrarily chosen. It has to be known, so that it's possible to calculate the output of the network when given the parameters in the same way as when given the actual weights and biases. Another possible representation that we will use often is to write this set as a vector $\theta \in \mathbb{R}^N$. A short notation for the output of the neural network will be explained in the next section.

1.3 Training of neural networks

In the previous sections we discussed how neural networks are built up. Now we will give an example of how they can be trained and used to perform specific tasks. Here, the term training refers to the search for a set of parameters θ that make a neural network with a fixed architecture behave in a desired way. Specifically, we will look at how they can be trained on data sets in a process referred to as

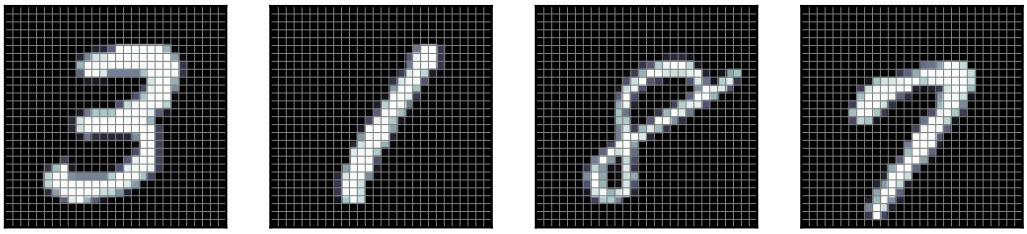


Figure 1.3: Displayed in this figure are 4 examples of input values of the MNIST dataset. This dataset represents handwritten digits from 0 to 9. Each input value consists of a 28 by 28 grid of grayscaled pixel values.

supervised learning. In supervised learning, the desired behavior of a network is defined by a dataset consisting of input and output values [9]. The network should behave in such a way that it produces the desired outputs when fed the corresponding inputs. Training consists of defining a measure of performance of the network, called the loss function, and then using mathematical iterative methods for minimizing the loss.

1.3.1 Datasets

First we need to assume that we have a given dataset containing input-output pairs that represent the task we want the neural network to perform. The neural network is supposed to learn which output is supposed to be generated from an input by evaluating the given inputs and desired outputs. We will call those desired outputs "**targets**" to distinct them from the actual outputs of the neural network. A good example of this would be the MNIST database, which was first used in 1994 in [10] and has since become a popular entry-level classification task for machine learning. Examples of input data for this database are shown in Fig. 1.3. This dataset consists of various handwritten digits from 0 to 9 represented by a 28 by 28 grid of grayscale pixel values. The values of the pixels in these grids act as the 784 input values for the neural network. The targets and output of the neural net-

work should be of the same shape. For example one might use a scalar output of the neural network that should be equal to the value of the digit drawn in the input pixel grid. In this case the targets are scalar values from 0 to 9 corresponding to their input pictures. Another way would be to use 10 output values together with the previously mentioned softmax function (see Section 1.2.2) to recieve the values as probabilities for the different numbers as output. We would then change our targets to vectors with 10 elements, where the entry corresponding to the handwritten digit in the input picture would be 1, every other value would be 0. For example, in this case, an image depicting the number 3 would correspond to a target of $(0, 0, 0, 1, 0, 0, 0, 0, 0, 0)$.

We will denote the data sets as mathematical sets of input-target pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ are the input values and $\mathbf{y}_i$ are the targets. Here, we referred to the total number of training samples as N. The mathematical dimension of $\mathbf{x}_i$ and $\mathbf{y}_i$ can vary and are dependent on the network architecture, but we will assume that they are vectors of the real numbers $\mathbf{x}_i \in \mathbb{R}^a$ $\mathbf{y}_i \in \mathbb{R}^b$. Due to this property we also sometimes refer to them as "points". If our inputs values are organized in a different manner, for example the MNIST data being matrices of $\mathbb{R}^{28\times 28}$, we can rearrange these numbers into a vector of $\mathbb{R}^{784}$ in any way we want, as long as every input is rearranged in the same manner.

Going further, we will denote an output of the neural network for an input $\mathbf{x}_i$ by $f_\theta(\mathbf{x}_i)$. This means that we mathematically describe the entire neural network as a function

$$f : \mathbb{R}^a \times \mathbb{R}^p \to \mathbb{R}^b \tag{1.3.1}$$

with p being the amount of parameters in the network.

1.3.2 Loss function

Thus far, we have defined what a neural network is and how the input data we want to train on is structured. In order to train our networks to learn the underlying function of our dataset, we now need to introduce a way to measure how well our network is performing. Once we can evaluate the performance of our network, we can then introduce ways to optimize that performance.

This measure of the performance of a neural network is called the "**loss function**" $\mathscr{L}$. It is also sometimes referred to as the cost function in literature [11]. Within this work, we will only consider loss functions of the form

$$\mathscr{L}\left(\{(f_\theta(\mathbf{x}_i), \mathbf{y}_i)\}_{i=1}^N\right) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f_\theta(\mathbf{x}_i), \mathbf{y}_i\right). \tag{1.3.2}$$

Let's call ℓ the subloss of the system. As a reminder, θ is a set containing the parameters of the neural network, f_θ is its output function. Sometimes it's more convenient to explicitly denote the dependency on theta, which would make the loss function

$$\mathscr{L}\left(\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \theta\right) = \frac{1}{N}\sum_{i=1}^{N} \ell_\theta\left(\mathbf{x}_i, \mathbf{y}_i\right). \tag{1.3.3}$$

How exactly the loss function should be defined cannot be generally stated. In any case, the value of the loss function should generally be smaller the closer the output of the neural network is to the corresponding targets.

In most of our cases we use loss functions with

$$\ell_\theta\left(\mathbf{x}_i, \mathbf{y}_i\right) = d\left(f_\theta(\mathbf{x}_i), \mathbf{y}_i\right), \tag{1.3.4}$$

where d represents the distance between the output vector of the neural network and the target vector according to different metrics. One very would be

$$\mathscr{L}\left(\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}, \theta\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(f_\theta(\mathbf{x}_i)_j - y_{i,j}\right)^2, \tag{1.3.5}$$

with $f_\theta(\mathbf{x}_i)_j$ and $y_{i,j}$ being the j-th component of the network output and target vectors. Further examples and loss functions of different forms can be found in [12].

1.3.3 Optimization

To quickly recap, we have now defined what a neural network is and that we want a fixed network architecture during training, for which the parameters can vary to optimize the performance. We also assume that we have a data set that we want our network to learn the structure of, and a loss function that measures how well the current setup of our network is performing. We will now talk about how we can change the parameters to optimize performance.

This is achieved by minimizing the value of the loss function. The simplest and most common algorithm to do this with is **Gradient Descent** (GD). Here, the rule for updating the parameters looks like [13]

$$\theta' = \theta - \eta\nabla_\theta\mathscr{L}\left[\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}, \theta\right], \tag{1.3.6}$$

where η is a parameter of training called the "learning rate", which is a scalar value for GD but could also vary through training and even be a higher rank tensor for more advanced optimization methods [14].

The idea behind GD is that the gradient with respect to θ points in the direction of the steepest ascent of the loss when the parameters are varied. If we subtract

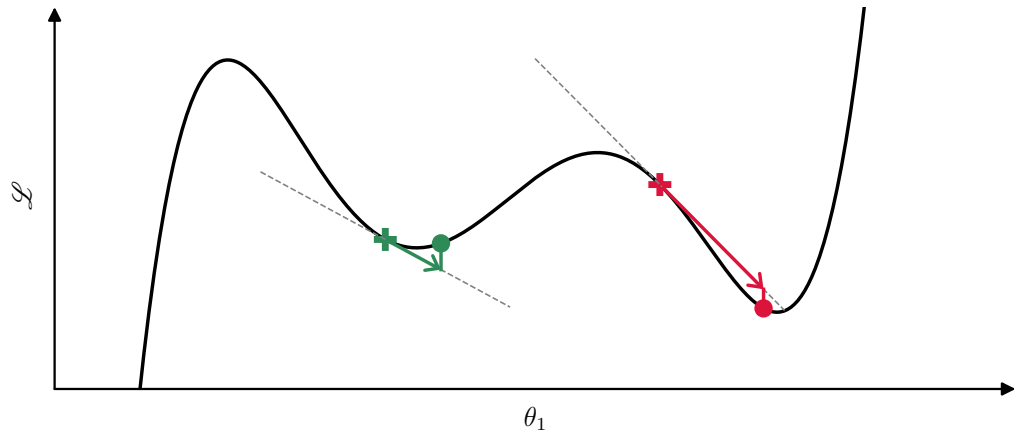


Figure 1.4: Visual explanation of the Gradient Descent algorithm. The dot markers show where a parameter that starts at the cross marker might end up if the algorithm is applied.

this gradient from the parameters we update the loss in the opposite direction, which is the direction of steepest descent. To visualize this, let's examine a single parameter θ_1. The update rule for this single parameter can be obtained from writing out the previous Eq. (1.3.6) in its full vectorial component form and then picking out the line of θ_1. It reads

$$\theta_1' = \theta - \eta \frac{\partial}{\partial \theta_1} \mathscr{L} \left[\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \theta \right] . \tag{1.3.7}$$

To explain the concept of this algorithm more visually let's take a look at Fig. 1.4. The cross markings symbolize the position of parameter θ_1 before the update step. The algorithm then calculates the derivative of the loss $\frac{\partial}{\partial \theta_1} \mathscr{L}$, which geometrically is the slope of the tangent. To update the parameter, the slope is multiplied with -1 times the learning rate η and added to θ_1, which is represented by the arrows. These arrows are adjusted in length so that the distance they cover in the direction of θ_1 is η times the derivative. One can see that the red arrow on the right covers more distance in the direction of θ_1 than the green one on the left, which

is a result of the corresponding slope being steeper. The circle markers then mark the updated parameter value θ_1'.

This process of updating the parameters is repeated several times. One update is called an update step. When using batches of the dataset for parameter updates, an epoch refers to a period of training after which the algorithm calculated gradients on each input-output data pair. The number of update steps in an epoch depends on the batch size.

This visual example already illustrates some of the drawbacks of the GD algorithm: Starting training with the green parameter will yield a higher final loss than starting with the red parameter. Also going further to the left would make the loss function smaller than both of these local minima. GD just tends to find the closest local minimum instead of a global one [15]. Another drawback results from the step size being finitely big, which makes the green parameter hop over the local minimum. Both of these problems can't be solved by simply optimizing the learning rate η in general.

1.3.4 Other optimizers

The standard GD calculates the gradient for parameter updates from the whole data set according to Eq. (1.3.6). In most cases it is sufficient or even beneficial to compute the loss function and its gradient for a subset of the whole data set [15], which is usually called mini-batches. The extreme end of this would be to calculate the gradient and update the parameters for one data point each, which is then called **stochastic Gradient Descent**. GD and the variations are very simple and fast to compute algorithms, but they can only find local minima and can quite easily hop over the actual minima. That's why other algorithms have been developed that try to solve or mitigate some of the problems of GD. For the training, which will be performed later in this thesis, the so-called ADAM optimizer was

used. The intricacies of this optimization method are beyond the scope of this work. We refer the interested reader to [16], where the details of this algorithm are discussed. Another optimization algorithm that can be used to overcome some of the problems of GD is the Natural Gradient Descent, which will be mentioned in later sections.

1.4 Which assumptions are actually necessary

In the previous sections we took a brief look at machine learning and neural networks. Although this barely scratches the surface of the methods that are used today, it's still more than what's needed for the upcoming chapters. The specific methods were given to explain how the NNs that will come up later in this thesis were trained, but are unnecessary restrictions that don't need to be made for the mathematical considerations coming up.

For those it is sufficient to view systems $f_\theta(\mathbf{x}_i)$ that depend on parameters θ, accept input vectors $\mathbf{x}_i$ of constant dimension and can be evaluated through the formalism of the loss function from equation Eq. (1.3.3). For the discussions of the NTK, we can even disregard the assumption of the loss function splitting up into a sum of ℓ functions. This means that the exact properties of the neural networks we looked at earlier can be ignored, allowing us to generalize the observations to many more network architectures or completely different systems than previously mentioned.

2 | Fisher Information

2.1 Use in Statistics

To introduce the **Fisher Information** (FI), we will start off with how it is defined and used in statistics.

Let's look at a **statistical model** $f(x_i|\theta)$ that represents how a parameter θ is related to the outcomes x_i of random variables X_i[17]. As an example of what this means, let us look at the statistical model of a Galton board. For readers who are not familiar with a what Galton boards are, there is an example photograph in Fig. 2.1. It's a famous mechanical model that visualizes binomial distributions, which are approximations of the normal distribution. If we place many balls at the top of the board and let them fall to the bottom, the amount of balls that end up in each cell are distributed according to the binomial distribution. In this case, x_i could assume the slot number which a ball can fall into. The i could label multiple throws into the board, but for now we'll assume that there is only one experiment i. To introduce a parameter that influences the distribution of balls, let's say one can throw from different spots above the Galton board which we now control with the the value of θ. For a known θ, the resulting value of the statistical model represents the probability distribution $f(x_i|\theta) = p_\theta(x_i)$ for the probability of the different x_i outcomes. As a sidenote, if we instead fixed the value of x_i and viewed the statistical model as a function of θ it would be called a **likelihood function**. A visual representation of the probability distributions for different θ can be seen in Fig. 2.2.

In general, the statistical models might be more complex, where θ contains several parameters, x being an element of a mathematical space other than $\mathbb{R}$ and the index i denoting various different experiments, all depending on the same param-

Figure 2.1: This figure shows a picture of a Galton board. Taken from [18].

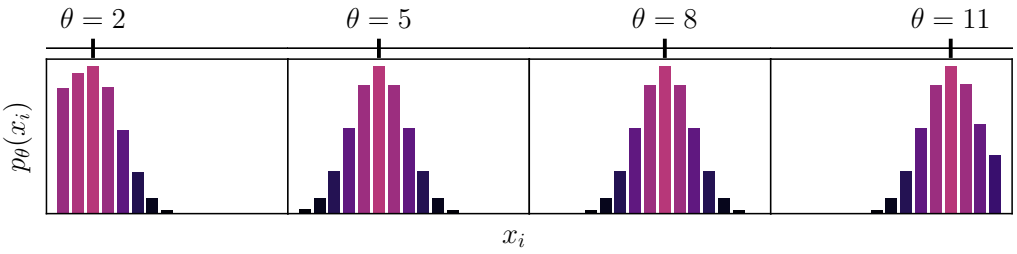


Figure 2.2: This figure shows the probability distributions for a Galton boards with different drop-in positions. The slots where the ball can end up are labeled by the value of x_i.

eter, but having different possible outcomes and probability distributions.
What's of interest for the FI are cases where the parameters are not known before conducting the experiment, and have to be approximated by the different outcomes x_i.

Before we introduce the FI, let's look at an example from [17]. Let's consider a biased coin where we denote the probability of heads ($x_i = 1$) with θ and the probability of tails ($x_i = 0$) with $1 - \theta$. We will now take a look at the outcome of n tosses, represented by the variable X^n. For example, an observed result for X^5 could be $x^5 = (1, 1, 0, 1, 0)$. Let's consider another variable Y, whose observation is the sum of the total head throws $y = \sum x^n$. In our example case of $x^5 = (1, 1, 0, 1, 0)$, this would result in a value of $y = 3$. The probability for this variable y is distributed according to the binomial distribution $f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. Here the binomial coefficient $\binom{n}{y}$ represents the different combinations that result in the same value of y. This is needed because there are 2^n different possibilities for x, while there are only n different possibilities for y.
If we now fix the outcome of y and look at the conditional probability for the different x^n that could have resulted in that y value, we get $p(x|y, \theta) = 1/\binom{n}{y}$. With $p(x|y, \theta)$ we denote the probability depending on x for fixed y and θ. Although the probability of y and x both depend on θ, the probability for x when y is fixed doesn't. After measuring y there is no information about θ left in the measurement of x. This means that y is fully descriptive of or "sufficient for" the parameter θ. Measuring y results in the same amount of information about the parameter θ as measuring the whole observation x. To quantify how much information a certain function of outputs $t(x)$, which is $y(x^n)$ in our example, contains about the parameters θ, Fisher introduced the Fisher information.

The Fisher information is defined as

$$I_{X,ij}(\theta) = \underset{x\in X}{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log f(x|\theta)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta_j}\log f(x|\theta)\right], \tag{2.1.1}$$

where we used the expectation E

$$\underset{x\in X}{E}[A(x)] = \begin{cases}\sum_{x\in X}(A(x)p(x)) & \text{if } X \text{ is discrete,}\\ \int_{x\in X} A(x)p(x)\mathrm{d}x & \text{if } X \text{ is continuous.}\end{cases} \tag{2.1.2}$$

We will later use an alternative notation where we denote $\log f$ as ℓ. As will be evident later, this notation does not interfere with the definition of ℓ in Eq. (1.3.3). Since the Fisher information is dependent of θ, we can assume the value of θ to be fixed during calculation, which makes $f(x|\theta)$ equal to the probability density $p_\theta(x)$.

For n independent experiments X^n, where $f(x^n|\theta) = \prod_{i=1}^n f(x_i|\theta)$, one can split the FI into

$$I_{X^n,ij}(\theta) = \prod_{i=1}^{n} I_{X_i,ij}(\theta). \tag{2.1.3}$$

A proof of this can be found in Section A.1.

For our example the FI yields $I_{X^n}(\theta) = I_Y(\theta) = n/(\theta(1-\theta))$[17]. This means that there is as much information about the θ contained in the measurement of Y as in the measurement of X^n, which coincides with Y being a sufficient measurement for θ.

To give another example of how the FI represents the information obtainable about a parameter from a measurement, let's consider the family of normal distributions

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}. \tag{2.1.4}$$

These will now act as our statistical model $p_\theta(x) = \mathcal{N}(x|\theta)$ where $\theta = \{\theta_1, \theta_2\} =$

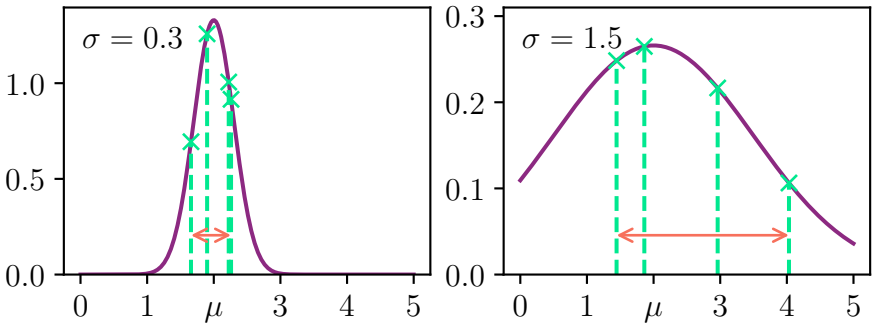


Figure 2.3: This figure shows two normal distributions centered around $\mu = 2$ with varying σ parameters. It also shows 4 samples chosen randomly according to the distribution. It's visible that for the case of a smaller variance σ, the points tend to be closer to the center and also less spread apart, which makes the information about μ contained in a measurement larger for a smaller variance.

$\{\mu, \sigma\}$. An observation would consist of a resulting value $x \in \mathbb{R}$, with it's probability distributed according to the statistical model. The FI from equation Eq. (2.1.1) can be derived as

$$I(\mu, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}. \tag{2.1.5}$$

We can now interpret the diagonal elements as measures of how much information a measurement contains about the corresponding parameter, and the off-diagonal elements as measurements of how similar the model changes when varying the corresponding parameters. To give a specific example, let's look at the diagonal component corresponding to μ, $I_{11}(\mu, \sigma) = 1/\sigma^2$. This value indicates that for smaller σ, random values drawn from the distribution contain more information about μ than samples drawn from distributions with larger σ. For a visual guide, consider Fig. 2.3. It can be seen that the smaller value of σ results in a narrower spread of randomly drawn values, as indicated by the orange arrow. The values also tend to be closer to the mean value μ the smaller the variance σ is. Therefore,

if we had to predict the value of μ from knowing only a few drawn samples, it would be easier to use the values drawn from the distribution with the smaller variance, because the information contained in the samples, measured by the FI, is greater there. Keep in mind that all of these drawn values are randomly distributed. Therefore it is also possible to have two sets of random samples where the samples from the larger variance are better at predicting μ, but statistically speaking the smaller variance tends to perform better.

To conclude this chapter, the Fisher information is used in statistics to measure the amount of information one can gather about a parameter θ by measuring the outcome of a probability distribution $p_\theta(x_i)$. It is defined in equation Eq. (2.1.1).

2.2 Fisher Information as Riemannian metric

Previously we talked about the FI in the context of statistics. You may wonder why we went over a statistical method to describe probabilities, when we previously only talked about machine learning and neural networks. The answer to this question is, that the Fisher information matrix, defined as

$$I_{ij}(\theta) = \underset{(\mathbf{x}_i,\mathbf{y}_i)\in D}{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\ell_\theta(\mathbf{x}_i,\mathbf{y}_i)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta_j}\ell_\theta(\mathbf{x}_i,\mathbf{y}_i)\right], \tag{2.2.1}$$

acts as the Riemannian metric describing the statistical manifold of the network regarding its loss. Note that we dropped the index of I that referred to the set over which the expectation will be calculated for visibility reasons. A brief introduction to this topic will be provided in this chapter. Keep in mind that this is mostly for intuition purposes and we will only cover a few important definitions. For more details, please refer to [19] where most of the following information is taken from. If you're not interested in mathematical details you can skip this chapter and go

directly to **FILL SOMETHING IN HERE**.

2.2.1 Differentiable manifolds

To state the definition, a n-dimensional manifold S is a topological space so that for every point you can define a neighborhood around that point which is homeomorphic to an open subset of $\mathbb{R}^n$. A good example of this would be the surface of the earth, where locally viewed in the scale that we usually see things, the earth appears flat, but on a global scale the earth is obviously a sphere. This results for example in the shortest path between to points not being a straight line in maps of the world as a whole. Also the angles of a triangle don't sum up to $180°$ as they would in a subspace of R^n. This results from conventional maps being subspaces of $\mathbb{R}^2$, although the earth is only homeomorphic to $\mathbb{R}^2$ in smaller local scales. If one tries to map the whole sphere into a map without gaps, one has to map the coordinates in a way that makes the shortest lines curved for example.
Let's come back to the statistical models $f(x|\theta)$ we talked about in the last section. We will treat the models considering fixed parameters as probability distributions $p_\theta(x)$ in this context. If the probabilities are sufficiently smooth in θ, which means that they are differentiable in θ as often as needed for further considerations, one can view the family of probabilities as a n-dimensional manifold, where the n different θ components play the role of the coordinate system of the manifold.
For example let's consider normal distributions

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}, \tag{2.2.2}$$

where $\theta = \{\theta_1, \theta_2\} = \{\mu, \sigma\}$. We can now consider this family of distributions as a manifold, displayed in Fig. 2.4. This is like a space, where every point in the space represents a distribution $p(x|\theta)$.

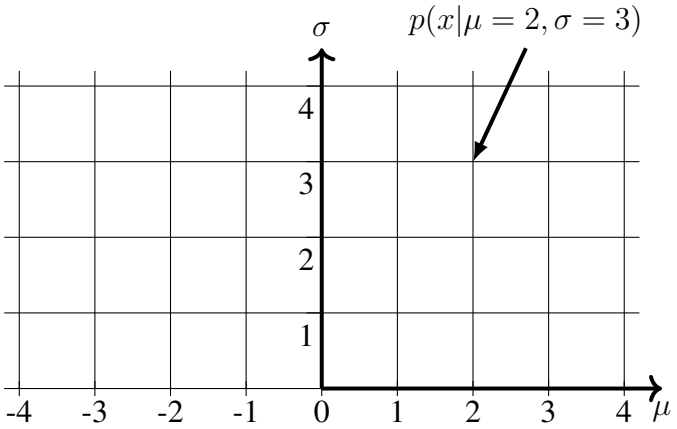


Figure 2.4: This figure illustrates the manifold of normal distributions. As coordinate system, μ and σ are used. Every point in this manifold represents a probability distribution, as indicated by the arrow.

It might also be clear that the coordinate system of a manifold is definable in multiple different ways. Although it's always given in our use case, let's therefore denote that in general when we have coordinates θ we also need a mapping ϕ, which maps coordinates to points on a manifold. This means that by applying $\phi(p)$ to a point $p \in S$ the resulting vector in $\mathbb{R}^n$ resembles the coordinates of that point. We can also apply the inverse of that mapping to a set of coordinates to the point in the manifold that's represented by those coordinates.

2.2.2 Tangent space

The tangent space T_p of a manifold at point p is a vector space obtained by linearization of the manifold around p. For intuition purposes, let's take a look at the tangent plane of a 2d-surface in Section 2.2.2. Here the tangent space is simply a plane that touches the surface in one point, with derivatives adjusted to match the surface at that point. For the general case of n-dimensional manifolds, it is obvious that tangent spaces aren't simply tangent planes of surfaces in every case, therefore let's introduce a way how to calculate a tangent space.

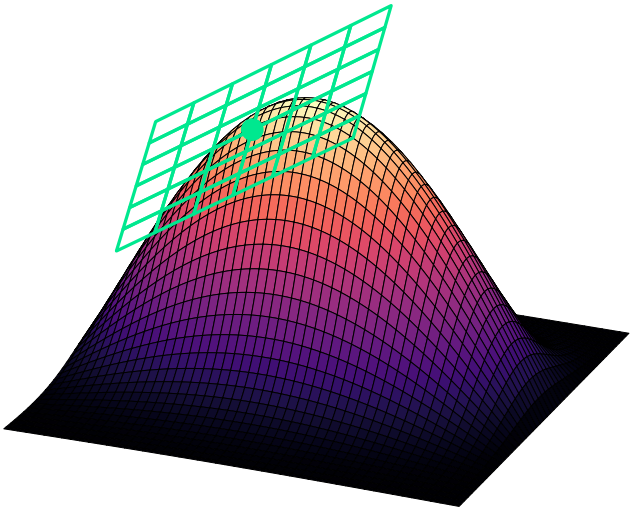


Figure 2.5: This figure contains an example of a tangent space of a manifold, which in this case is a 2D-surface.

First we will define curves $c(t)$ that are continuous mappings from an interval $[a, b] \in \mathbb{R}$ into the manifold S. In the parametric representation, the curve is given by $\theta(t)$. Now we can define what a tangent vector is.
Imagine a smooth real function $f(\theta) : S \to \mathbb{R}$. We can now restrict this function to our predefined curve c by $f \circ c : [a, b] \to \mathbb{R}$. We'll denote this via $f(\{\theta(t)\})$ in the coordinate expression. The derivative Cf of this function is then given by

$$Cf = \frac{\mathrm{d}f \circ c}{\mathrm{d}t} = \sum_{i=1}^{n} \frac{\partial f}{\partial \theta_i} \frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t} = \underbrace{\left(\sum_{i=1}^{n} \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \frac{\partial}{\partial \theta_i} \right)}_{\text{Operator } C} f. \tag{2.2.3}$$

Therefore we can associate a directional derivative operator C with each curve. The only dependence of this operator regarding the curve is $\frac{\mathrm{d}\theta_i}{\mathrm{d}t}$, where we might also clarify that this derivative depends itself on the point where it is calculated at. If the manifold is infinitely differentiable, the set of these mappings C at a fixed point on the manifold forms a n-dimensional vector space, called the "**tangent**

space" T_p of that point p.

To make this more clear, let's look at the simplest basis for this vector space, which we'll call the "natural basis" of the tangent space. For this we'll consider curves $c_1, c_2, \dots c_n$ through a point p_0, where the curves are defined as

$$c_i(t) = \{\theta_1^0, \theta_2^0, \dots, \theta_i^0 + (t - t_0), \dots \theta_n^0\}, \tag{2.2.4}$$

so that $c_i(t_0) = \{\theta_1^0, \theta_2^0, \dots \theta_n^0\} = p_0$ for every curve. The tangent vectors C_i then are simply the derivatives regarding their corresponding coordinate $C_i f = \partial/\partial\theta_i f$. We will denote this in short by $C_i = \partial_i$. The n vectors ∂_i are linearly independent and form the natural basis for the tangent space. This means that any tangent vector A can be represented by

$$A = \sum_{I=1}^{n} A_i \partial_i, \tag{2.2.5}$$

with components with respect to the natural basis A_i. If we are given a curve $c(t)$ going through $c(t_0)$ which derivative operator C in t_0 is equivalent to the vector A, we can find the natural basis representation of A as

$$A_i = \left.\frac{\mathrm{d}\theta_i}{\mathrm{d}t}\right|_{t_0}. \tag{2.2.6}$$

Now let's take a look at the case of manifolds of statistical models. First, we will revisit the definition

$$\ell(x|\theta) = \log f(x|\theta), \tag{2.2.7}$$

and assume that for fixed θ, the n-functions $\partial_i \ell(x|\theta)$ are linearly independent. This means that we can construct a vector space by defining

$$T_\theta^{(1)} = \{A(x)|A(x) = A_i \partial_i \ell(x|\theta)\}. \tag{2.2.8}$$

We do this because there is a natural isomorphism between the tangent space T_θ and this space $T_\theta^{(1)}$ through

$$\partial_i \in T_\theta \leftrightarrow \partial_i \ell(x|\theta) \in T_\theta^{(1)}. \tag{2.2.9}$$

We will call $T_\theta(1)$ the "**1-representation**" of our tangent space for the statistical models. We will now use this vector to define an inner product on the tangent space and it's 1-representation.

2.2.3 Riemannian metric and Fisher Information

When the inner product of the tangent spaces T_p is defined, the manifold is called a **Riemannian manifold**.

Let's first consider the inner product of the 1-representation space. Let $A(x)$ and $B(x)$ be 1-representations of A and $B \in T_\theta$. It is intuitive to define the inner product as

$$\langle A(x), B(x) \rangle = \underset{x \in X}{E}\Big[A(x)B(x)\Big], \tag{2.2.10}$$

with the expectation value $E[\cdot]$. Since the tangent space is isomorphic to its 1-representation, the inner product also translates via

$$\langle A, B \rangle = \langle A(x), B(x) \rangle. \tag{2.2.11}$$

This also means that we can calculate the inner product of the basis vectors as

$$\begin{aligned} g_{ij}(\theta) := \langle \partial_i, \partial_j \rangle &= \langle \partial_i \ell(x|\theta), \partial_j \ell(x|\theta) \rangle \\ &= \underset{x \in X}{E} \Big[\partial_i \ell(x|\theta), \partial_j \ell(x|\theta) \Big]. \end{aligned} \tag{2.2.12}$$

The resulting object $g_{ij}(\theta)$ is called the **Riemannian metric tensor** of the manifold. We can see that the Riemannian metric tensor for the statistical model is equivalent to the Fisher Information. It might be of interest to note here that we assume that ℓ only depends explicitly on θ. If we denote these as implicit dependencies, we have to replace the partial derivatives with absolute ones.
The inner product of two vectors can now be expressed with the metric tensor as

$$\langle A, B \rangle = \sum_{i,j} A_i B_j g_{ij}(\theta) \tag{2.2.13}$$

in the component form.
Using this representation of the inner product, we can define various things. For example, the length of a vector A is defined as $|A|^2 = \sum_{i,j} A_i A_j g_{ij}$, the orthogonality of two vectors when their inner product is zero, and the distance between two points $\theta^{(0)}$ and $\theta^{(1)}$ along the curve c is defined by

$$s = \int_{t_0}^{t_1} \sum_{i,j} \sqrt{g_{ij} \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \frac{\mathrm{d}\theta_j}{\mathrm{d}t}} \mathrm{d}t. \tag{2.2.14}$$

This also introduces the concept of **Riemannian geodesics**. These are the curves that connect two points via the minimal distance between the two.
Another representation of the metric tensor or the FI is

$$g_{ij}(\theta) = - \underset{x \in X}{E} \Big[\partial_i \partial_j \ell(x|\theta) \Big]. \tag{2.2.15}$$

A proof of this is denoted in Section A.2.

2.2.4 Scalar curvature and Christoffel symbols

Here we will only give the definitions needed to compute the scalar curvature of a statistical manifold. A full understanding requires much more mathematics than we will go over here. If you are interested in this, please see [19] and [20].
Note that we don't use covariant and contravariant indices, since we don't need them in the scope of this thesis. The notation of uppercase or lowercase indices is just to be consistent with the notation from general relativity and should be of no further concern for our experiments. The parameters with lowercase indices defined previously translate directly to the ones with uppercase indices here. We will also use Einstein notation for these equations. First we define the Christoffel-symbols which can be calculated from the Riemannian metric via

$$\Gamma^i_{jk} = \frac{1}{2} g^{im} \left(\frac{\partial g_{mk}}{\partial \theta^l} + \frac{\partial g_{ml}}{\partial \theta^k} - \frac{\partial g_{kl}}{\partial \theta^m} \right), \tag{2.2.16}$$

where $g^{ij} = (g^{-1})_{ij}$ are the components of the inverse of the metric. From those we can define the Riemannian curvature tensor as

$$R^i_{jkl} = \partial_k \Gamma^i_{jl} - \partial_l \Gamma^i_{jk} + \Gamma^i_{mk} \Gamma^m_{jl} - \Gamma^i_{ml} \Gamma^m_{jk}. \tag{2.2.17}$$

Finally, we can state the Ricci scalar curvature as

$$R = g^{ij} R^m_{imj}. \tag{2.2.18}$$

The Ricci scalar can be viewed as a measure of how much the manifold at the point of computation differs from flat space. For euclidean spaces, the Ricci scalar is 0 [20]. How the Ricci scalar can be a measure of complexity of a neural network

will be mentioned at the end of Section 2.3.

2.2.5 Application to neural networks

Now that we've laid down the mathematical fundamentals of how the FI can be considered the metric of a statistical manifold, we can apply this to our neural network optimization by introducing a way of viewing them as statistical models. Let's first define the statistical model by the probability $p_\theta(\mathbf{x}_i, \mathbf{y}_i) = e^{-\ell_\theta(\mathbf{x}_i, \mathbf{y}_i)}$, where ℓ is the subloss function defined in Eq. (1.3.3). You can think of this as a kind of probability that the network characterized by the parameters θ and the subloss function ℓ can generate the outputs $\mathbf{y}_i$ from the inputs $\mathbf{x}_i$. In reality, of course, it's not a probability that the network reproduces the output, since that would be either true or false. It's rather a measure of how close the network's prediction is to the actual target. We can think of it and use it as a probability because of the exponential function, which results in a probability of 1 if the network output matches the target (which means $\ell = 0$) and shrinks towards 0 for larger discrepancies.

Now we can also see that the definitions of two different ℓ don't interfere with each other. In the context of a statistical model, ℓ is defined as the logarithm of $f(x|\theta)$, in the context of neural network training, ℓ is defined as the subloss (see Eq. (1.3.3)) that we used for the probability above. This means that in the case of a statistical manifold of neural network training, the negative subloss $-\ell_\theta(\mathbf{x}_i\mathbf{y}_i)$ is equal to the logarithm of the statistical model $\ell(\mathbf{x}_i, \mathbf{y}_i|\theta)$. Since we will only use ℓ to calculate the FI, where the two minus signs cancel, we can think of the ℓ in the definition of the FI in Eq. (2.1.1) as being the subloss of a network defined in Eq. (1.3.3). How the components of the resulting matrix can be interpreted will be discussed at the end of Section 2.3.

2.3 Intuitive explanations

In the previous sections we introduced the Fisher Information, first in a statistical context and later as the metric of a statistical manifold to be able to apply some of it's insights to neural network training.
Since the information provided in the last sections has been mathematically abstract and lacking intuition, let's quickly recap the basics of what we need to know and understand about the Fisher Information for the rest of this work.
First let's state the definition again. The FI is defined as

$$\begin{aligned} I_{i,j} &= \underset{x\in X}{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log f(x|\theta)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta_j}\log f(x|\theta)\right] \\ &= \underset{(\mathbf{x}_i,\mathbf{y}_i)\in D}{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\ell_\theta(\mathbf{x}_i,\mathbf{y}_i)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta_j}\ell_\theta(\mathbf{x}_i,\mathbf{y}_i)\right], \end{aligned} \tag{2.3.1}$$

where the notation in the first line is used in the context of statistics and the one from the second line in the context of the manifold of neural network training.
We first introduced the FI as a statistical measure of how much information the measurement of a random variable contains about its underlying parameters.
For example let's say our experiment consists of throwing a biased coin, where the probability of heads is the underlying parameter of the random variable of the coin toss. We can now conduct an experiment with the goal of estimating the probability of the biased coin. Let's say we make 10 throws twice. The first time, we will use the full observation for the parameter estimation. We will exactly know the results of the coin toss *for every single one* of the 10 throws. The second time, we will only know how often heads came up *in total* for the 10 throws. We will lose the information about when exactly during these 10 trials we measured heads or tails. The Fisher Information with respect to the parameter yields the same value for both of these observables. This therefore means that the information about the

probability of heads contained in the measurement is the same for both observations. To estimate the bias of the coin, it is perfectly sufficient to only write down the total number of heads instead of the entire observation.

This how the Fisher Information was first introduced. Going further, we also introduced another way of obtaining the FI as the Riemannian metric of the statistical manifold corresponding to a neural network.

For this we considered a neural network along with a dataset and loss function as a statistical model. We did this by introducing a kind of probability $p_\theta(\mathbf{x}_i, \mathbf{y}_i) = \mathrm{e}^{-\ell_\theta(\mathbf{x}_i, \mathbf{y}_i)}$. We can view this as the probability that our network with parameters θ produces the correct outputs $\mathbf{y}_i$ when given the inputs $\mathbf{x}_i$. In this view, we consider the family of probabilities p_θ that are functions taking network input-output pairs as input. That family of probabilities forms a statistical manifold. The manifold is a space with coordinates $\theta = \{\theta_1, \cdots, \theta_n\}$, where every point in the space corresponds to a probability function that depends on the θ coordinates of our space. It is locally isometric to a subset of $\mathbb{R}^n$. Globally, euclidean geometry doesn't apply, which for example makes the shortest paths between points in the manifold curves in the coordinate system. When every point in the space is a probability function by itself, it's very hard to properly imagine lines between points and even the seemingly simple concept of distance becomes very abstract and confusing. Therefore, let's not dive too deeply into visualizations and first think about how we will measure distance and characterize curvature in this space.

To define distance in the manifold, we first introduced the concept of a tangent spaces. A tangent space is, again hard to grasp, an abstract vector space defined at a point in the manifold, where the vectors are differential operators corresponding to the derivatives along curves through the point. The reason we defined it is because through defining an inner product $\langle \cdot, \cdot \rangle$ on the tangent spaces, we can obtain

a metric on the manifold through $g_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle$. As mentioned before, the basis vectors $\mathbf{e}_i$ on the tangent spaces are abstract concepts and it's not trivial to find a definition of an inner product intuitively. That's why we introduced a new space, isomorph to the tangent space, where defining the inner product feels natural. The ismorphism was introduced as

$$\partial_i \leftrightarrow \partial_i \ell(x|\theta) \tag{2.3.2}$$

maps the derivative operators onto actual derivatives of the logarithm of the statistical model f. To obtain the inner product of two derivative operators we then defined

$$\langle \partial_i, \partial_j \rangle = \langle \partial_i \ell(x|\theta), \partial_j \ell(x|\theta) \rangle, \tag{2.3.3}$$

where we naturally assume the inner product between the two distributions as

$$\langle \partial_i \ell_\theta(\mathbf{x}_i, \mathbf{y}_i), \partial_j \ell_\theta(\mathbf{x}_i, \mathbf{y}_i) \rangle = \underset{(\mathbf{x}_i, \mathbf{y}_i) \in D}{E} \big[\partial_i \ell_\theta(\mathbf{x}_i, \mathbf{y}_i) \cdot \partial_j \ell_\theta(\mathbf{x}_i, \mathbf{y}_i) \big]. \tag{2.3.4}$$

Now we arrived at the definition of the metric of our manifold, which is equivalent to the Fisher Information matrix

$$I_{ij} = g_{ij} = \underset{(\mathbf{x}_i, \mathbf{y}_i) \in D}{E} \big[\partial_i \ell_\theta(\mathbf{x}_i, \mathbf{y}_i) \cdot \partial_j \ell_\theta(\mathbf{x}_i, \mathbf{y}_i) \big]. \tag{2.3.5}$$

This equation is a bit easier to grasp, which means that we can finally try to get a bit of intuition about the connections of the concepts mentioned before.
Distance between two points in a curved manifold along a curve c is defined as

$$s = \int_{t_0}^{t_1} \sum_{i,j} \sqrt{g_{ij} \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \frac{\mathrm{d}\theta_j}{\mathrm{d}t}} \mathrm{d}t. \tag{2.3.6}$$

If we only move along one coordinate θ_i on the curve, the distance reduces to

$$s = \int_{t_0}^{t_1} \sqrt{g_{ii}} \left| \frac{\mathrm{d}\theta_i}{\mathrm{d}t} \right| \mathrm{d}t. \tag{2.3.7}$$

Therefore the diagonal components of the FI tell us how far we move across the manifold when we change parameter θ_i. Through inspection of the equation for the metric (Eq. (2.3.5)), it's clear that the distance between points created by changing a parameter relates to the expected change in the logarithm of the probability distribution. We can directly map the parameter space onto the manifold of probabilities, but we need to reconsider the notion of distance for the manifold because distance in the euclidean parameter space doesn't generally translate to distance in the manifold, which is a measure of difference between the probabilities.

Finally let's examine the components of the metric for the neural network. The diagonal components are the expectation values of the squared derivative of the loss. This means that the diagonal values represent how much our subloss changes when we vary the corresponding parameter. The off-diagonal values represent how similar the change in the subloss function is under changes in the two corresponding parameters. This information about the change of the loss regarding the parameters is now of high interest when considering neural network training. In general, calculating the FI computationally is very costly though, because the number of parameters can be very high for complex tasks. That's why the results of this work consider a few observations resulting from the fisher matrix and investigate some possibilities to obtain them from other computationally easier observables of training instead.

The Ricci scalar from Section 2.2.4 is a measure of how much a manifold differs from flat space at a certain point. For neural networks, curvature means that changes in different parameters result in similar changes in the loss function or

that changes in some parameters result in different amounts of change of the loss. We will investigate the curvature of the manifold of the neural network loss during training later **add chapter reference**.

2.4 Investigation of physical phase transitions using Fisher Information

The Fisher Information can also be useful in physical context. It is possible to use it to find phase transitions in thermodynamic systems. We will later look at the possibility of applying this to neural network training in search of processes equivalent to phase transitions, which may lead to more insight into what exactly influences training.

To explain how the FI can be used to find phase transitions let's briefly review [21].

Given an equilibrated physical system in a large thermal heat bath, statistical models of those systems usually deal with Gibbs measures of the form

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_i \theta_i X_i(x)\right). \tag{2.4.1}$$

Here, x represent the microstates of the system, X_i are time-independent functions called "collective variables" and θ_i represent the time-dependent thermodynamic variables. These θ_i could be, for example, temperature, pressure, magnetic fields etc. They will be used as parameters θ of the FI.

Using the thermodynamic variables one can construct thermodynamic potentials. One example used in the paper is the Helmholtz free energy

$$A = -k_B T \ln Z(\theta), \tag{2.4.2}$$

where we consider $k_B T = 1/\beta$ to be one of the thermodynamic variables.
Now it's stated that a classification of phase transitions typically requires an examination of the derivatives of the thermodynamic potential. Specifically, there are cases where an order parameter ϕ^i describing the phase transition is representable as a negative derivative of the potential over some thermodynamic variable θ^i. In this case, the diagonal components Fisher Information defined by the probability distributions in Eq. (2.4.1) and Eq. (2.1.1) can be written as

$$I_{ii} = \beta \frac{\partial \phi^i}{\partial \theta^i}. \tag{2.4.3}$$

There are second order phase transitions, where the order parameter (which is a derivative of the thermodynamic potential) changes continuously while it's derivative diverges. Using this relationship, one can identify those phase transitions by searching for divergences in the diagonal components of the Fisher Information.
In addition to that, [22] visits the same thermodynamic metric described by the Fisher Information. Here the scalar curvature corresponding to the metric $\mathscr{R}$ is introduced as a measure of complexity for the physical systems. It is stated that for all models that they've considered so far, $\mathscr{R}$ diverges at, and only at, the phase transition.
This gives us two ways of finding phase transitions which we can also apply to the Fisher Information of neural network training, to possibly search for analogues of physical phase transitions in the training.

3 | Neural Tangent Kernel

3.1 Derivation from Gradient Flow

3.1.1 What we call time

A simple and intuitive way to introduce the NTK is via "**Gradient Flow**", which is an assumption related to the GD algorithm from Section 1.3.3. To quickly recap the update step from stochastic gradient descent, it is defined as

$$\theta' = \theta - \eta \nabla_\theta \mathscr{L} \left(\{ (f_\theta(\mathbf{x}_i), \mathbf{y}_i) \}_{i=1}^N \right). \tag{3.1.1}$$

The "flow" aspect arises when we start to ignore the discrete nature of these update steps and assume that the θ parameters change continuously. A visual demonstration of how this changes the evolution of the parameters can be seen in Fig. 3.1. Here, the dashed markers represent the evolution under regular GD, while the solid line represents the evolution for gradient flow. To do this, we first introduce a notion of "time" into our system. We try to visualize the optimization process as an evolution of our parameters θ through this variable called time, which converts updating the parameters into moving further along the timeline of our parameters. We translate $\theta \to \theta(t)$ and $\theta' \to \theta(t + \Delta t)$. This time in our system doesn't work exactly the same way as physical time, but since the process of calculating better parameters and changing them is always associated with the expenditure of physical time, it is intuitive to refer to our system's propagation variable as "time".
Returning to the GD algorithm, since the learning rate η affects the size of our update step, we will refer to η as the amount of time it takes to update a parameter

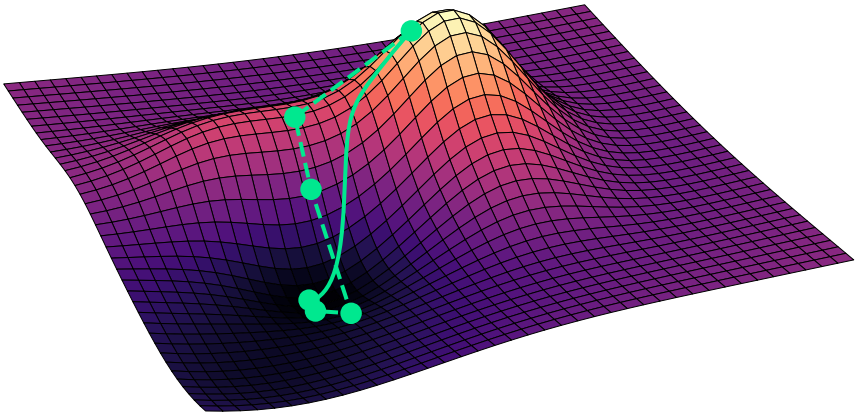


Figure 3.1: This graph shows the effect of assuming gradient flow. The dashed line along with the circle markers show the positions during regular GD with finite step sizes, the solid line shows the path of the parameters under gradient flow.

$\theta' \to \theta(t+\eta)$. The whole GD algorithm then becomes

$$\theta(t+\eta) = \theta(t) - \eta \nabla_{\theta(t)} \mathscr{L}\left(\{(f_{\theta(t)}(\mathbf{x}_i), \mathbf{y}_i)\}_{i=1}^N\right) \tag{3.1.2}$$

which we can rewrite to

$$\frac{\theta(t+\eta) - \theta(t)}{\eta} = -\nabla_{\theta(t)} \mathscr{L}\left(\{(f_{\theta(t)}(\mathbf{x}_i), \mathbf{y}_i)\}_{i=1}^N\right). \tag{3.1.3}$$

When observing the term on the left side, readers who are familiar with calculus might recognize that it looks similar to the definition of a derivative in time

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = \lim_{\eta \to 0} \frac{\theta(t+\eta) - \theta(t)}{\eta}. \tag{3.1.4}$$

This means that for very small learning rates we can approximate the GD as

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\nabla_{\theta(t)} \mathscr{L}\left(\{(f_{\theta(t)}(\mathbf{x}_i), \mathbf{y}_i)\}_{i=1}^N\right) \tag{3.1.5}$$

with the partial derivative of θ along the assumed time variable.

For visual simplicity reasons, let's define the j-th component of the network output for the i-th input point $f_{\theta(t)}(\mathbf{x}_i)_j$ as F_{ij} and assume Einstein summation for the rest of the chapter. This means that when an index is occuring twice, we don't denote a hidden summation over all possible values for this index (for example $a_k b_k = \sum_k a_k b_k$).

Because it will be convenient later, we also spell out one component of the ∇_θ derivation of the loss function further by using the chain rule as

$$\begin{aligned}\frac{\mathrm{d}\theta_k}{\mathrm{d}t} &= -\frac{\mathrm{d}}{\mathrm{d}\theta_k}\mathscr{L}\left(\{(f_{\theta(t)}(\mathbf{x}_i), \mathbf{y}_i)\}_{i=1}^N\right) \\ &= -\frac{\partial \mathscr{L}}{\partial F_{ij}} \cdot \frac{\mathrm{d}F_{ij}}{\mathrm{d}\theta_k}.\end{aligned} \tag{3.1.6}$$

3.1.2 Derivation of the NTK

This notion of time affects not only the parameters, but also everything that depends on them. For example, since the network output of a fixed architecture for a given input data point only depends on the parameters of the network, it can also be mathematically viewed as dependent on the time $f_\theta(\mathbf{x}_i) \to f_{\theta(t)}(\mathbf{x}_i)$. This means we can also calculate the derivative of one of the network outputs $f_{\theta(t)}(\mathbf{x}_i)_j = F_{ij}$ to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F_{ij} &= \frac{\partial F_{ij}}{\partial \theta_k}\frac{\mathrm{d}\theta_k}{\mathrm{d}t} \\ &= \frac{\partial F_{ij}}{\partial \theta_k}\left(-\frac{\partial \mathscr{L}}{\partial F_{ij}}\frac{\mathrm{d}F_{ij}}{\mathrm{d}\theta_k}\right) \\ &= -\underbrace{\frac{\partial F_{ij}}{\partial \theta_k}\frac{\partial F_{lm}}{\partial \theta_k}}_{=:\,\Lambda_{iljm}}\frac{\partial \mathscr{L}}{\partial F_{lm}}.\end{aligned} \tag{3.1.7}$$

In the last line, we used the fact that F_{ij} only depend explicitly on θ, which means that the total and partial derivative ar interchangeable. The rank 4 hypermatrix Λ is what we call the Neural Tangent Kernel for GD. We sorted the indices of this matrix so that the first two refer to the input points of f and the last two refer to the components of the output dimensions of the neural network. Note that this NTK is derived directly from the update algorithm of the GD for infinitely small η, the equations above don't hold for other optimization systems.
Another way to derive the NTK using an approximation of $\Delta\mathscr{L}$ for small η can be seen around page 196 of [14]. The NTK derived there also works for tensorial gradient descent with a learning rate tensor η_{ij}.

3.2 Interpretation

Through Eq. (3.1.7) the NTK is defined as

$$\Lambda_{ij\alpha\beta} = \nabla_\theta f_\theta(\mathbf{x}_i)_\alpha \cdot \nabla_\theta f_\theta(\mathbf{x}_j)_\beta. \tag{3.2.1}$$

For now, let's assume that the neural network has only one output $f_\theta(\mathbf{x}_i)_\alpha \to f_\theta(\mathbf{x}_i)$, which gets rid of the α and β indices for us and makes the NTK a regular matrix. We can use this matrix to calculate the time derivative of the neural network output similar to Eq. (3.1.7) by

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{\theta(t)}(\mathbf{x}_i) = \sum_j \Lambda_{ij} \left(-\frac{\partial \mathscr{L}}{\partial f_{\theta(t)}(\mathbf{x}_j)} \right). \tag{3.2.2}$$

This means that the evolution of the network output for input $\mathbf{x}_i$ is influenced by the outputs for other input values $\mathbf{x}_j$ through the NTK. We can investigate this further by taking a look at the definition of the NTK above in Eq. (3.2.1).

A mathematical kernel K is defined as a function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i), \tag{3.2.3}$$

with another so-called "feature map" function ϕ that maps the points $\mathbf{x}_i$ into a higher dimensional inner product space called the "feature space". The kernel K assigns a scalar value to the two points by comparing their "features" via a scalar product in the feature space. For more specific information on Kernels and how they are used, please refer to [23].
In our case the feature map is ∇_θ which maps our scalar network output onto a vector that contains the derivatives of the network output with respect to all of the various parameters. This is our feature space, because what we're interested in is how moving in θ space changes the output values of our network. If we compare those two mappings we get a value that measures how closely the direction that increases $f_\theta(\mathbf{x}_i)$ most effectively matches with the direction that increases $f_\theta(\mathbf{x}_j)$ most effectively (which are the directions the gradients point to). For example, $\Lambda_{ij} = 0$ means that varying θ in the direction that changes $f_\theta(\mathbf{x}_i)$ most effectively results in 0 change for $f_\theta(\mathbf{x}_j)$.

Coming back to Eq. (3.2.2) the right side is the negative loss derivative with respect to $f_\theta(\mathbf{x}_j)$. Since we update the parameters in a way that minimizes the loss most effectively in gradient flow, the negative loss derivative with respect to $f_\theta(\mathbf{x}_j)$ describes how beneficial increasing $f_\theta(\mathbf{x}_j)$ is for our system. If we now multiply this with the NTK, which is a kernel that describes how similar $f_\theta(\mathbf{x}_i)$ and $f_\theta(\mathbf{x}_j)$ behave when changing theta, and sum everything up, we get the change of the function value $f_\theta(\mathbf{x}_i)$ as result.
In θ space, we can directly relate the evolution of θ to its negative loss derivative,

because the parameters themselves are what we update with our algorithm. For the derivative of the output values we need the NTK as well, because we don't change the output values directly. If the loss derivative with respect to an output value tells the system that it *should increase this output value*, the system *changes the parameters* in a way that increases this output value. The difference to the derivative of θ is that the change of parameters now doesn't reflect in just one output, but in every other output value as well. That's where the NTK arises in the equation. The NTK is, in a way, a translation of the GD into the space of outputs of the neural network.

All of the explanations above apply to the 4 dimensional hypermatrix form of the NTK for multidimensional neural network output as well, which can be seen when comparing Eq. (3.2.2) to Eq. (3.1.7).

4 | Results

4.1 Fisher Trace - NTK relation

For a neural network with N parameters, the Fisher Information is a matrix with N^2 entries. To illustrate how large those resulting matrices are, let's consider a network containing 3 hidden layers with a respective width of 128 neurons that gets trained on the 28×28 matrices of the MNIST dataset. This network has over 130000 parameters, which results in over 10^{10} entries in the FI. When saving the entries as 32-bit floating point numbers, the matrix would take over 67GB of space. Calculating this matrix in under 1 hour would require over 2 million entries to be calculated per second. Calculating the full matrix to optimize training is therefore computationally inefficient. That's why we're going to look at an equation for calculating the trace of the Fisher Information, or Fisher Trace for short, through other mathematical objects.

Let's first inspect the Fisher Trace by using the chain rule of derivation as

$$\begin{aligned}
\mathrm{tr}(I) &= \sum_{\alpha=1}^{N_P} I_{\alpha\alpha} \\
&= \sum_{\alpha=1}^{N_P} \left\{ \underset{(\mathbf{x}_i,\mathbf{y}_i)\in D}{E} \left[\frac{\mathrm{d}}{\mathrm{d}\theta_\alpha} \ell(f_\theta(\mathbf{x}_i),\mathbf{y}_i) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta_\alpha} \ell(f_\theta(\mathbf{x}_i),\mathbf{y}_i) \right] \right\} \\
&= \sum_{\alpha=1}^{N_P} \left\{ \frac{1}{N} \sum_{i=1}^{N_D} \left[\sum_{a=1}^{N_O} \left(\frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_a} \frac{\mathrm{d} f_\theta(\mathbf{x}_i)_a}{\mathrm{d}\theta_\alpha} \right) \cdot \sum_{b=1}^{N_O} \left(\frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_b} \frac{\mathrm{d} f_\theta(\mathbf{x}_i)_b}{\mathrm{d}\theta_\alpha} \right) \right] \right\} \\
&= \frac{1}{N} \sum_{i=1}^{N_D} \sum_{a=1}^{N_O} \sum_{b=1}^{N_O} \left[\frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_a} \frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_b} \cdot \underbrace{\sum_{\alpha=1}^{N_P} \left(\frac{\mathrm{d} f_\theta(\mathbf{x}_i)_a}{\mathrm{d}\theta_\alpha} \frac{\mathrm{d} f_\theta(\mathbf{x}_i)_b}{\mathrm{d}\theta_\alpha} \right)}_{\Lambda_{iiab}} \right],
\end{aligned} \tag{4.1.1}$$

with the amount of parameters N_P, the amount of dataset pairs N_D and the output dimension of the network N_O. By identifying the entries of the NTK in the last line and simplifying the notation we can rewrite the relation as

$$\mathrm{tr}(I) = \underset{(\mathbf{x}_i,\mathbf{y}_i)\in D}{E}\left[\sum_{a,b}\left(\frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_a}\frac{\partial \ell}{\partial f_\theta(\mathbf{x}_i)_b}\cdot \Lambda_{iiab}\right)\right]. \tag{4.1.2}$$

Here it is important to note that there are two main parts on the right hand side: One is the loss derivative and the other is the NTK. Going further we will conduct some experiments to investigate which of those two terms is the dominant driving force in the evolution of the Fisher Trace. Specifically, we will look at the time evolution of the Fisher Trace in comparison the trace of the NTK to see how similar they evolve during training.

4.1.1 Comparison of traces of NTK and Fisher Information

To test how similar the evolution of the traces of NTK and Fisher Information are, we trained multiple networks on the MNIST dataset explained in Section 1.3.1 and recorded the traces of the FI and NTK. We used networks consisting of 2 hidden layers that were equipped with the ReLU activation function. The network widths varied between 128 and 784. The losses used were the L_p-Norm loss [24]

$$d_p(f_\theta(\mathbf{x}), \mathbf{y}) = \sqrt[p]{\sum_{i=1}^{N_O} |f_\theta(\mathbf{x}_i) - \mathbf{y}_i|^p}, \tag{4.1.3}$$

the mean-power loss of order n

$$d_n(f_\theta(\mathbf{x}), \mathbf{y}) = \frac{1}{N_O}\sum_{i=1}^{N_O}(f_\theta(\mathbf{x}_i) - \mathbf{y}_i)^n, \tag{4.1.4}$$

and the softmax-ross-entropy loss [12]

$$d(f_\theta(\mathbf{x}), \mathbf{y}) = \sum_{i=1}^{N_O} \mathbf{y}_i \log(\sigma_i(f_\theta(\mathbf{x}))), \tag{4.1.5}$$

where we used the definition for distance d from Eq. (1.3.4), and a softmax function σ_i from Eq. (1.2.2). The experiment was conducted for orders of n or p equal to 2, 4 and 6. We trained each combination of architecture and loss function for a total of 5 times and calculated mean values and standard deviations of the traces using the results. The networks were trained using the Adam optimizer [16].
As an illustrative example, the results for L_2-norm loss, Mean power loss of order $n = 2$, as well as the result for Cross Entropy loss for a network with a width of 128 neurons are depicted in Fig. 4.1. The solid line represents the mean of the 5 experiments, the translucent area marks one standard deviation above and under the mean value.

It is visible that for the L_2 norm loss of order, the traces evolve almost proportionally to each other, but for the other loss functions they evolve significantly different. This indicates that the loss derivatives in Eq. (4.1.2) play major roles and can't be disregarded. The trace of the NTK doesn't seem to be a good approximation of the Fisher Trace in general. For which cases the NTK can be used as a good approximation for the Fisher Trace needs to be investigated further and can't be answered from the data generated here. More results from the experiment for different losses and also a larger network width can be found in Appendix B. From there it seems like the deviations between the two traces increase for larger network size.

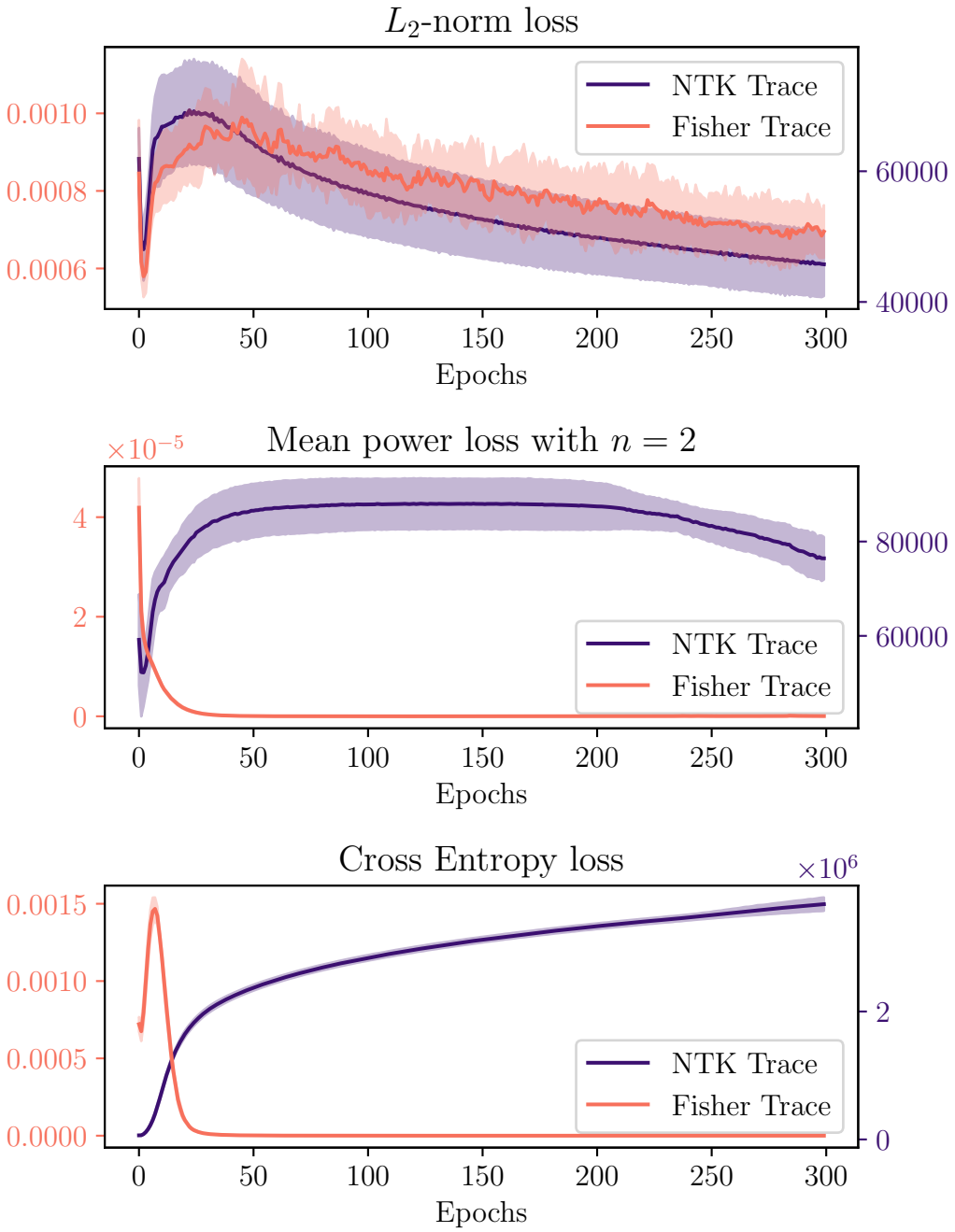


Figure 4.1: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 128 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.

Literature

[1] Esther Inglis-Arkell. Mar. 7, 2012. URL: https://gizmodo.com/the-very-first-robot-brains-were-made-of-old-alarm-cl-5890771 (visited on 08/23/2023).

[2] A. M. TURING. "I.COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236 (Oct. 1950), pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. URL: https://doi.org/10.1093/mind/LIX.236.433.

[3] URL: https://www.oxfordlearnersdictionaries.com/us/definition/english/machine-learning (visited on 08/23/2023).

[4] SS Khare and AR Gajbhiye. "Literature Review on Application of Artificial Neural Network (Ann) In Operation of Reservoirs". In: *International Journal of computational Engineering research (IJCER) IJCER| June 2013| VOL 3 ISSUE 6* (1943), p. 63.

[5] Hermann Haken. *Principles of brain functioning: a synergetic approach to brain activity, behavior and cognition*. Vol. 67. Springer Science & Business Media, 2013.

[6] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–133. DOI: https://doi.org/10.1007/BF02478259.

[7] Iqbal H Sarker. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions". In: *SN Computer Science* 2.6 (2021), p. 420.

[8] Bolin Gao and Lacra Pavel. “On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning”. In: (2018). arXiv: 1704.00805 [math.OC].

[9] Vladimir Nasteski. “An overview of the supervised machine learning methods”. In: *Horizons. b* 4 (2017), pp. 51–62.

[10] L. Bottou et al. “Comparison of classifier methods: a case study in handwritten digit recognition”. In: *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*. Vol. 2. 1994, 77–82 vol.2. DOI: 10.1109/ICPR.1994.576879.

[11] S. Amari and S.C. Douglas. “Why natural gradient?” In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP ’98 (Cat. No.98CH36181)*. Vol. 2. 1998, 1213–1216 vol.2. DOI: 10.1109/ICASSP.1998.675489.

[12] Qi Wang et al. “A Comprehensive Survey of Loss Functions in Machine Learning”. In: *Annals of Data Science* 9 (2 2022), pp. 2198–5812. DOI: 10.1007/s40745-020-00253-5.

[13] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[14] Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory*. Cambridge University Press, May 2022. DOI: 10.1017/9781009023405. URL: https://doi.org/10.1017%2F9781009023405.

[15] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. “An alternative view: When does SGD escape local minima?” In: *International conference on machine learning*. PMLR. 2018, pp. 2698–2707.

[16] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: (2017). arXiv: 1412.6980 [cs.LG].

[17] Alexander Ly et al. “A Tutorial on Fisher information”. In: *Journal of Mathematical Psychology* 80 (2017), pp. 40–55. ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2017.05.006. URL: https://www.sciencedirect.com/science/article/pii/S0022249617301396.

[18] *atemateca (IME/USP)/Rodrigo Tetsuo Argenton*. URL: https://commons.wikimedia.org/wiki/File:Galton_box.jpg#/media/Datei:Galton_box.jpg (visited on 08/24/2023).

[19] Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*. 1st ed. ISBN: 978-0-387-96056-2. DOI: https://doi.org/10.1007/978-1-4612-5056-2.

[20] Bernard Schutz. *A First Course in General Relativity*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511984181.

[21] Mikhail Prokopenko et al. “Relating Fisher information to order parameters”. In: *Phys. Rev. E* 84 (4 Oct. 2011), p. 041116. DOI: 10.1103/PhysRevE.84.041116. URL: https://link.aps.org/doi/10.1103/PhysRevE.84.041116.

[22] W Janke, DA Johnston, and Ralph Kenna. “Information geometry and phase transitions”. In: *Physica A: Statistical Mechanics and its Applications* 336.1-2 (2004), pp. 181–186.

[23] Colin Campbell. “An introduction to kernel methods”. In: *Studies in Fuzziness and Soft Computing* 66 (2001), pp. 155–192.

[24] Tien D Bui et al. “LP norm relaxation approach for large scale data analysis: a review”. In: *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*. Springer. 2018, pp. 285–292.

[25] Nan Du et al. “Glam: Efficient scaling of language models with mixture-of-experts”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5547–5569.

[26] Piyush Kaul and Brejesh Lall. “Riemannian Curvature of Deep Neural Networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.4 (2020), pp. 1410–1416. DOI: 10.1109/TNNLS.2019.2919705.

[27] Daniel Jannai et al. *Human or Not? A Gamified Approach to the Turing Test*. 2023. arXiv: 2305.20010 [cs.AI].

[28] S. Amari and S.C. Douglas. “Why natural gradient?” In: 2 (1998), 1213–1216 vol.2. DOI: 10.1109/ICASSP.1998.675489.

[29] Enzo Grossi and Massimo Buscema. “Introduction to artificial neural networks”. In: *European Journal of Gastroenterology and Hepatology* 19.12 (Dec. 2007), pp. 1046–1054. DOI: 10.1097/MEG.0b013e3282f198a0.

A | Some mathematical proofs

A.1 Proof of Eq. (2.1.3)

For all variables X^n that satisfy

$$f(x^n|\theta) = \prod_{\alpha=1}^{n} f(x_\alpha|\theta) \tag{A.1.1}$$

we can prove

$$I_{X^n,ij} = \sum_{\alpha} I_{X_\alpha,ij} \tag{A.1.2}$$

by

$$
\begin{aligned}
I_{X^n,ij} &= \underset{x^n \in X^n}{E} \left\{ \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x^n|\theta) \frac{\mathrm{d}}{\mathrm{d}\theta_j} \log f(x^n|\theta) \right\} \\
&= \sum_{x^n \in X^n} \left\{ \left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x^n|\theta) \right] \left[\frac{\mathrm{d}}{\mathrm{d}\theta_j} f(x^n|\theta) \right] [f(x^n|\theta)]^{-1} \right\} \\
&\overset{A.1.1}{=} \sum_{x^n \in X^n} \left\{ \frac{\mathrm{d}}{\mathrm{d}\theta_i} \left[\prod_\alpha f(x_\alpha|\theta) \right] \frac{\mathrm{d}}{\mathrm{d}\theta_j} \left[\prod_\beta f(x_\beta|\theta) \right] \left[\prod_\gamma f(x_\gamma|\theta) \right]^{-1} \right\} \\
&= \sum_{x^n \in X^n} \left\{ \left(\prod_\alpha f(x_\alpha|\theta) \right) \left(\sum_\alpha \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right) \right. \\
&\qquad \left. \left(\prod_\beta f(x_\beta|\theta) \right) \left(\sum_\beta \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right) \left[\prod_\gamma f(x_\gamma|\theta) \right]^{-1} \right\} \\
&= \underset{x^n \in X^n}{E} \left\{ \left[\sum_\alpha \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right] \left[\sum_\beta \frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\beta|\theta) \right] \right\} \\
&\overset{(*)}{=} \underset{x^n \in X^n}{E} \left\{ \sum_\alpha \left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right] \left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right] \right\} \\
&= \sum_\alpha \underset{x^n \in X^n}{E} \left\{ \left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right] \left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} \log f(x_\alpha|\theta) \right] \right\} \\
&= \sum_\alpha I_{X_\alpha,ij}.
\end{aligned}
\tag{A.1.3}
$$

The equality at $(*)$ holds because for all $\alpha \neq \beta$

$$\begin{aligned}
&\underset{x^n \in X^n}{E}\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log f(x_\alpha|\theta)\right]\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log f(x_\beta|\theta)\right]\right\} \\
&\propto \sum_{x_\alpha}\sum_{x_\beta}\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\alpha|\theta)\right]\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\beta|\theta)\right]\right\} \\
&= \sum_{x_\alpha}\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\alpha|\theta)\right]\sum_{x_\beta}\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\beta|\theta)\right]\right\} \\
&= \sum_{x_\alpha}\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\alpha|\theta)\right]\sum_{x_\beta}\frac{\mathrm{d}}{\mathrm{d}\theta_i}\left[f(x_\beta|\theta)\right]\right\} \\
&= \sum_{x_\alpha}\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i} f(x_\alpha|\theta)\right]\sum_{x_\beta}\frac{\mathrm{d}}{\mathrm{d}\theta_i}\left[1\right]\right\} \\
&= 0.
\end{aligned} \tag{A.1.4}$$

This prove still holds when using continuous variables instead of discrete ones.

A.2 Proof of Eq. (2.2.15)

Here we will prove Eq. (2.2.15) which states

$$\underset{x\in X}{E}\Big[\partial_i \log p(x|\theta)\cdot\partial_j \log p(x|\theta)\Big] = -\underset{x\in X}{E}\Big[\partial_i\partial_j \log p(x|\theta)\Big]. \tag{A.2.1}$$

To prove this we will evaluate the right side by

$$\begin{aligned}
-\underset{x\in X}{E}\Big[\partial_i\partial_j \log p(x|\theta)\Big] &= -\underset{x\in X}{E}\left\{\partial_i\left[\partial_j \log p(x|\theta)\right]\right\} \\
&= -\underset{x\in X}{E}\left\{\partial_i\left[\frac{\partial_j p(x|\theta)}{p(x|\theta)}\right]\right\} \\
&= -\underset{x\in X}{E}\left[\frac{\partial_i\partial_j p(x|\theta)}{p(x|\theta)} - \frac{\partial_i p(x|\theta)\partial_j p(x|\theta)}{p(x|\theta)^2}\right] \\
&= -\underset{x\in X}{E}\left[\frac{\partial_i\partial_j p(x|\theta)}{p(x|\theta)} - \partial_i \log p(x|\theta)\partial_j \log p(x|\theta)\right] \\
&= \underset{x\in X}{E}\Big[\partial_i \log p(x|\theta)\cdot\partial_j \log p(x|\theta)\Big],
\end{aligned} \tag{A.2.2}$$

Where the last equation holds because

$$\begin{aligned}
-\underset{x\in X}{E}\left[\frac{\partial_i\partial_j p(x|\theta)}{p(x|\theta)}\right] &= \sum_{x\in X}\left[\partial_i\partial_j p(x|\theta)\right] \\
&= \partial_i\partial_j \sum_{x\in X}\left[p(x|\theta)\right] \\
&= -\partial_i\partial_j \underset{x\in X}{E}\left[1\right] \\
&= 0.
\end{aligned} \tag{A.2.3}$$

Here we assumed that we can swap the sum and the derivatives. For continuous variables, the swapping of the integral and the derivatives has to be assumed for the equation to hold.

B | MNIST experiment for trace comparison

This section contains more results of the experiment from Section 4.1.1. The results for loss functions of order 4 and 6 for the network of width 128 are depicted in Fig. B.1 and Fig. B.2. The results for loss functions of order 2, 4 and 6 for the network of width 784 are depicted in Fig. B.3, Fig. B.4 and Fig. B.5. The cross entropy loss doesn't vary with order but is still depicted for reference in every plot.

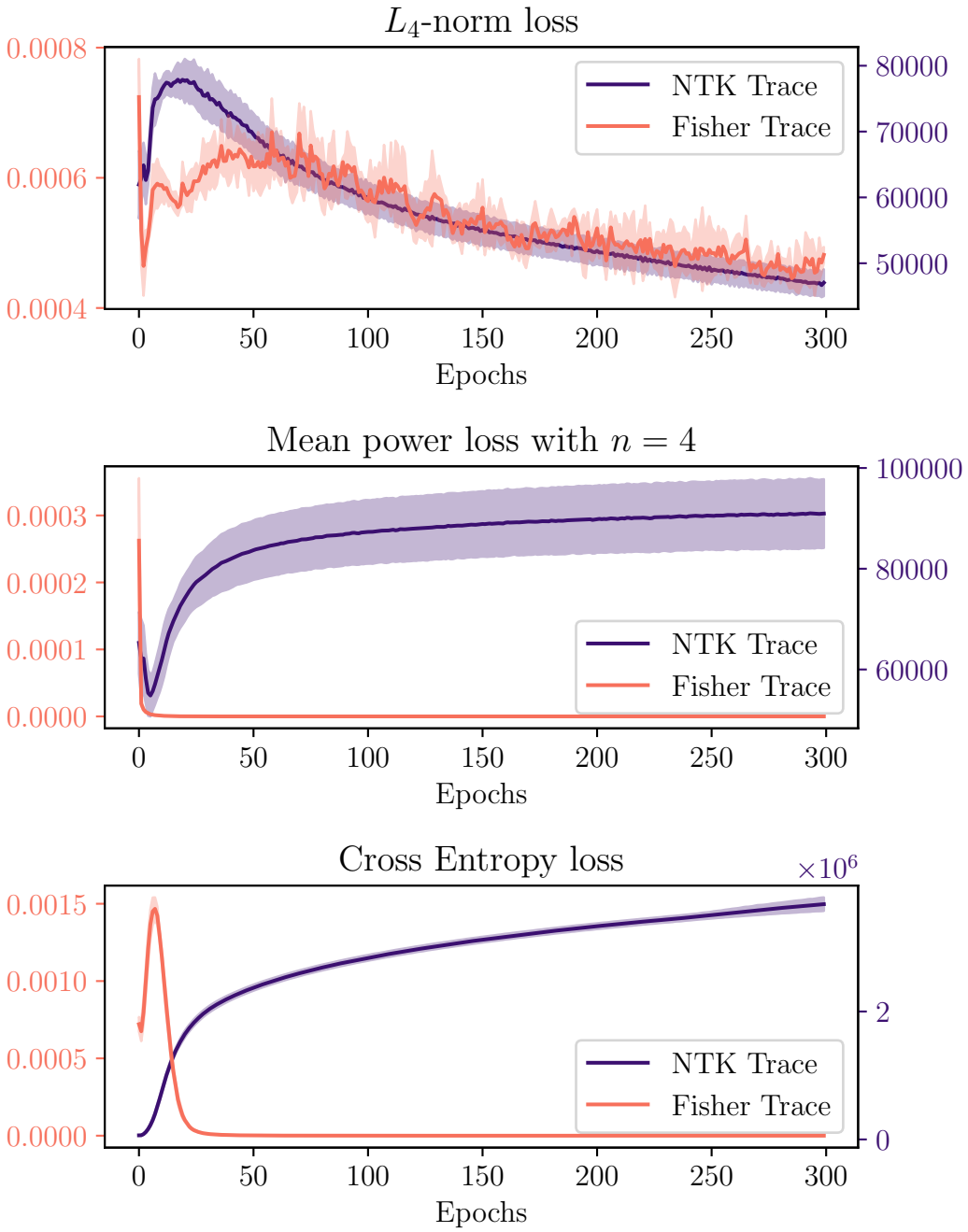


Figure B.1: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 128 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.

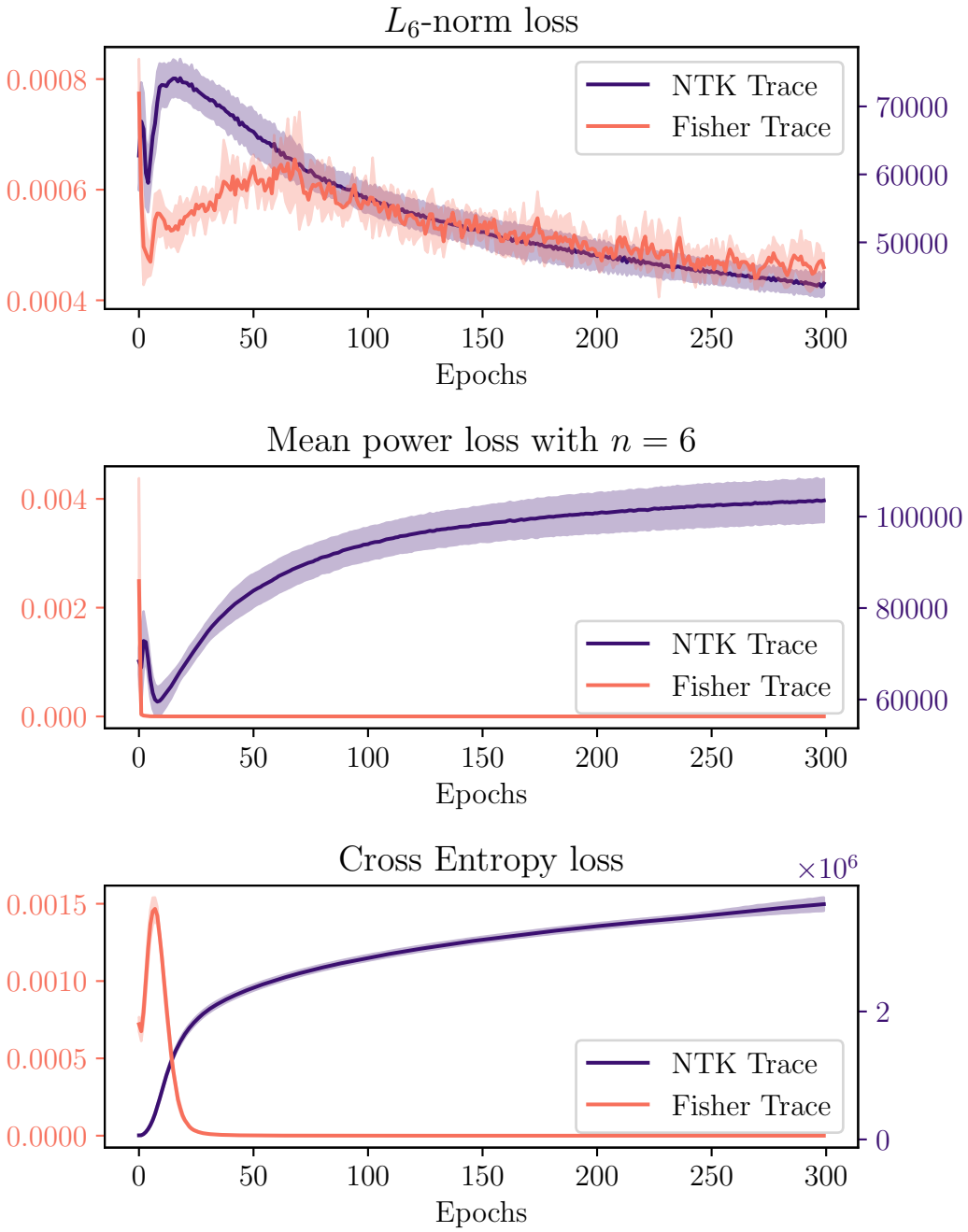


Figure B.2: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 128 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.

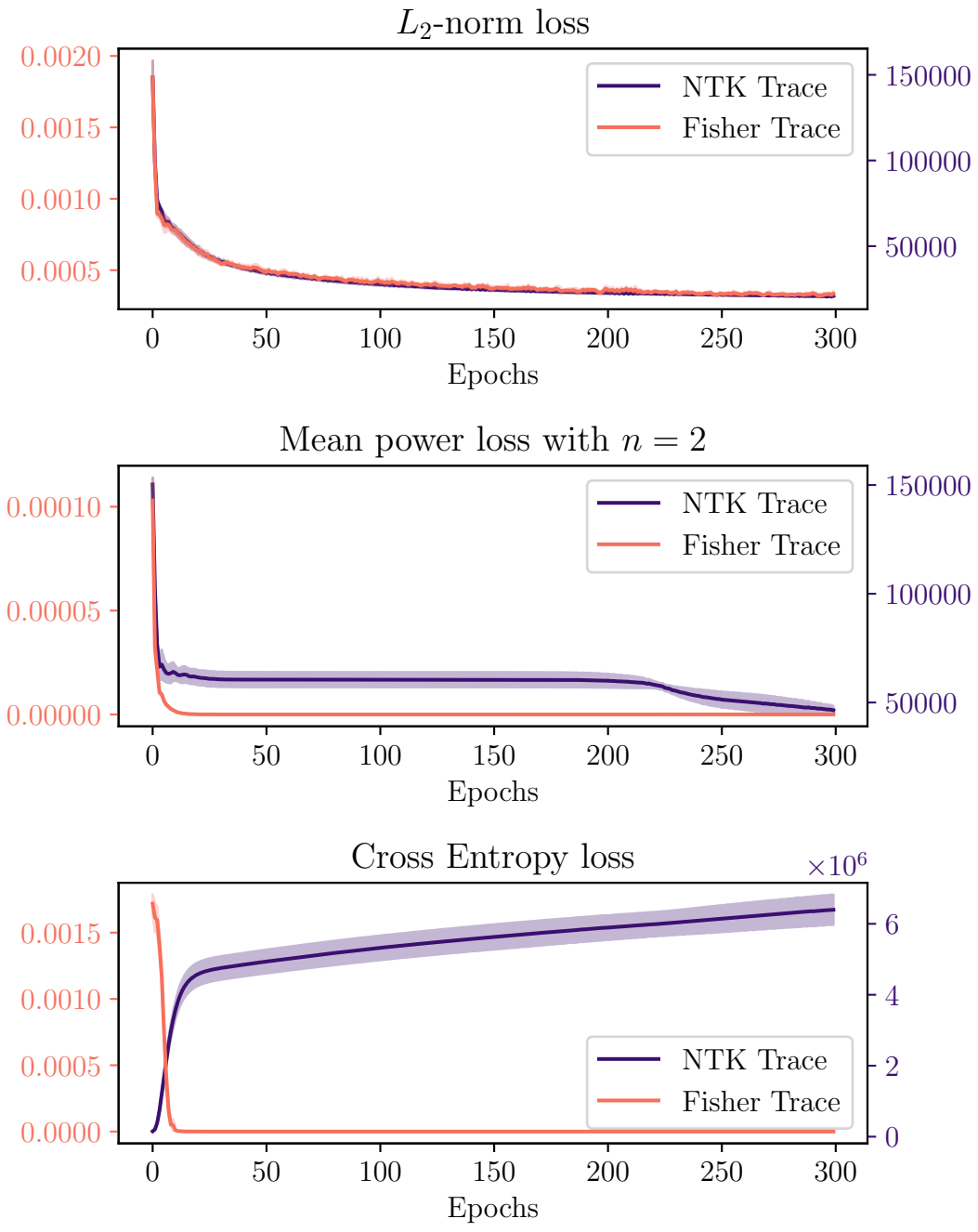


Figure B.3: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 784 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.

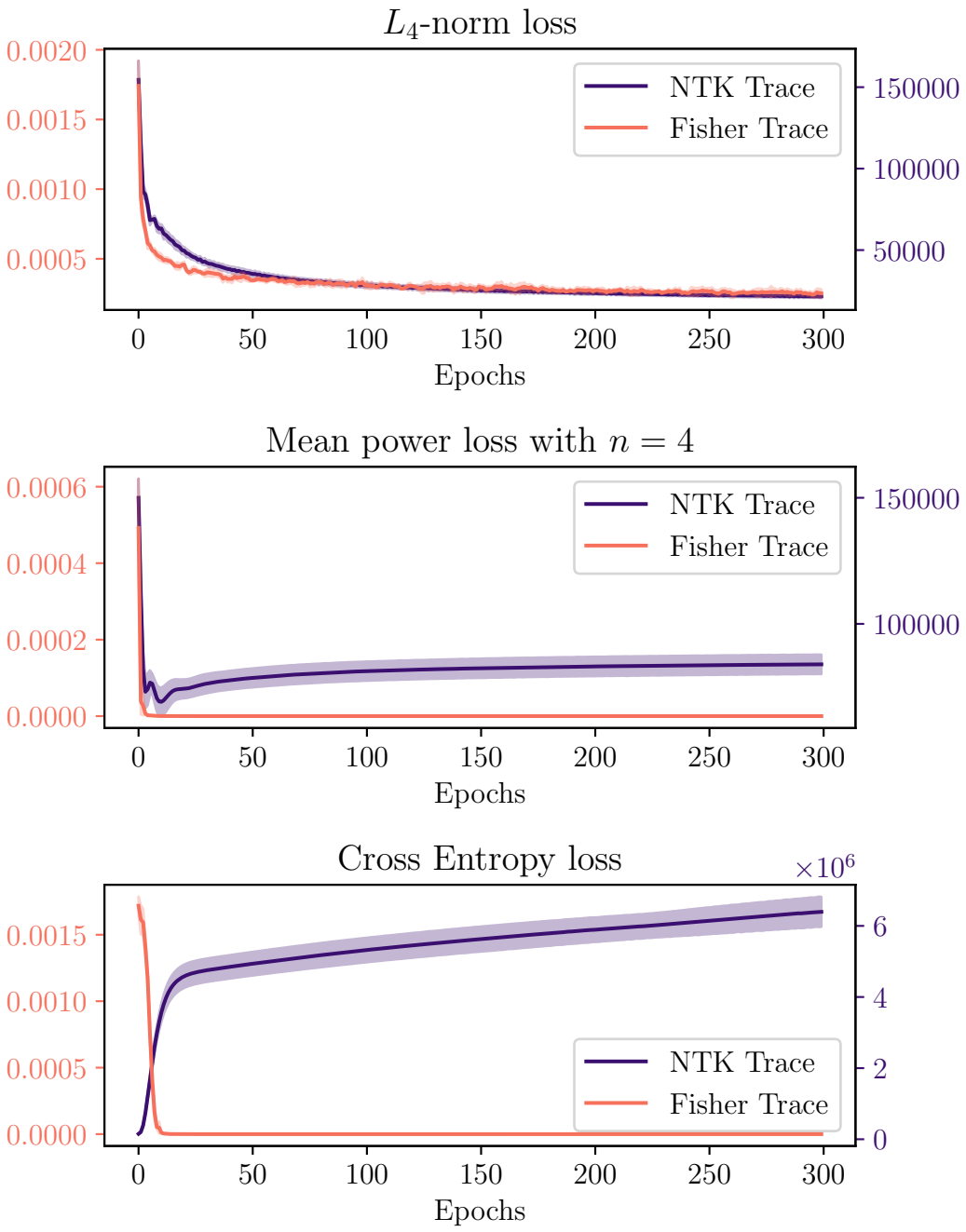


Figure B.4: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 784 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.

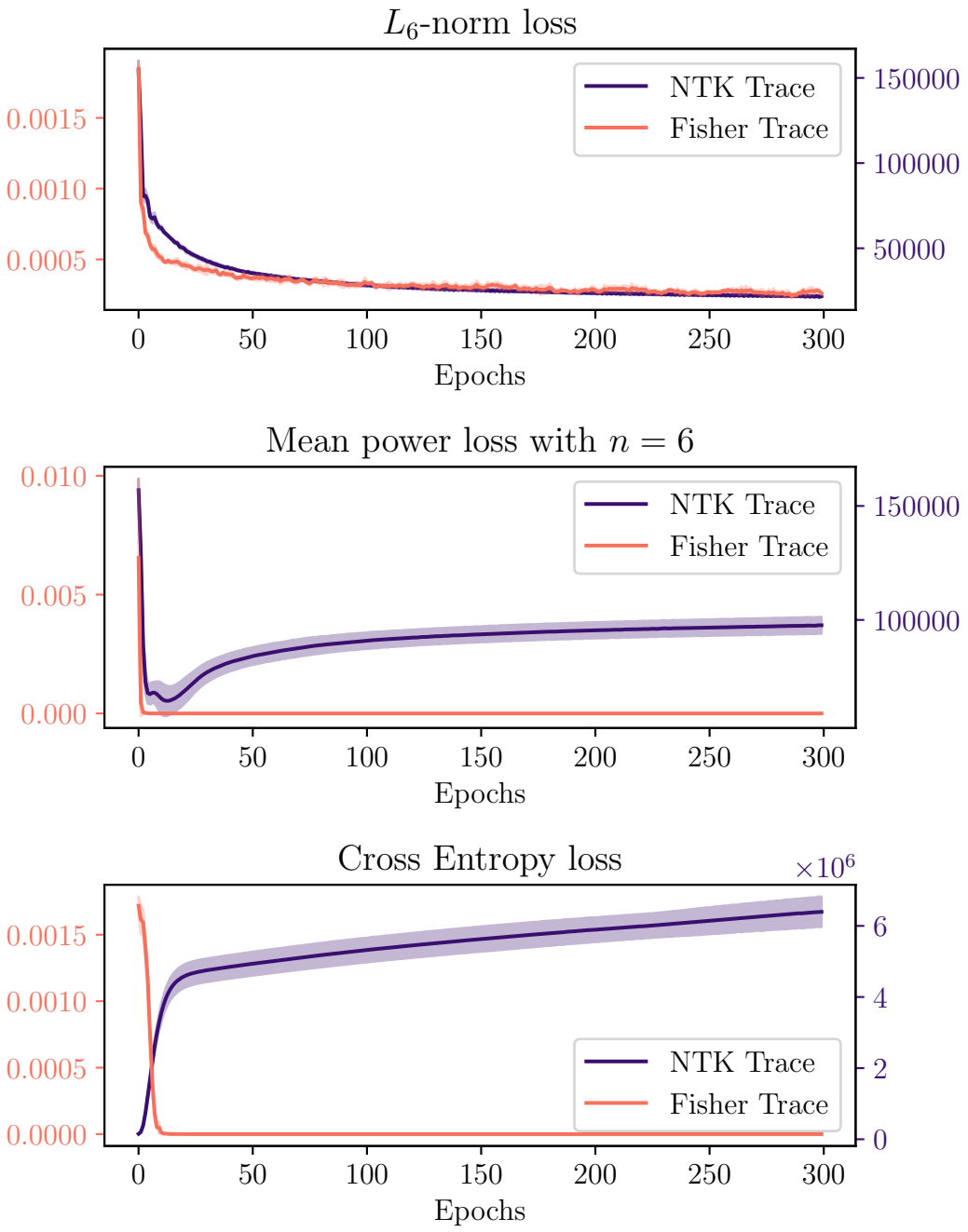


Figure B.5: Trace of NTK and Fisher information during 300 epochs of training on the MNIST dataset using the Adam optimizer. The network consisted of 2 hidden layers with a *width of 784 neurons* that were equipped with the ReLU activation function. The loss functions used for optimization are denoted in each subplot. The solid line represents the mean value of 5 experiments, the translucent area represents one standard deviation from the mean value.