An Information Theoretic View on Learning of Artificial Neural Networks

Emilio Rafael Balda and Arash Behboodi and Rudolf Mathar
Institute for Theoretical Information Technology
RWTH Aachen University, D-52056 Aachen, Germany
{balda, behboodi, mathar}@ti.rwth-aachen.de

***Abstract*—Deep learning based on Artificial Neural Networks (ANNs) has achieved great successes over the last years. However, gaining insight into the fundamentals and explaining their functionality is an open research area of high interest. In this paper, we use an information theoretic approach to reveal typical learning patterns of ANNs. For this purpose the training samples, the true labels, and the estimated labels are considered as random variables. Then, the mutual information and conditional entropy between these variables are studied. We show that the learning process of ANNs consists of essentially two phases. First, the network learns mostly about the input samples without significant improvement in the accuracy, thereafter the correct class allocation becomes more pronounced. This is based on investigating the conditional entropy of the estimated class label given the true one in the course of training. We next derive bounds on the conditional entropy as a function of the error probability, which provide interesting insights into the learning behavior of ANNs. Theoretical investigations are accompanied by extensive numerical studies on an artificial data set as well as the MNIST and CIFAR benchmark data using the widely known networks LeNet-5 and DenseNet. Amazingly, in all cases the bounds are nearly attained in later stages of the training phase, which allows for an analytical measure of the training status of an ANN.**

***Index Terms*—Neural networks, Fano's inequality, machine learning**

I. Introduction

Multi-layer ANNs and deep learning algorithms have achieved amazing success in a variety of tasks, which were previously deemed to be notoriously difficult. To name a few among many, we mention large scale pattern recognition [1], speech analysis [2] and reinforcement learning as in the AlphaGo challenge [3]. Ignited by their success in practical tasks, many attempts have been made during recent years to develop a satisfactory theory for learning of ANNs and in particular for the effectiveness of gradient-based back-propagation training.

Among many existing works, the information bottleneck approach of [4] sparked researchers' interest in information theoretic quantities to explain learning of ANNs. Based on the information bottleneck method [5], multi-layer ANNs were considered as a Markov chain of random variables representing the input, the output and the hidden layers. The mutual information between the hidden layers, the input and the output are studied during training and it is shown that they reside close to the information bottleneck curve. Additionally it is observed that a compression phase happens at later stages of training for deep neural networks with a sigmoid activation function. This work gave rise to an ongoing discussion about the claims of the paper [6], [7]. Despite these discussions, it seems that the asymptotic behavior of mutual information for the hidden layers is correctly described by an information theoretic method.

Applying information theoretic methods for learning extends to recent works where the generalization error of learning algorithms is studied. The framework was initially introduced in [8] to deal with the selection bias of learning algorithms. In that work, it is shown that the generalization error can be upper bounded by the mutual information between the training dataset and the output of the learning algorithm. The framework was extended later in [9], and further in [10] to provide a more tight upper bound using generic chaining techniques.

In the present paper, we adopt an information theoretic view to understand the learning process in ANNs. The main metrics to assess the learning progress will be information theoretic quantities like mutual information and conditional entropy. In particular, we investigate how much information about the training samples and true labels is contained in the output of the ANN during the training process. One interesting conclusion can be drawn from observing the information content of the output labels. The learning of ANNs appears to progress in two distinct phases. In the early stage of training, the output labels obtain mostly information about the training samples and they learn about the true labels mainly in later stages of training. This finding suggests that the learning is divided into a predominantly non-discriminative stage where the output learns mostly about the input samples, followed by a predominantly discriminative stage in which the true labels are learned.

Moreover, from Fano's inequality we derive an upper bound on the conditional entropy of the estimated labels given the true ones in terms of the error probability. If the learning process is set up properly, then we observe experimentally that the conditional entropy approaches the upper bound in later stages of the learning process.

The paper is organized as follows. In Section II, the information theoretic framework for learning is introduced. The

This work was partially supported by the DFG grant SCoSNeL-MA 1184/36-1 in the priority program CoSIP.

relation between the expected error of a learning algorithm and conditional entropy is discussed in Section III. Extensive numerical experiments are conducted in Section IV. Besides a Fully Connected Neural Network (FCNN) on an artificial dataset, we apply widely known ANNs, namely LeNet-5 and DenseNet to the benchmark data MNIST and CIFAR, in order to verify our hypotheses on learning behavior.

A. Notation

Vectors are denoted by bold characters $\mathbf{x}, \mathbf{y}$ and matrices by capitals $\mathbf{A}, \mathbf{B}$. Random variables and vectors are denoted by X, Y and bold symbols $\mathbf{X}, \mathbf{Y}$ respectively.

II. Information Theoretic Learning

Before setting up the formulation of the learning problem, we introduce some well known concepts from information theory. The central notion of information theory is entropy. For a discrete random variable[1] X supported by some countable set $\mathcal{X}$, the Shannon entropy of X is denoted by $H(X)$ and it is given by

$$H(X) = -\sum_{x\in\mathcal{X}} \mathbb{P}(x)\log\mathbb{P}(x) = \mathbb{E}(\log\frac{1}{\mathbb{P}(X)}),$$

where $\mathbb{P}(x)$ is the probability mass function. One can equally define the conditional entropy $H(X|Y)$ for two discrete random variables X and Y as

$$H(X|Y) = \mathbb{E}(\log\frac{1}{\mathbb{P}(X|Y)}).$$

The mutual information between X and Y is given by

$$I(X;Y) = H(X) - H(X|Y).$$

The operational meaning of these notions will be interpreted in the next section.

A. A Model for Learning of Artificial Neural Networks

Once the number of layers, neurons and the choice of nonlinearities are fixed, ANNs can be characterized by the set of weights $(\mathbf{W}_1, \ldots, \mathbf{W}_L)$ and biases $(\mathbf{b}_1, \ldots, \mathbf{b}_L)$. At layer l, the input-output relation is given by

$$\mathbf{T}_l = \sigma_l(\mathbf{W}_l\mathbf{T}_{l-1} + \mathbf{b}_l),$$

where $\sigma_l(\cdot)$ is the so called activation function at layer l. The output y is generated by feeding forward input $\mathbf{x} \in \mathbb{R}^p$ through the layers. This relation is simply denoted by $y = g_{\boldsymbol{\theta}}(\mathbf{x})$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ including all design parameters, i.e., weights and biases. We consider classification networks in this work with input given as a vector $\mathbf{x} \in \mathbb{R}^p$ and output label $y \in \{0, \ldots, K-1\}$.

We assume that future input to an ANN $g_{\boldsymbol{\theta}}(\cdot)$ is modeled by a random variable $\mathbf{X} \in \mathbb{R}^p$ with corresponding class label $Y \in \{0, \ldots, K-1\}$. Y may be function of $\mathbf{X}$ as $Y = g(\mathbf{X})$ or may be subject to additional random effects. The distribution of $\mathbf{X}$ is normally unknown and because of the large state space barely estimable. For a fixed parameter $\boldsymbol{\theta}$, the decision of an ANN about the class label of the input $\mathbf{X}$ is given by $\hat{Y} = g_{\boldsymbol{\theta}}(\mathbf{X})$, a random variable itself. Of course it may happen that $\hat{Y} \neq Y$.

[1]Throughout the paper, random variables are assumed to be discrete.

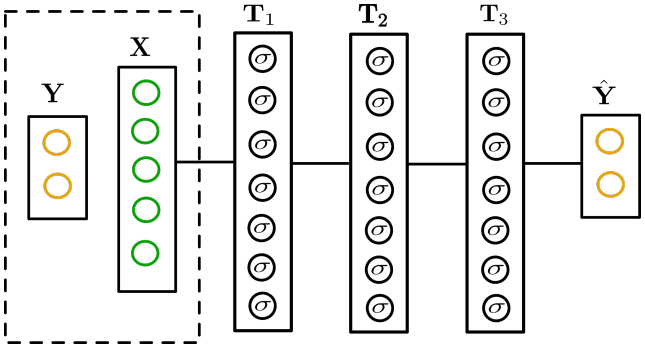

Fig. 1. Multi-layer neural networks with the correct label Y and the approximate label $\hat{Y}$

In the training phase of a classification network a large sample of n independent observations

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

is drawn from $(\mathbf{X}, Y)$ as a realization of i.i.d. random variables $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ with the same distribution as $(\mathbf{X}, Y)$.

Within the framework of statistical learning theory [11], [12], the pairs $(\mathbf{x}, y)$ belong to the instance space $\mathcal{Z} = \mathbb{R}^p \times \{0, \ldots, K-1\}$. The hypothesis space $\mathcal{H}$ is defined for ANNs as the set of functions $\{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ that can be used for classification and are parameterized by $\boldsymbol{\theta}$. The goal is to determine $\boldsymbol{\theta}$ such that the training error is minimized and a good generalization holds. The following

$$\hat{R}(g_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(g_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{y}_i \neq y_i),$$

is called empirical risk in statistical learning theory. The goal of training is to minimize the empirical risk by determining

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \hat{R}(g_{\boldsymbol{\theta}}).$$

The Empirical Risk Minimization (ERM) does not require any assumption about the distribution of the instance space. There are two errors that are particularly important in learning, the expected error and the generalization error. The expected error of an ANN is defined as

$$R(g_{\boldsymbol{\theta}}) = \mathbb{P}(g_{\boldsymbol{\theta}}(\mathbf{X}) \neq Y) = \mathbb{P}(\hat{Y} \neq Y).$$

This is also called the risk of $g_{\boldsymbol{\theta}}$. When the trained ANN is tested by random inputs, the risk measures how well the network performs on the unseen data. It might happen that the trained network overfits, that is, there is a certain function in the hypothesis space that gives a very low training error but a high risks. To account for this issue, the generalization error of $g_{\boldsymbol{\theta}}$ should be considered, which is defined as the difference between the empirical risk and the expected error, namely

$$\text{gen}(g_{\boldsymbol{\theta}}) = |R(g_{\boldsymbol{\theta}}) - \hat{R}(g_{\boldsymbol{\theta}})|.$$

The generalization error controls the difference between the expected error and the training error. A well trained ANN yields simultaneously a small generalization error and small expected error. If the training set S is randomly generated, the individual losses $\mathbf{1}(\hat{Y}_i \neq Y_i)$ are i.i.d. Bernoulli random variables and therefore the generalization error can be studied by controlling the deviation of the empirical risk $\hat{R}(g_{\boldsymbol{\theta}})$ from its expectation which is the expected error $R(g_{\boldsymbol{\theta}}) = \mathbb{E}(\hat{R}(g_{\boldsymbol{\theta}}))$.

The study of the generalization error is a central topic in statistical learning theory and there are many works on bounding the generalization error in a probabilistic way as a function of the number of samples n and certain measures of complexity of the hypothesis space $\mathcal{H}$, such as the Vapnik-Chervonenkis (VC) dimension. In this work we focus mostly on the expected error and we refer the interested readers to the classic references [11]–[13]. It is worth mentioning recent works where the generalization error is bounded using the mutual information $I(S;W)$ between the training set S considered as a random sample and a random variable W on the hypothesis space $\mathcal{H}$ [9], [10], [14].

In this work we analyze the learning process of ANNs by the information theoretic concept of mutual information between the true and estimated class label $I(Y;\hat{Y})$. Mutual information measures the amount of information that random variables contain about each other. It describes the reduction in the uncertainty of one random variable due to the knowledge of the other. Hence, successful training of ANNs should result in maximizing $I(Y;\hat{Y})$. We start to develop the corresponding analytical framework. Since $\hat{Y} = g_{\boldsymbol{\theta}}(\mathbf{X})$ with a deterministic function $g_{\boldsymbol{\theta}}$ it holds that

$$H(\hat{Y}|\mathbf{X}) = H(g_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X}) = 0.$$

Hence, the mutual information $I(Y;\hat{Y})$ can be written as

$$\begin{aligned} I(Y;\hat{Y}) &= H(\hat{Y}) - H(\hat{Y}|Y) - H(\hat{Y}|\mathbf{X}) \\ &= I(\mathbf{X};\hat{Y}) - H(\hat{Y}|Y). \end{aligned} \tag{1}$$

When training an ANN one should observe with the course of training epochs that the difference $I(\mathbf{X};\hat{Y}) - H(\hat{Y}|Y)$ is maximized by

- enlarging $I(\mathbf{X};\hat{Y})$, the mutual information between input and estimated label, and
- decreasing $H(\hat{Y}|Y)$, the conditional entropy of the estimated label given the true label.

After training, the former has to be large, the latter, interpreted as the conditional uncertainty, has to be small in order to maximize the difference.

We will observe how these quantities evolve during the learning phase of ANNs. For this purpose $I(\mathbf{X};\hat{Y})$ and $H(\hat{Y}|Y)$ will be estimated during the training phase. The estimation of information theoretic quantities will be discussed later. To anticipate an extremely interesting observation, in the training phase of the network first $I(\mathbf{X};\hat{Y})$ and $H(\hat{Y}|Y)$ in (1) are rapidly increased in concert so that $I(Y;\hat{Y})$ in (1) stays close to zero. In later phases $I(\mathbf{X};\hat{Y})$ increases only slowly while $H(\hat{Y}|Y)$ drops rapidly to a value close to zero. This behavior is typical all investigated examples and will be interpreted later. In the next section, we characterize the relation between the expected error $R(g_{\boldsymbol{\theta}})$ and other information theoretic quantities.

III. Bounds on the Expected Error

In this section, we investigate the relation between information theoretic quantities in (1) and the expected error of ANNs. There is no one-to-one correspondence as as may be seen from the following example. If the output $\hat{Y}$ of an ANN is a permuted version of the correct labels Y, the conditional entropy is zero, despite a bad training error. There are many works connecting the expected error $\mathbb{P}(Y \neq \hat{Y})$ to the conditional entropy $H(Y|\hat{Y})$, for instance see [15]–[17] and references therein. We start with the following result.

Proposition 1. *For a neural network $g_{\boldsymbol{\theta}}$, let $\hat{Y} = g_{\boldsymbol{\theta}}(\mathbf{X})$ be the output and $Y \in \{0,\dots,K-1\}$ the corresponding class label. The conditional entropy is upper bounded by a function of the expected error as*

$$\max\{H(Y|\hat{Y}), H(\hat{Y}|Y)\} \leq \Psi(R(g_{\boldsymbol{\theta}})), \tag{2}$$

where the function $\Psi(\cdot)$ is defined as

$$\Psi(x) = x\log(K-1) + h_b(x), \quad x \in [0,1]$$

with $h_b(x) = -x\log(x) - (1-x)\log(1-x)$ the binary entropy function.

Proof. Follows from Fano's inequality [18, Lemma 3.8]. □

This establishes formally that a small expected error implies a small conditional entropy. The function $\Psi(\cdot)$ is strictly increasing for $x \in [0, 1-\frac{1}{K}]$ and strictly decreasing otherwise. Fano's inequality is widely used in machine learning to provide an implicit lower bound on the expected error (for instance see [14] and references therein). In the context of ANNs, Proposition 1 implies that the conditional entropy $H(\hat{Y}|Y)$ cannot exceed $\Psi(R(g_{\boldsymbol{\theta}}))$. In Section IV, we will demonstrate experimentally that the points $(R(g_{\boldsymbol{\theta}}), H(\hat{Y}|Y))$ come very close to the curve $(x, \Psi(x))$ in later stages of the training process.

The case that the true labels are deterministically given by $\mathbf{X}$ is described by $Y = g(\mathbf{X})$ for some function g. There is zero error if $g_{\boldsymbol{\theta}}$ approximates g perfectly. Inequality (2) implies that in this case $H(\hat{Y}|Y) = 0$. In the remainder of this section we suppose that the class labels are not a deterministic function of the input $\mathbf{X}$ but that random errors may occur. This corresponds to the case that the expert, who allocates labels to input samples in the training set, makes mistakes from time to time. The following model describes the situation. Let $\tilde{Y} = g(\mathbf{X})$ be the true class label and

$$Y = (\tilde{Y} + B) \mod K, \tag{3}$$

where B is an independent random variable taking values in $\{0,\dots,K-1\}$. Proposition 1 provides a way to find a lower bound on the minimum expected error.

Proposition 2. *Consider the random variables $(\mathbf{X}, Y)$ where $Y \in \{0, \dots, K-1\}$ is a corrupted version of true labels $\tilde{Y} = g(\mathbf{X})$ according to* (3)*. Then, the expected error $R(g_{\boldsymbol{\theta}})$ is lower bounded by*

$$\Phi(H(B)) \leq R(g_{\boldsymbol{\theta}}),$$

where $\Phi : [0, \log K] \to [0, 1 - \frac{1}{K}]$ is the inverse function of $\Psi(x)$ on $[0, 1 - \frac{1}{K}]$.

Proof. For independent noise B, we have

$$H(Y|\hat{Y}) \geq H(Y|\hat{Y}, \tilde{Y}) = H(B|\hat{Y}, \tilde{Y}) = H(B).$$

From Proposition 1 it follows that $H(B) \leq \Psi(R(g_{\boldsymbol{\theta}}))$. Applying $\Phi(x)$, $x \in [0, 1 - \frac{1}{K}]$, completes the proof. □

As a special case we consider that the noise B is governed by the following distribution with parameter $p \in [0, 1 - \frac{1}{K}]$:

$$\mathbb{P}(B = i) = \begin{cases} 1 - p, & i = 0 \\ \frac{p}{K-1}, & i \in \{1, \dots, K-1\} \end{cases}. \tag{4}$$

By distribution (4) the correct class label occurs with probability $1 - p$, and any of the $K - 1$ labels occurs with the same probability $\frac{p}{K-1}$. The entropy of B is given by $h_b(p) + p \log(K-1) = \Psi(p)$. Therefore, Proposition 2 implies

$$R(g_{\boldsymbol{\theta}}) \geq \Phi(\Psi(p)) = p.$$

Proposition 2 provides a lower bound on the expected error. Nevertheless, as we discussed above, minimizing the expected error of a learning algorithm amounts to maximizing the mutual information $I(Y; \hat{Y})$. The following theorem formalizes this intuition when the ANN is successfully trained for the noise model in (4).

Theorem 1. *Consider the system model of Proposition 2 with uniformly distributed true labels $\tilde{Y}$ and the noise model* (4)*. If an ANN achieves the minimum expected error, i.e., $R(g_{\boldsymbol{\theta}}) = p$ then the following holds*

$$I(\mathbf{X}; \hat{Y}) = H(\hat{Y}) = \log K \tag{5}$$

$$H(\hat{Y}|Y) = \Psi(p)\,. \tag{6}$$

$$I(Y; \hat{Y}) = \log K - \Psi(p) \tag{7}$$

Proof. The proof of Proposition 2 shows that

$$H(Y|\hat{Y}) \geq H(Y|\hat{Y}, \tilde{Y}) = H(B). \tag{8}$$

On the other hand, since $R(g_{\boldsymbol{\theta}}) = p$, from Proposition 1 we have $H(Y|\hat{Y}) \leq \Psi(p)$. Therefore $H(Y|\hat{Y}) = \Psi(p)$. Since the random variable $\tilde{Y}$ corresponding to the true labels is uniformly distributed over $\{0, 1, \dots, K-1\}$, the entropy $H(\tilde{Y})$ is equal to $\log K$. Hence, equality (7) follows, namely the mutual information $I(Y; \hat{Y})$ writes as $I(Y; \hat{Y}) = \log K - \Psi(p)$. We proof the other two equalities using the following lemma.

Lemma 1. *If $R(g_{\boldsymbol{\theta}}) = p$ for the noise model* (4) *with $p < 1 - \frac{1}{K}$, it holds that*

$$\mathbb{P}(\hat{Y} = \tilde{Y}) = 1.$$

Proof. Note that $\mathbb{P}(\hat{Y} \neq Y) = R(g_{\boldsymbol{\theta}})$ and let $\delta \triangleq \mathbb{P}(\hat{Y} = \tilde{Y})$. For notational convenience we introduce $\tilde{Y} \oplus i$ to denote $(\tilde{Y} + i) \mod K$. Using the fact that the noise B is independent of other random variables, we have

$$\begin{aligned}
1 - p &= \mathbb{P}(\hat{Y} = Y) \\
&= \sum_{i=0}^{K-1} \mathbb{P}(\hat{Y} = Y, \hat{Y} = \tilde{Y}, B = i) + \mathbb{P}(\hat{Y} = Y, \hat{Y} \neq \tilde{Y}, B = i) \\
&= \mathbb{P}(\hat{Y} = \tilde{Y}, B = 0) + \sum_{i=1}^{K-1} \mathbb{P}(\hat{Y} = Y, \hat{Y} \neq \tilde{Y}, B = i) \\
&= P(B = 0)\delta + \sum_{i=1}^{K-1} \mathbb{P}(\hat{Y} = \tilde{Y} \oplus i, B = i) \\
&= P(B = 0)\delta + \frac{p}{K-1} \sum_{i=1}^{K-1} \mathbb{P}(B = i) \\
&= (1 - p)\delta + \frac{p}{K-1}(1 - \delta).
\end{aligned}$$

Since $p \neq 1 - \frac{1}{K}$, we have $\delta = 1$, i.e., $\mathbb{P}(\hat{Y} = \tilde{Y}) = 1$. □

The previous lemma implies that $H(\hat{Y}) = H(\tilde{Y}) = \log K$ and therefore (5) is proven. Finally, (6) follows from the other two equalities. □

The above propositions provide a way to lower bound the best expected error by the conditional entropy. Interestingly, our empirical studies show that these bounds are sharp, particularly Theorem 1 precisely predicts the mutual information and the conditional entropy for a successful training.

IV. Experiments

In this section, we empirically study the behavior of information theoretic quantities $I(\mathbf{X}; \hat{Y})$, $I(Y; \hat{Y})$ and $H(\hat{Y}|Y)$ during the training for different datasets and ANNs. The first issue is to properly estimate these quantities. Since $I(\mathbf{X}; \hat{Y}) = H(\hat{Y})$, it is sufficient to obtain a good estimation of the joint distribution of $(\hat{Y}, Y) \in \{0, \dots, K-1\}^2$ in order to approximate the mutual information and conditional entropies. A naive estimator directly computes the entropy from the empirical distribution of N independent observations of $(\hat{Y}, Y)$. It is shown in [19] that the approximation error incurred by this method is of order K^2/N. Hence, this approach yields a good approximation if $N \gg K^2$. This holds particularly for our experiments where the number of classes does not exceed 10 while the number of test examples are much larger than 10^2. We use this method to estimate the information theoretic quantities. When the number of classes K is large, one can consider other methods such as [20], [21].

A. Experiment Setup

For our experiments, we use the following three datasets:

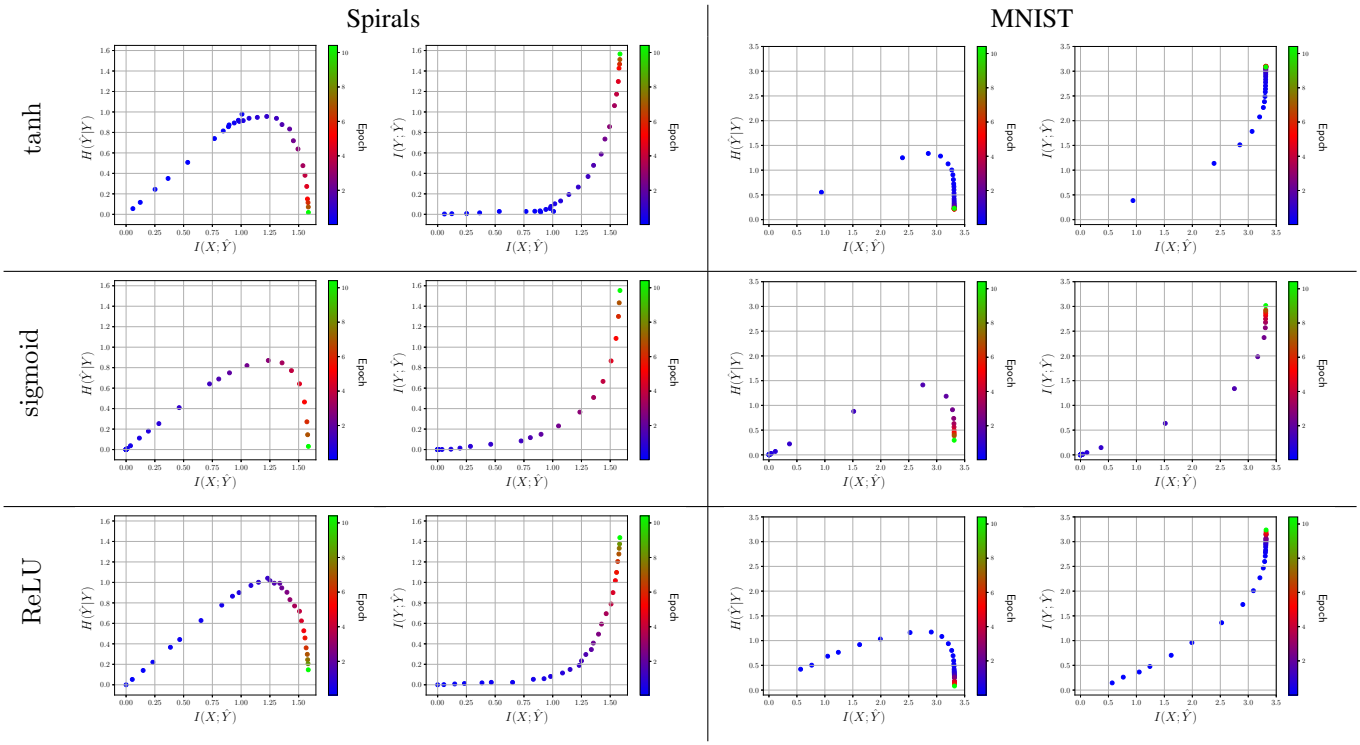

Fig. 2. Information theoretic quantities during the learning process for scenario 1.

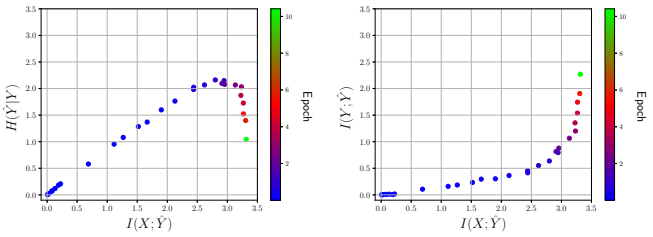

Fig. 3. Information theoretic quantities during the learning process for CIFAR-10 in scenario 1. These values are computed by averaging over 5 independent realizations of DenseNet.

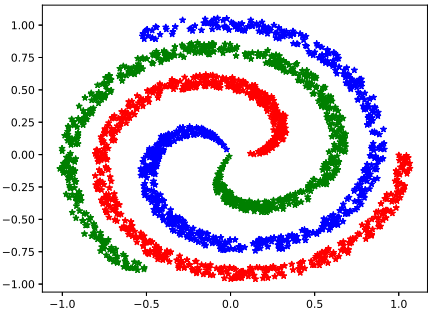

Fig. 4. Multi-class spiral classification dataset with 3 classes, represented using different colors.

- **Spirals:** The spirals dataset consists of two-dimensional points belonging to one of three spirals shown in Figure 4. This corresponds to $(\mathbf{X}, \tilde{Y})$ generated by

$$\mathbf{X} = \begin{pmatrix} (\sqrt{a}+b)\cos\left(2\pi a + \frac{2\pi}{3}\tilde{Y}\right) \\ (\sqrt{a}+b)\sin\left(2\pi a + \frac{2\pi}{3}\tilde{Y}\right) \end{pmatrix},$$

  where $a \in [0,1]$, $b \in [0, 0.1]$ and $\tilde{Y} \in \{0,1,2\}$ are independent uniformly distributed random variables. Moreover, this dataset is divided into a training set of $50\,000$ samples and a test set of $2\,000$.
- **MNIST:** This is a well known dataset for handwritten digit recognition [22]. It consists of $55\,000$ training images and $10\,000$ test images.
- **CIFAR-10:** This dataset consists of tiny RGB images belonging to 10 categories [23]. It contains $50\,000$ images for training and $10\,000$ for testing.

A fully connected ANN with four hidden layers of five neurons each, referred to as FCNN, is trained on the spirals dataset. For the MNIST dataset LeNet-5 [24] is used. To train these networks we let the learning rate $\gamma \in \mathbb{R}$ start at a given $\gamma_{\max} \in \mathbb{R}$ and then decay by 40% per epoch until reaching some given minimum learning rate $\gamma_{\min} < \gamma_{\max}$, that is $\gamma = \max\{\gamma_{\max} 0.6^{\lfloor \text{epoch} \rfloor}, \gamma_{\min}\}$. For the CIFAR-10 dataset we train a 100 layer DenseNet architecture as done in [25], but we stop the training after 10 epochs instead of the original 300 used by the authors. We train FCNN and LeNet-5 using various hidden layer activation functions, learning rates, and mini-batch sizes. The different configurations used for these experiments are summarized in Table I.

TABLE I
SIMULATION PARAMETERS

Dataset	Activation	Batch Size	γ_{max}	γ_{min}	Test Acc.
Spirals	tanh	128	10^{-1}	10^{-2}	99.7%
	sigmoid	128	10^{-1}	10^{-5}	99.6%
	ReLU	700	10^{-1}	10^{-5}	97.8%
MNIST	tanh	128	10^{-2}	10^{-2}	97.1%
	sigmoid	128	10^{-2}	10^{-4}	96.3%
	ReLU	128	10^{-2}	10^{-4}	99.1%
CIFAR-10	ReLU	64	10^{-1}	10^{-1}	80.2%

B. Experiment Results

Based on the above setup, the ANNs are trained to learn the noisy labels from the data. The true labels are corrupted by independent additive noise as in (3) with the noise distribution given in (4). We distinguish following scenarios:

- **Scenario 1 (noiseless labels):** This scenario corresponds to the case where $p = 0$. In other words, the best expected error is given by $R(g_{\boldsymbol{\theta}}) = 0$.
- **Scenario 2 (noisy labels):** This is the general setting where $p \in (0, 1 - \frac{1}{K}]$, thus $R(g_{\boldsymbol{\theta}}) \geq p$.

We are interested in studying the behavior of ANNs as they learn classification tasks to perfection. More precisely, we are only interested in studying those ANNs that managed to achieve the best performance during the training. For FCNN and LeNet-5 we only consider ANNs that achieved less than $0.05 + p$ error on their corresponding test set, and assume that they are correctly trained. For CIFAR-10, where Denset is only trained for 10 epochs instead of 300, we assume that ANNs with less than $0.2 + p$ test error are well trained. In Figure 2 the average behavior over 100 independent realizations of correctly trained ANNs during their learning process, for Scenario 1, is depicted. In these figures, a similar trend can be observed regarding the evolution of the information theoretic quantities during training. Regardless of the non-linearity and the dataset used, we observe that the learning process consists of two phases. The first phase occurs at the beginning of learning, where $I(\mathbf{X}; \hat{Y})$ increases even at the expense of increasing $H(\hat{Y}|Y)$. A possible explanation of this phenomenon is that, at the beginning of training it is more important to improve the information flow between $\mathbf{X}$ and $\hat{Y}$ than learning about the labels. In other words, learning about the input distribution is more important at early stages of the learning process. Note that learning about the distribution of $\mathbf{X}$ may be done in an unsupervised manner since the labels Y are not needed for this task. This behavior continues until a certain value of $I(\mathbf{X}; \hat{Y})$ is reached, from that point onward the second phase starts. In the spirals dataset the first learning phase ends around $I(\mathbf{X}; \hat{Y}) \approx 1.25$ for all used activation functions. The same behavior holds true for the MNIST and CIFAR-10 (see Figure 3) datasets, where the second phase of learning starts at $I(\mathbf{X}; \hat{Y}) \approx 3$. From this result we may conjecture the existence of a fundamental relation between the prediction task and a typical value of $I(\mathbf{X}; \hat{Y})$ where the second learning phase starts, which seems to be nearly independent of the ANNs architecture. As said, the second phase of learning starts when $I(\mathbf{X}; \hat{Y})$ is large enough. Then, minimizing $H(\hat{Y}|Y)$ plays a more significant role than maximizing $I(\mathbf{X}; \hat{Y})$. This phase can be intuitively seen as the discriminative phase of training, where ANNs learn to master the prediction task. The behavior of $I(\mathbf{X}; \hat{Y})$, $I(Y; \hat{Y})$ and $H(\hat{Y}|Y)$ during the learning process appears to be independent of the particular activation function. Hence, in Scenario 2 we focus on $\tanh$ for the spirals dataset, ReLU for the MNIST dataset, and assume that similar results hold for other activation functions. In Figure 5, the trajectory of information theoretic quantities is observed when the probability p is changed in (4). We first see that $I(Y; \hat{Y})$ and $H(\hat{Y}|Y)$ approach their corresponding bounds given in (7) and (6) as the training goes on. This supports the assumption that the models considered have in fact learned correctly since the expected error achieves the lower bound given by Fano's inequality. Note that regardless of p, the mutual information $I(\mathbf{X}; \hat{Y})$ approaches its maximum, i.e., $I(\mathbf{X}; \hat{Y}) = \log K$ in all experiments as it was predicted in (5). The first phase of learning ends at the value of $I(\mathbf{X}; \hat{Y})$ where $H(\hat{Y}|Y)$ reaches its maximum. This value of $I(\mathbf{X}; \hat{Y})$ where the first phase ends seems to increase with p in all simulations. This supports the idea that the channel between $\mathbf{X}$ and $\hat{Y}$ should be good enough, that is $I(\mathbf{X}; \hat{Y})$ should be above some threshold value before assigning the labels to data. As the labels get more noisy, we need a better channel between $\mathbf{X}$ and $\hat{Y}$ in order to learn correctly. Proposition 1 gives an upper bound on $H(\hat{Y}|Y)$ in terms of the expected error $R(g_{\boldsymbol{\theta}})$, thus a lower bound on $I(\hat{Y}; Y)$. We assume that the test error is a good estimate of $R(g_{\boldsymbol{\theta}})$, thus we use it to compute these bounds. Figure 6 shows that the pair $(R(g_{\boldsymbol{\theta}}), H(\hat{Y}|Y))$ approaches the dashed curve $(x, \Psi(x))$ for correctly trained ANNs and various values of p in Scenario 2. The result suggests that if ANNs are properly trained and the expected error $R(g_{\boldsymbol{\theta}})$ tends to p, the values of $H(\hat{Y}|Y)$ and $I(\hat{Y}; Y)$ approach their corresponding bounds given in Proposition 1. When DenseNet is used for classification of the CIFAR dataset, after only 10 epochs the pair $(R(g_{\boldsymbol{\theta}}), H(\hat{Y}|Y))$ starts to approach the curve even though the error is still far from the lower bound p. This is an indication that the model is learning correctly. Note that the model eventually reaches less than 6% test error after 300 epochs [25]. These experiments show that observing the trajectory of conditional entropy $H(\hat{Y}|Y)$, mutual information $I(\hat{Y}; Y)$ and expected error $R(g_{\boldsymbol{\theta}})$ can serve as a method for verifying the correct learning of ANNs.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[3] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. v. d. Driessche, T. Graepel, and D. Hassabis,

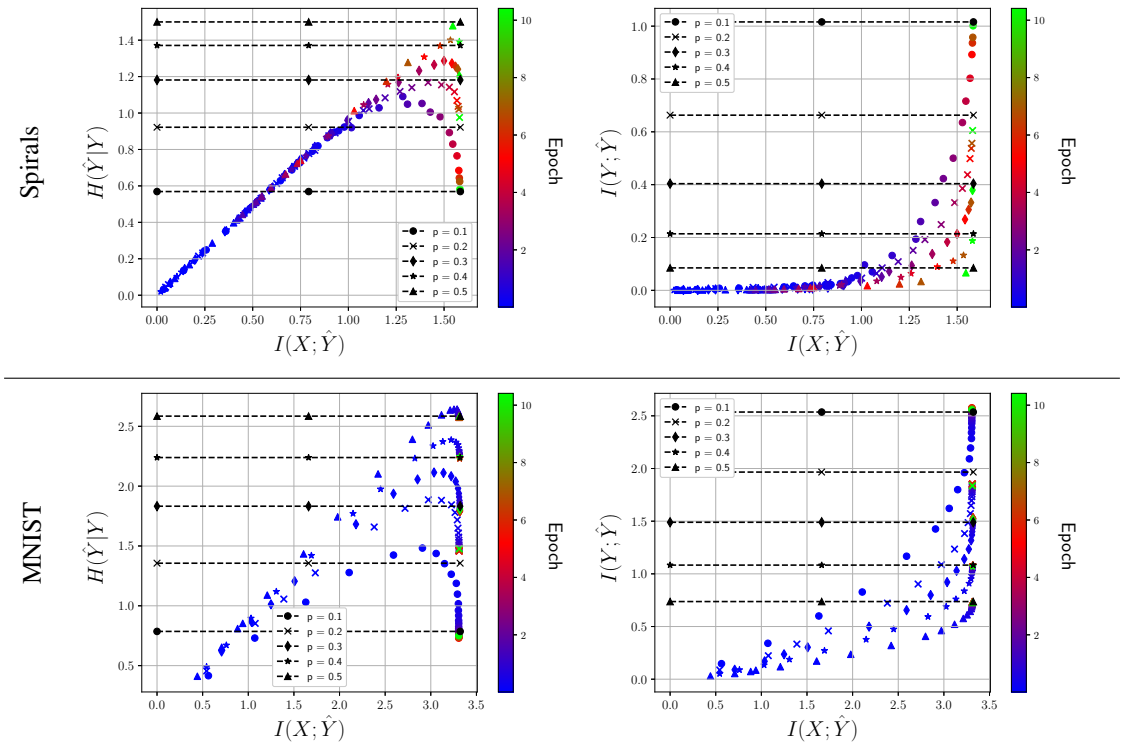

Fig. 5. Information theoretic quantities during the learning process for scenario 2. The black dashed lines represent the ideal limits, given in (6) and (7), to which $H(\hat{Y}|Y)$ and $I(Y;\hat{Y})$ converge. Various marker shapes are used to distinguish between experiments with different values of p.

"Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[4] R. Shwartz-Ziv and N. Tishby, "Opening the Black Box of Deep Neural Networks via Information," *arXiv preprint arXiv:1703.00810*, Mar. 2017.

[5] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.

[6] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," in *International Conference on Learning Representations*, 2018.

[7] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," *arXiv preprint arXiv:1805.09785*, 2018.

[8] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *arXiv preprint arXiv:1511.05219*, Nov. 2015.

[9] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2521–2530.

[10] A. R. Asadi, E. Abbe, and S. Verd, "Chaining Mutual Information and Tightening Generalization Bounds," *arXiv preprint arXiv:1806.03803*, Jun. 2018.

[11] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to Statistical Learning Theory," in *Advanced Lectures on Machine Learning*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Feb. 2003, pp. 169–207.

[12] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms*. New York, NY, USA: Cambridge University Press, 2014.

[13] M. Anthony and P. L. Bartlett, *Neural network learning: theoretical foundations*. Cambridge: Cambridge Univ. Press, 2009.

[14] M.-J. Zhao, N. Edakunni, A. Pocock, and G. Brown, "Beyond Fano's Inequality: Bounds on the Optimal F-score, BER, and Cost-sensitive Risk and Their Implications," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1033–1090, Apr. 2013.

[15] S.-W. Ho and S. Verdu, "Conditional entropy and error probability," in *2008 IEEE International Symposium on Information Theory*, Jul. 2008, pp. 1622–1626.

[16] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, Jan. 1994.

[17] I. Sason and S. Verd, "Arimoto rényi Conditional Entropy and Bayesian m-Ary Hypothesis Testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, Jan. 2018.

[18] I. Csiszr and J. Krner, *Information theory: coding theorems for discrete memoryless systems*, 2nd ed. Cambridge ; New York: Cambridge University Press, 2011.

[19] G. A. Miller, "Note on the bias of information estimates," *Information theory in psychology: Problems and methods*, vol. 2, no. 95, p. 100, 1955.

[20] T. Schürmann, "Bias analysis in entropy estimation," *Journal of Physics A: Mathematical and General*, vol. 37, no. 27, p. L295, 2004.

[21] E. Archer, I. M. Park, and J. W. Pillow, "Bayesian entropy estimation for countable discrete distributions," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2833–2868, 2014.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[23] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[24] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, contour and grouping in computer vision*. Springer, 1999, pp. 319–345.

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

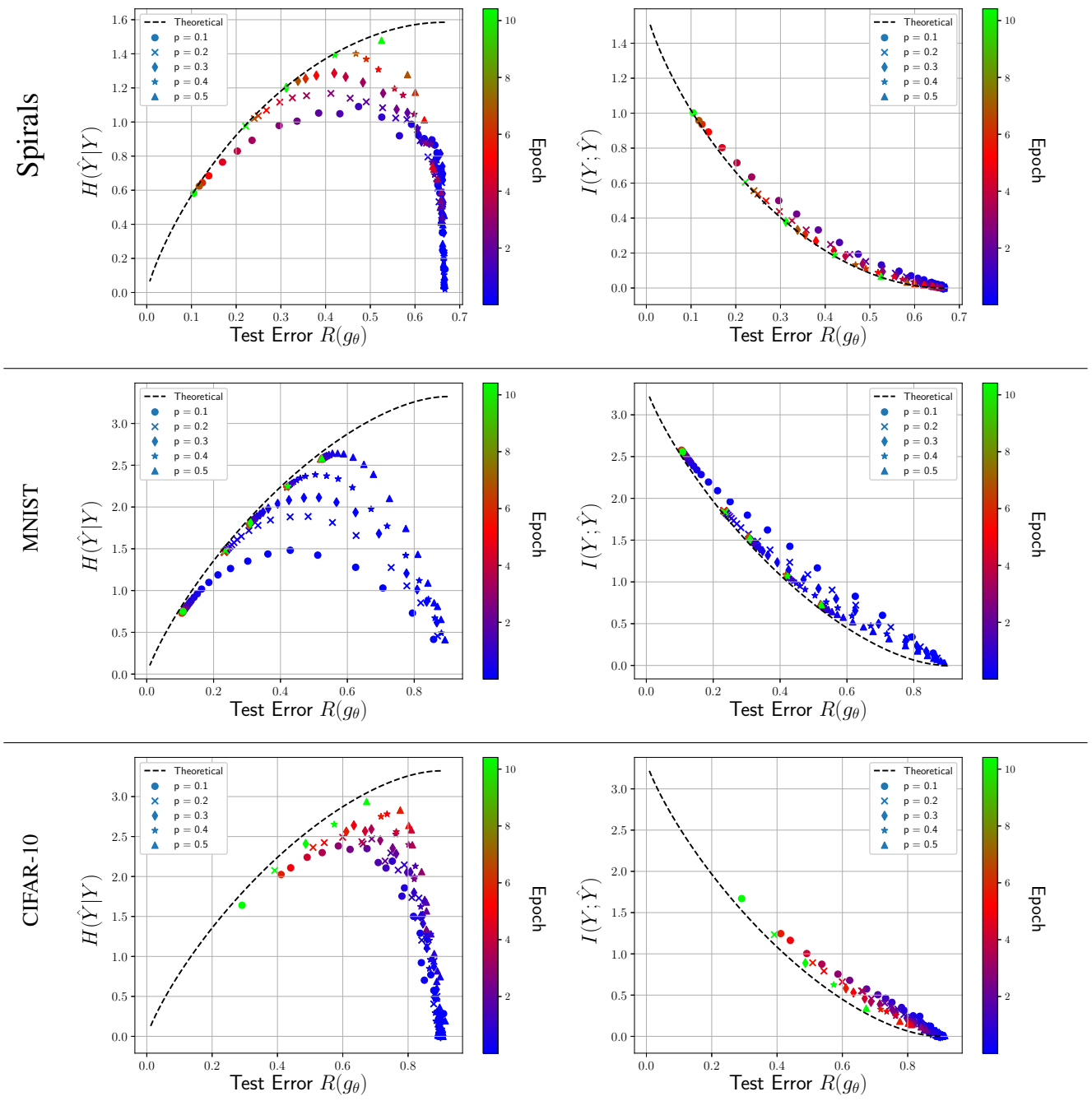

Fig. 6. Information theoretic quantities during the learning process for scenario 2. The black dashed lines depict the bounds obtained in Proposition 1. Various marker shapes are used to distinguish between experiments with different values of p.