University of Illinois at Urbana-Champaign

Statistical Learning Theory
Course Project Report

ECE 543

Learning Through The Lens of Information Bottleneck

Author:
Dariush Kari

Instructor:
Professor Bruce Hajeck

May 10, 2019

Contents

Abstract

In this report, we will investigate the concept of Information Bottleneck (IB) and its connection to machine learning. Regarding this concept, a learning algorithm seeks to find the most compressed while informative (about the output or labels) representation of the input (features). The mathematical guarantees for this claim are explored, which use the tools we have learned in the Statistical Learning Theory course. Furthermore, the IB notion has been used to explain the deep neural networks' good performance, although it has become controversial in the machine learning community. Hence, we briefly discuss its application in the deep learning.

1 Introduction

An essential issue in the learning algorithms is the notion of relevance, which is linked to the concept of sufficient statistics [2]. The concept of minimal sufficient statistics has been used to obtain the simplest representation of the input that contains the relevant components about a desired output. However, the minimal sufficient statistics does not always admit a bounded dimensionality [2]. The concept of Information Bottleneck (IB) can be considered as an information theoretic generalization of the minimal sufficient statistics [4], in which a compressed representation of the input is obtained that has a certain amount of relevant information about the output. The compression is quantified by the mutual information between the input and the representation, whereas the relevance (informativeness) is quantified by the mutual information between the representation and the output. In the sequel, we show how this concept is related to the learning algorithms.

In learning algorithms and more specifically classification tasks, there is an input X and a corresponding output Y, and we seek to find a mapping between the input and output via observing a finite sample generated by a fixed and unknown distribution $p(x, y)$. Therefore, the learning algorithms find the most important features of the input based on observing a finite number of samples. Usually there are many features in the input, not all of which are related to the output, i.e., the entropy of the input is usually much larger than that of the output. Therefore, it has been of interest to find the most relevant features of the input. However, the relevant features can be captured by using the IB concept. In fact, using the IB, one achieves a compressed representation of the input by throwing away the unrelated information about the output, and this representation can be further processed to predict the output. We illustrate this discussion in the following examples, which can help the reader have a better idea of what will be discussed in the next sections.

Suppose we are going to design an algorithm for detecting a "cat" in a given input image. When representing the image as a vector of real numbers, each pixel of the

image has a random value indicating its color. Thus, there are a large number of possible random images that can be fed into the algorithm, yielding a very high input entropy. For example, if the image is black and white (i.e., the pixels are 0 or 1) with N pixels generated at random from a uniform distribution on the set of all possible images, its entropy equals N, i.e., we need N bits to exactly reconstruct the image. However, the output is either 0 (corresponding to "there is no cat in the image") or 1 ("there is at least 1 cat in the image"). As a result, the output entropy is only 1. We observe that the input entropy in this example is much higher than the output. Furthermore, note that as long as there is a cat in the image, the output is 1, regardless of the other objects present in the image. Roughly speaking, the algorithm should "ignore" everything but the cat in the image, i.e., it should generalize well.

As another example, consider an algorithm whose job is to recognize the gender (male or female) of a person who is talking on a background music. Again, the input consists of a background music and a person's voice, while the output is binary. Therefore, the algorithm have to be able to disregard the irrelevant parts (e.g., remove the background music) and provide a useful representation of the input data.

As the final example consider a digital communications system. In digital communications, usually a message (generated at the transmitter side) is transmitted over a noisy channel, through which a random noise is added to the desired signal. At the receiver, there is an algorithm to estimate the transmitted signal by observing the noisy version of it. The output (desired signal) has a lower entropy than the observed noisy signal (input to the algorithm at the receiver's end) as it seeks to approximate the noiseless transmitted signal, which usually comes from a finite alphabet. For instance, the message can be a sequence of 0's or 1's, while the received signal is contaminated by Gaussian noise. Hence, the received signal corresponding to each transmitted bit is a real number that needs an infinite precision (infinite number of bits) to be determined uniquely.

In all of these examples, we seek to find an algorithm that can find an "appropriate" representation of the input that can discard the irrelevant information of the input. Note that not in all classification tasks is there a deterministic relationship between the input and the output. For example, if the labels are noisy due to the fact that human labelers cannot agree on a certain label for a specific input. Hence, we consider a probabilistic mapping between the input and output.

This report is organized as follows. Section 2 is devoted to concept of information bottleneck. Section 3 provides the mathematical connection between the IB and classification task and provides the upper bounds on the performance and general-

ization errors. Section 4 interprets the deep neural networks using the IB concept. Finally, we conclude the report with some remarks in Section 5.

1.1 Notations

Throughout this report, random variables (assumed to be discrete) are shown by capital letters and their realization are shown by small letters. We denote the input and output by $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, respectively. Also, $T \in \mathcal{T}$ indicates a representation of the input. the mutual information between two random variables A and B is denoted by $\mathbb{I}(A;B)$. The approximation of the arbitrary quantity a is depicted by $\widehat{a}$. $\mathbb{E}$ denotes the expectation.

2 Information Bottleneck (IB)

Information Bottleneck [4] has been introduced as a tradeoff between two objectives in machine learning algorithms, (a) compressing the input (for good generalization) and (b) minimizing the training error, and has been used to obtain some generalization bounds on the learning algorithms [2]. This concept introduced the idea of using information plane to observe and qualitatively analyze the performance of a learning algorithm. In this sense, an algorithm in fact seeks to find an optimal representation of the input with respect to the output that can be well generalized. As an example consider a classification algorithm, in which the input data has a higher entropy than the output. Note that when the algorithm tries to minimize the training error, it inherently seeks for the representations (extracted features) of the input that preserve as much as possible the relevant information about the output labels [4]. Obviously, these representations cannot provide more information about the output, than the original input, due to the Data Processing Inequality (DPI). However, while the input provides the maximum possible information on the output (among all possible representations), it also contains irrelevant information such as the noise in the input data. This irrelevant information can affect the generalization performance and cause over-fitting. Nevertheless, the over-fitting can be avoided by discarding the irrelevant information and preserving only the part of information relevant to the output labels.

In order to address the tradeoff between minimizing the irrelevant information and maximizing the relevant information, one may use the concept of mutual information between the input and the representation and between the representation and the output. If we denote the input by X, the representation by T, and the output by Y, the information bottleneck tradeoff can be formulated as

$$\min_{T} \mathbb{I}(X;T) - \beta\, \mathbb{I}(T;Y), \tag{1}$$

where β is a positive constant controlling the tradeoff [4]. Note that T is a function of X. However, if the function is invertible, the entropy of X and T will be the same, which is not desired. Therefore, a good representation of X is obtained through a non-invertible (possibly stochastic) function of X. In addition, given X, the representation T and the output Y become statistically independent. Thus, these random variables construct a Markov chain as $Y - X - T$.

According to the Markov chain and using the DPI, we see that $I(T;Y) \leq I(X;Y)$. Furthermore, denoting by $p(T|X)$ the conditional probability of T given X, which defines the mapping between the representation and the output, one can find the solution to (1) by iteratively minimizing the following objective function over the set of corresponding probabilities [4].

$$\min_{p(t|x),p(t),p(y|t)} \mathbb{I}(X;T) + \beta D_{KL}(p(y|x)\|p(y|t)), \tag{2}$$

Note that due to the Markov chain, $p(y|t) = \sum_{x\in\mathcal{X}} p(y|x)p(x|t)$ and the following iterations provide a solution to (2) [4].

$$\begin{aligned} p_{k+1}(t|x) &= \frac{p_k(t)}{Z_k(x,\beta)} \exp\left[-\beta D_{KL}(p(y|x)\|p_k(y|t)\right] \\ p_{k+1}(t) &= \sum_{x\in\mathcal{X}} p_k(t|x)p(x) \\ p_{k+1}(y|t) &= \sum_{x\in\mathcal{X}} p(y|x)p(x|t), \end{aligned}$$

where $Z_k(x,\beta)$ is a normalization factor.

The optimization problem in (1) expresses a trade off between $\mathbb{I}(X;T)$ and $\mathbb{I}(T;Y)$, and by adjusting different values of β, different solutions are obtained, in each of which we have

$$\frac{\delta\mathbb{I}(T;Y)}{\delta\mathbb{I}(X;T)} = \beta^{-1},$$

where $\delta I/\delta J$ indicates the functional derivative of I with respect to J. By obtaining solutions for different values of β, we can determine the possible pairs of $(\mathbb{I}(X;T), \mathbb{I}(T;Y))$, which can be shown in a 2-D information plane [4].

3 Learning and Generalization Analysis

In this section, we focus on the connection of the information bottleneck and classification, and provide the analysis for the performance and generalization [2]. The analyses show that if there is enough number of samples, one can approximate the mutual information in the IB formulation as accurate as possible.

3.1 Finite Sample Analysis

Since we do not have access to the joint distribution of the input and output in a learning task, we obtain the upper bounds on the error in approximating the quantities that are used in IB framework. These analyses along with their interpretations are presented in the next four theorems, all of which are from [2].

Theorem 1. Let T be any arbitrary random mapping of X, determined by $p(t|x)$, and let $\mathcal{S}$ denote a sample of size m. With probability at least $1-\delta$, where $\delta \in (0,1)$, we have

$$|\mathbb{I}(X;T) - \widehat{\mathbb{I}}(X;T)| \leq \frac{(|\mathcal{T}|\log(m) + \log(|\mathcal{T}|))\sqrt{\log(4/\delta)}}{\sqrt{2m}} + \frac{|\mathcal{T}|-1}{m}$$

and also simultaneously

$$|\mathbb{I}(Y;T) - \widehat{\mathbb{I}}(Y;T)| \leq \frac{((3|\mathcal{T}|+2)\log(m))\sqrt{\log(4/\delta)}}{\sqrt{2m}} + \frac{(|\mathcal{Y}|+1)(|\mathcal{T}|+1)-4}{m}$$

Proof. To prove this theorem, we first find upper bounds on $|\widehat{\mathbb{I}}(X;T) - \mathbb{E}\widehat{\mathbb{I}}(X;T)|$, and $|\mathbb{E}\widehat{\mathbb{I}}(X;T) - \mathbb{I}(X;T)|$ and then use the triangle inequality to obtain the stated bound. We repeat the similar steps for $\mathbb{I}(Y;T)$.

We are going to bound the difference in $\widehat{H}(T)$ and $\widehat{H}(T|x)$, occurred by changing only one sample, and then use the McDiarmid's inequality. Note that by replacing (x,y) with (x',y'), the empirical distribution $p(x,y)$ changes by at most $1/m$, which in turn means that $\widehat{p}(x)$ and $\widehat{p}(x')$ also change by at most $1/m$. As a result, for a fixed $t \in \mathcal{T}$, $\widetilde{p}(t) = \sum_x p(t|x)\widehat{p}(x)$ also changes by at most $1/m$. Also,

$$\widehat{H}(T) = -\sum_{t\in\mathcal{T}} \widetilde{p}(t)\log(\widetilde{p}(t)),$$

and since according to lemma 2 (that appears in the proof of Theorem 3), for any natural m and for $a \in [0, 1-1/m]$ and $\Delta \leq 1/m$,

$$(a+\Delta)\log(a+\Delta) - a\log(a) \leq \frac{\log(m)}{m},$$

it yields that $\widehat{H}(T)$ changes by at most $|\mathcal{T}|\log(m)/m$.

In order to bound the difference in $\widehat{H}(T|X)$, note that

$$\widehat{H}(T|X) = \sum_x \widehat{p}(x)H(T|X=x).$$

Because $H(T|X=x) \leq \log(|\mathcal{T}|)$, by changing a single instance, $\widehat{H}(T|X)$ changes by at most $\log(|\mathcal{T}|)/m$. Overall, $\widehat{\mathbb{I}}(X;T) = \widehat{H}(T) - \widehat{H}(T|X)$ changes by at most $(|\mathcal{T}|\log(m) + \log(|\mathcal{T}|))/m$. Now, using the McDiarmid's inequality, with probability at least $1-\delta_1$

$$|\widehat{\mathbb{I}}(X;T) - \mathbb{E}\widehat{\mathbb{I}}(X;T)| \leq \frac{(|\mathcal{T}|\log(m) + \log(|\mathcal{T}|))\sqrt{\log(2/\delta_1)}}{\sqrt{2m}}.$$

For $\widehat{\mathbb{I}}(T;Y) = H(T) + H(Y) - H(T,Y)$, by similar reasoning, we achieve

$$|\widehat{\mathbb{I}}(Y;T) - \mathbb{E}\widehat{\mathbb{I}}(Y;T)| \leq \frac{(3|\mathcal{T}|+2)\log(m)\sqrt{\log(2/\delta_2)}}{\sqrt{2m}}.$$

In order to complete the proof, we use the following lemma [2].

Lemma 1. For a random variable X with empirical estimate $\widehat{H}(.)$ on its entropy, based on an i.i.d. sample of size m, we have

$$|\mathbb{E}\widehat{H}(X) - H(X)| \leq \frac{|\mathcal{X}|-1}{m}.$$

By using this lemma for X, T, Y and (Y,T), and using the triangle inequality we can achieve the upper bounds on $|\widehat{\mathbb{I}} - \mathbb{I}|$ terms. Then by setting $\delta_1 = \delta_2 = \delta/2$, we observe that the upper bounds in the theorem hold simultaneously with probability at least $(1-\delta_1)(1-\delta_2) = 1 - \delta + \delta^2/4 \geq 1-\delta$. □

Note that the bounds in 1, do not depend on $|\mathcal{X}|$ (note that $p(t|x)$ is assumed to be known). Moreover, the dependence on $|\mathcal{Y}|$ is also through a factor of $\frac{1}{m}$, while the dominant term is a factor of $\frac{1}{\sqrt{m}}$. Also note that the previous bound holds uniformly for all Ts, which means that we do not use the sample to obtain T. Nevertheless, a tighter bound can be obtained for each specific T, as presented in the next theorem. Before stating the theorem, we define some simplifying notations. Assume that the elements of $\mathcal{X}$ and $\mathcal{Y}$ are all ordered. Also, assume that $\boldsymbol{p}(T=t|x) = [p(T=t|x_1), ..., p(T=t|x_{|\mathcal{X}|})]$ and $\boldsymbol{H}(T|x) = [H(T|x_1), ..., H(T|x_{|\mathcal{X}|})]$ be vectors. $\widehat{\boldsymbol{H}}$ is also defined similarly. Furthermore, for any real valued vector $\mathbf{a}$, we define the function V as

$$V(\mathbf{a}) = \sum_{i=1}^{n} \left(a_i - \frac{1}{n}\sum_{j=1}^{n} a_j \right)^2.$$

In addition, we define $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\phi(x) = \begin{cases} 0 & x = 0 \\ x\log(1/x) & 0 < x \leq 1/e \\ 1/e & x > 1/e, \end{cases}$$

which is an increasing concave function.

Theorem 2. With the assumptions in Theorem 1, it holds with probability at least $1-\delta$ that

$$|\mathbb{I}(X;T)-\widehat{\mathbb{I}}(X;T)| \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\ V(\boldsymbol{H}(T|x))}{m}}+\sum_t \phi\left(\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\ V(\boldsymbol{p}(T=t|x))}{m}}\right)$$

and also simultaneously

$$|\mathbb{I}(Y;T)-\widehat{\mathbb{I}}(Y;T)| \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\ V(\widehat{\boldsymbol{H}}(T|x))}{m}}+2\sum_t \phi\left(\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\ V(\boldsymbol{p}(T=t|x))}{m}}\right)$$

Proof. First note that $|\mathbb{I}(X;T)-\widehat{\mathbb{I}}(X;T)| \leq |H(T|X)-\widehat{H}(T|X)|+|H(T)-\widehat{H}(T)|$. Moreover, since for two probability distributions p and $\widehat{p}$, we always have (for any scalar a) $0=a-a=a\sum_x p(x)-a\sum_x \widehat{p}(x)=\sum_x a(p(x)-\widehat{p}(x))$,

$$|H(T|X)-\widehat{H}(T|X)| \leq \|\boldsymbol{p}-\widehat{\boldsymbol{p}}\|\|\boldsymbol{H}(T|x)-a\|,$$

which, by choosing $a=\frac{1}{|\mathcal{X}|}\sum_x H(T|x)$, changes to

$$|H(T|X)-\widehat{H}(T|X)| \leq \|\boldsymbol{p}-\widehat{\boldsymbol{p}}\|\sqrt{V(\boldsymbol{H}(T|x))}.$$

In addition, it is straightforward to show that following lemma holds.

Lemma 2. For $a,b\in[0,1]$,

$$|a\log(a)-b\log(b)| \leq \phi(|a-b|).$$

Hence

$$\begin{aligned}
|H(T)-\widehat{H}(T)| &= |\sum_t p(t)\log(p(t))-\widehat{p}(t)\log(\widehat{p}(t))| \\
&\leq \sum_t \phi(|p(t)-\widehat{p}(t)|) \\
&= \sum_t \phi(\sum_x p(t|x)|p(x)-\widehat{p}(x)|) \\
&\leq \sum_t \phi(\sqrt{V(\boldsymbol{p}(T=t|x))}\|\boldsymbol{p}(x)-\widehat{\boldsymbol{p}}(x)\|).
\end{aligned}$$

Thus, we obtain

$$|\mathbb{I}(X;T)-\widehat{\mathbb{I}}(X;T)| \leq \sum_t \phi(\sqrt{V(\boldsymbol{p}(T=t|x))}\|\boldsymbol{p}(x)-\widehat{\boldsymbol{p}}(x)\|)+\|\boldsymbol{p}(x)-\widehat{\boldsymbol{p}}(x)\|\sqrt{V(\boldsymbol{H}(T|x))}.$$

Following the similar steps we achieve

$$\begin{aligned}|\mathbb{I}(Y;T)-\widehat{\mathbb{I}}(Y;T)| \leq &\sum_t \phi(\sqrt{V(\boldsymbol{p}(T=t|x))}\|\boldsymbol{p}(x)-\widehat{\boldsymbol{p}}(x)\|)\\ &+\sum_y p(y)\sum_t \phi(\sqrt{V(\boldsymbol{p}(T=t|x))}\|\boldsymbol{p}(x|y)-\widehat{\boldsymbol{p}}(x|y)\|)\\ &+\|\boldsymbol{p}(y)-\widehat{\boldsymbol{p}}(y)\|\sqrt{V(\widehat{\boldsymbol{H}}(T|y))}.\end{aligned}$$

However, note that if $\widehat{\rho}$ is an empirical estimation of the true distribution ρ based on m samples, we have with probability at least $1-\delta$

$$\|\widehat{\rho}-\rho\|_2 \leq \frac{2+\sqrt{2\log(1/\delta)}}{\sqrt{m}}.$$

In order to bound $|\boldsymbol{p}(x)-\widehat{\boldsymbol{p}}(x)\|$, $|\boldsymbol{p}(y)-\widehat{\boldsymbol{p}}(y)\|$, and $|\boldsymbol{p}(x|y)-\widehat{\boldsymbol{p}}(x|y)\|$ (for each fixed $y\in\mathcal{Y}$), it is sufficient to replace δ with $\frac{\delta}{|\mathcal{Y}|+2}$ (similar to what we did in the proof of Theorem 1). By this substitution and noting that

$$2+\sqrt{2\log(\frac{|\mathcal{Y}|+2}{\delta})} \leq \sqrt{C\log(\frac{|\mathcal{Y}|}{\delta})}$$

for a constant C, the bounds in Theorem 2 are proven. □

According to Theorem 2, the smoother $\boldsymbol{p}(T=t|x)$, the tighter the bound will be. In the extreme case, if $p(T=t|x)=p(T=t)$, i.e., T is independent of X, the generalization bounds become zero. However, in this case T carries no information about X or Y, i.e., $\mathbb{I}(X;T)=\mathbb{I}(T;Y)=\widehat{\mathbb{I}}(X;T)=\widehat{\mathbb{I}}(T;Y)=0$. Next theorem proven in [2] provides a simpler but looser upper bound by worst case assumptions on the statistical dependency of T on X.

Theorem 3. With the assumptions in Theorem 1, it holds with probability at least $1-\delta$ that

$$|\mathbb{I}(X;T)-\widehat{\mathbb{I}}(X;T)| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(\sqrt{|\mathcal{T}||\mathcal{X}|}\log(m)+|\mathcal{X}|^{1/2}\log(|\mathcal{T}|)\right)+\frac{1}{e}|\mathcal{T}|}{\sqrt{m}}$$

and also simultaneously

$$|\mathbb{I}(Y;T)-\widehat{\mathbb{I}}(Y;T)| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(\sqrt{|\mathcal{T}||\mathcal{X}|}\log(m)+\frac{1}{2}|\mathcal{Y}|^{1/2}\log(|\mathcal{T}|)\right)+\frac{2}{e}|\mathcal{T}|}{\sqrt{m}}.$$

We see that when $|\mathcal{T}| \ll |\mathcal{Y}|$, the bound is tight, which means that we can approximate the mutual information accurately even when there is not enough number of samples to accurately approximate $p(x,y)$.

3.2 A Learning Theoretic Perspective

The IB concept optimizes a tradeoff between $\mathbb{I}(X;T)$ and $\mathbb{I}(Y;T)$. On the other hand, the learning algorithms usually optimize a tradeoff between a risk function and a generalization term. In this section we show how $\mathbb{I}(Y;T)$ is related to the risk minimization, while $\mathbb{I}(X;T)$ serves as a generalization term. To illustrate this connection we use the example of document categorization. Here, $\mathcal{Y}$ is the set of possible document types (e.g., "academic", "political", etc.) and $\mathcal{X}$ is a set of possible words. When a new document $D = \{x_1, ..., x_n\}$ is provided, the empirical distribution of T given D is obtained by

$$\widetilde{p}(t) = \sum_{i=1}^{n} p(t|x_i)\widehat{p}(x_i),$$

while

$$\widehat{p}(t|y) = \sum_{x} p(t|x)\widehat{p}(x|y).$$

Then, assuming that the document is sampled according to $p(t|y)$, the most probable class is

$$y^* = \arg\min_{y} D_K L[\widetilde{p}(t)\|\widehat{p}(t|y)],$$

which reveals that $\mathbb{I}(Y;T)$ is a reasonable objective function in this case.

Assume that the true class is y_1. Moreover assume that $\widehat{p}(t|y_1) = p(t|y_1)$ and $\widehat{p}(t|y_2) = p(t|y_2)$. Then , regarding the Stein's lemma in information theory, the error incurred by predicting y_2, shown by α_n, satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log(\alpha_n) = -D_{KL}[p(t|y_2)\|p(t|y_1)].$$

Using the union bound, we achieve the following upper bound on the error probability

$$\sum_{y\neq y_1} \exp(-nD_{KL}[p(t|y)\|p(t|y_1)]).$$

Note that due to the convexity of KL divergence, we have

$$\begin{aligned}
\mathbb{I}(T;Y) &= \mathbb{E}_y D_{KL}(p(t|y)\|p(t)) \\
&= \mathbb{E}_y D_{KL}(p(t|y)\|\mathbb{E}_{y'} p(t|y')) \\
&\leq \mathbb{E}_{y,y'} D_{KL}[p(t|y)\|p(t|y')].
\end{aligned}$$

As a result, $-n\mathbb{I}(T;Y)$ will be an upper bound on the term inside the exponential under the summation. Therefore, the higher $\mathbb{I}(T;Y)$ is, the less the error probability will be, i.e., $-\mathbb{I}(T;Y)$ can be considered as a risk function.

The next theorem connects $\mathbb{I}(X;T)$ to the error in approximating $\mathbb{I}(Y;T)$. Based on this theorem, $\mathbb{I}(X;T)$ can be considered as a generalization term, which is desired to be as small as possible.

Theorem 4. Under the same conditions as the previous theorems, with probability at least $1-\delta$,

$$|\mathbb{I}(Y;T)-\widehat{\mathbb{I}}(Y;T)| \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)}{m}}\left(C_1\log(m)\sqrt{|\mathcal{T}|\mathbb{I}(X;T)}+C_2|\mathcal{T}|^{3/4}\mathbb{I}(X;T)^{1/4}+C_3\widehat{\mathbb{I}}(X;T)\right)$$

where C is the same constant as in Theorem 1, and C_1, C_2, and C_3 depend only on $p(x)$ and $p(y)$.

Proof. Since this proof is very long while it does not contain new tricks related to the course, we refer the reader to [2]. However, a sketch of the proof is provided. First, an upper bound based on the L_1 norm is obtained for $|\mathbb{I}(Y;T)-\widehat{\mathbb{I}}(Y;T)|$. Then, the key step is relating the L_1 norm to the KL divergence using the Lemma 12.6.1 in Cover and Thomas's information theory book, i.e.,

$$\|p(x|t)-p(x)\|_1 \leq \sqrt{2\log(2)D_{KL}(p(x|t)\|p(x))}.$$

□

This finally introduces the $\mathbb{I}(X;T)$ in the upper bound.

4 IB in Deep Learning

It has been suggested by [5, 3] that deep neural networks (DNNs) inherently optimize an IB cost. The layered structure of the DNNs, generates a Markov chain of consecutive representations of the input, in which the output of each layer is the input to the next layer. It has been observed that in a fully connected network with sigmoid activation functions performed on a binary classification task, and using the SGD for training, the DNN first increases $\mathbb{I}(Y;T)$ and then in many more iterations, it tries to reduce $\mathbb{I}(X;T)$. This is believed to be the reason why DNN do not over-fit. However these ideas have been challenged by [1], the discussion of which is out of the scope of this project.

5 Conclusion

This report studies the information bottleneck concept, its connection to machine learning algorithms. Furthermore, an interpretation of the deep neural networks using the IB framework is represented. Regarding the IB concept, each learning algorithm inherently tries to find the most compressed representation of the input

that preserve the necessary information about the output, where maintaining the relevant information is linked to the risk minimization, while compression is related to the generalization. Also, DNNs seem to obtain a representation of the input corresponding to an information bottleneck problem by consecutively processing the input in each layer. However, there are still many concerns about the correctness of this idea about DNNs.

References

[1] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

[2] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

[3] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[4] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[5] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pages 1–5. IEEE, 2015.