Learning and Generalization with the Information Bottleneck

Ohad Shamir[†], Sivan Sabato[†*], and Naftali Tishby[†‡]

[†] School of Computer Science and Engineering, [‡] Interdisciplinary Center for Neural Computation,
The Hebrew University, Jerusalem 91904, Israel
[*] IBM Research Laboratory in Haifa, Haifa 31905, Israel
{ohadsh,sivan_sabato,tishby}@cs.huji.ac.il

Abstract. The Information Bottleneck is an information theoretic framework that finds concise representations for an 'input' random variable that are as relevant as possible for an 'output' random variable. This framework has been used successfully in various supervised and unsupervised applications. However, its learning theoretic properties and justification remained unclear as it differs from standard learning models in several crucial aspects, primarily its explicit reliance on the joint input-output distribution. In practice, an empirical plug-in estimate of the underlying distribution has been used, so far without any finite sample performance guarantees. In this paper we present several formal results that address these difficulties. We prove several finite sample bounds, which show that the information bottleneck can provide concise representations with good generalization, based on smaller sample sizes than needed to estimate the underlying distribution. The bounds are non-uniform and adaptive to the complexity of the specific model chosen. Based on these results, we also present a preliminary analysis on the possibility of analyzing the information bottleneck method as a learning algorithm in the familiar performance-complexity tradeoff framework. In addition, we formally describe the connection between the information bottleneck and minimal sufficient statistics.

1 Introduction

The Information Bottleneck (IB) method, introduced in [23], is an information-theoretic framework for extracting relevant components of an 'input' random variable X, with respect to an 'output' random variable Y. This is performed by finding a *compressed,* non-parametric and model-independent representation T of X, that is most *informative* about Y. Formally speaking, the notion of compression is quantified by the mutual information between T and X, while the informativeness is quantified by the mutual information between T and Y. A scalar Lagrange multiplier β smoothly controls the tradeoff between these two quantities.

The method has proven to be useful for a number of important applications (see [24, 8, 21] and references therein), but its learning theoretic justification has remained unclear, for two main reasons: (i) The method assumes that the joint distribution of X and Y is known, and uses it explicitly. This stands in contrast to most finite-sample based machine learning algorithms. In practice, the empirical co-occurrence distribution is used to calculate a plug-in estimate of the IB functional, but without explicit regularization, finite-sample generalization bounds or error guarantees of any kind. Moreover, it was not clear what is left to be learned if it is assumed that this distribution is known. (ii) IB is formally related to classical information theoretic problems, such as Rate-Distortion theory and Coding with Side-Information. It is, however, unclear why maximizing mutual information about Y is useful for any "natural" learning theoretic model, and in particular how it is related to classification error.

In this paper we provide rigorous answers to some of the above issues concerning the IB framework. We focus on a learning theoretic analysis of this framework, where X and Y are assumed to be discrete, and the empirical distribution of $p(x, y)$ is used as a plug-in for the true distribution. We develop several finite sample bounds, and show that despite this use of plug-in estimation, the IB framework can actually generalize quite well, with realistic sample sizes that can be much smaller than the dimensionality of this joint distribution, provided that we are looking for a reasonably *simple* representation T of our data. In fact, it is exactly the reliance of the framework on explicit manipulation of the joint distribution that allows us to derive non-uniform bounds that are adaptive to the complexity of the specific model chosen. In addition, we present a preliminary analysis regarding the question in which settings the information bottleneck can be seen as a standard learning algorithm, trading off a risk-like term and a regularization term controlling the generalization. Finally, we discuss its utility as a natural extension of the concept of minimal sufficient statistics for discrimination.

The paper is organized as follows. In Sec. 2, we formally present the information bottleneck framework and the notation used in the paper. We then turn to analyze its finite sample behavior in Sec. 3. Sec. 4 discusses the characteristics of the information bottleneck as a learning algorithm, while its relation to minimal sufficient statistics is considered in Sec. 5. Selected proofs are presented in Sec. 6; Full proofs can be found in [19]. We finish with a discussion in Sec. 7.

2 The Information Bottleneck Framework

In this section we explain and formally describe the basic information bottleneck (IB) framework. This framework has several variants and extensions, both to multivariate variables and to continuous representations (see [20, 4] for more details), but these are not the focus of this paper.

The IB framework attempts to find a simple representation of one random variable X through an auxiliary variable T, which is relevant to another random variable Y. Let us first exemplify how the IB method can be used for both supervised and unsupervised learning. Consider the area of text analysis. A typical unsupervised problem can be clustering documents based on their word-statistics in order to discover similarities and relationships between them. In this case the X variable is taken as the document identity (typically considered as "bags of words") and the Y as the words in the documents. In this case, the T variable will be clusters of documents with similar word-statistics, based, for instance, on the "the two sample problem" [13] similarity measure.

In a typical supervised application in this domain, X can denote the words while Y are topic-labels of the documents. Here T are clusters of words that are (approximately) sufficient for document categorization [24]. In all the applications a variable β allows us to smoothly move between a low resolution - highly compressed - solution, to a solution with higher resolution and more information about Y. This form of dimensionality reduction, a special case of the information bottleneck, was introduced under the name of distributional clustering in [16], and has proven to be quite effective in analyzing high dimensional data [2, 9].

In this work, we assume that X and Y take values in the finite sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, and use x and y respectively to denote elements of these sets. The basic quantity that is utilized in the IB framework is Shannon's mutual information between random variables, which for discrete variables is formally defined as:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right).$$

Mutual information is well known to be the unique measure of informativeness, up to a multiplicative constant, under very mild assumptions [5]. The IB functional is built upon the relationship between minimal sufficiency and information. It captures a tradeoff between minimality of the representation of X, achieved by minimizing $I(X;T)$, and sufficiency of information on Y, achieved by constraining the value of $I(Y;T)$. The auxiliary variable T is thus determined by the minimization of the IB-Lagrangian

$$\mathcal{L}_{IB}[p(t|x)] = I(X;T) - \beta I(Y;T) \tag{1}$$

with respect to the mapping $p(t|x)$. T is subject to the Markovian relation $T - X - Y$, and $p(t|x)$ is subject to the obvious normalization constraints. The tradeoff parameter β is a positive Lagrange multiplier associated with the constraint on $I(Y;T)$. Formally, T is defined over some space $\mathcal{T}$, but the elements of this space are arbitrary - only the probabilistic relationships between T and X, Y are relevant.

The solutions of this constrained optimization problem are characterized by the *bottleneck equations*,

$$\begin{cases} p(t|x) = \frac{p(t)}{Z(\beta,x)}\exp(-\beta\,\mathrm{D_{KL}}[p(y|x)\|p(y|t)]) \\ p(t) \quad = \sum_{x\in\mathcal{X}} p(t|x)p(x) \\ p(y|t) = \sum_{x\in\mathcal{X}} p(y|x)p(x|t)\ , \end{cases} \tag{2}$$

where $\mathrm{D_{KL}}$ is the Kullback-Leibler divergence and $Z(\beta, x)$ is a normalization function. These equations need to be satisfied simultaneously, given $p(x, y)$ and β. In [23] it is shown that alternating iterations of these equations converge - at least locally - to a solution for any initial $p(t|x)$, similar to the Arimoto-Blahut algorithm in information theory [5]. In [3] it is shown that the set of achievable $p(x, y, t)$ distributions form a strictly convex set in the $(I(X;T), I(Y;T))$ plane, bounded by a smooth optimal function - *the information curve* - similar to the rate-distortion function in source coding. By increasing the value of β one can move smoothly along this curve from the trivial, $I(X;T) = I(Y;T) = 0$ solution at the origin, all the way to the most complex solution where T captures all the relevant information from X and $I(X;T) = H(X)$, $H(X)$ denoting the entropy of X. In addition, as β is increased, $I(Y;T)$ increases and T captures more information on Y. Due to the data-processing inequality, $I(Y;T) \leq I(X;Y)$, with equality only when T becomes an exact sufficient statistic for Y. The trade-off inherent in Eq. (1) forces us to find a simple representation T of X, which preserves only those aspects of X which are informative, i.e. relevant, about Y.

It should be emphasized that despite superficial similarities, IB is *not* a hidden variable model. In such models, we assume that the joint distribution $p(x, y)$ can be factorized using an auxiliary random variable T, forming a Markovian relation $X - T - Y$. In IB, we make no generative assumption on the distribution, and the Markovian relation is $T - X - Y$. Namely, T is a generic compression of X, and the information-curve is characterized by the joint distribution $p(x, y)$ independently of any modeling assumptions.

An important observation is that the effective cardinality of an optimal T is not fixed and depends on β. When $\beta \leq 1$, even a trivial T of cardinality 1 will optimize Eq. (1), since we always have $I(Y;T) \leq I(X;T)$. On the other hand, as β increases, more emphasis is put on informativeness with respect to Y, and the cardinality of T will increase, although the cardinality of an optimal T need not exceed the cardinality of X, as proven in [10].

In order to optimize Eq. (1) we need to calculate the quantities $I(X;T)$ and $I(Y;T)$ for any chosen T and β. Since T is defined only via X, we need to know $p(x, y)$ in order to calculate these two quantities. In most applications, however, $p(x, y)$ is unknown. Instead, we assume that we have an i.i.d sample of m instances drawn according to $p(x, y)$, and we use this sample to create a maximum-likelihood estimate of the distribution using $\hat{p}(x, y)$, the empirical distribution of the sample. Following current practice, this empirical estimate is then plugged into the calculation of $I(X;T)$ and $I(Y;T)$ instead of the true joint distribution, and Eq. (1) is optimized using this plug-in estimate. In general, we use the $\hat{}$ symbol to denote quantities calculated using $\hat{p}(x, y)$ instead of $p(x, y)$. Thus, instead of calculating $I(X;T)$ and $I(Y;T)$ precisely, we rely on the empirical estimates $\hat{I}(X;T)$ and $\hat{I}(Y;T)$ respectively. In this work we investigate how much these empirical estimates can deviate from the true values when we optimize for T - in other words, whether this plug-in practice is justified. Note that the sample size m is often smaller than the number of bins $|\mathcal{X}||\mathcal{Y}|$, and thus $\hat{p}(x, y)$ can be a very poor approximation to $p(x, y)$. Nevertheless, this is precisely the regime we are interested in for many applications, text categorization to name one.

3 Finite Sample Analysis

We begin our analysis by focusing on the finite-sample behavior of the IB framework, and in particular on the relationship between $I(X;T)$ and $I(Y;T)$ that appear in Eq. (1) and their empirical estimates $\hat{I}(X;T)$ and $\hat{I}(Y;T)$.

Our first result shows that for any *fixed* T defined as a random mapping of X via $p(t|x)$, it is possible to determine the value of the objective function Eq. (1) within reasonable accuracy based on a random sample. The proof outline is provided in Sec. 6.1. The full proof can be found in [19].

Theorem 1. *Let T be a given probabilistic function of X into an arbitrary finite target space, determined by $p(t|x)$, and let $\mathcal{S}$ be a sample of size m drawn from the joint probability distribution $p(X,Y)$. For any confidence parameter $\delta \in (0,1)$, it holds with a probability of at least $1-\delta$ over the sample $\mathcal{S}$ that*

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

and that

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

Note that the theorem holds for any fixed T, not just ones which optimize Eq. (1). In particular, the theorem holds for any T found by an IB algorithm, even if T is not a globally optimal solution.

The theorem shows that estimating the objective function for a certain solution T is much easier than estimating $p(x,y)$. Indeed, the bound does not depend on $|\mathcal{X}|$, which might even be countably infinite. In addition, it depends on $|\mathcal{Y}|$ only as a second-order factor, since $|\mathcal{Y}|$ is multiplied by $1/m$ rather than by $1/\sqrt{m}$. The complexity of the bound is thus mainly controlled by $|\mathcal{T}|$. By constraining $|\mathcal{T}|$ to be small, or by setting β in Eq. (1) to be small enough so that the optimal T has low cardinality, a tight bound can be achieved.

Thm. 1 provides us with a bound on a certain pre-specified T, where the sample $\mathcal{S}$ is not part of the process of selecting T. The next theorem is a full generalization bound, determined by the sample when it is used as a training set by which T is selected.

For presenting the theorem compactly, we will use some extra notation. Let $x_1, \ldots, x_{|\mathcal{X}|}$ be some fixed ordering of the elements of $\mathcal{X}$, and $y_1, \ldots, y_{|\mathcal{Y}|}$ be an ordering of the elements of $\mathcal{Y}$. We use the shorthand $\mathbf{p}(T=t|x)$ to denote the vector $(p(t|x_1), \ldots, p(t|x_{|\mathcal{X}|}))$. Similarly, we denote the vector $(\hat{H}(T|y_1), \ldots, \hat{H}(T|y_{|\mathcal{Y}|}))$ by $\hat{\mathbf{H}}(T|y)$ where $\hat{H}(T|y_i)$ is the entropy of $\hat{p}(T|y_i)$. Furthermore, the vector $(H(T|x_1), \ldots, H(T|x_{\mathcal{X}}))$ is denoted by $\mathbf{H}(T|x)$, where $H(T|x_i)$ is the entropy of $p(T|x_i)$. Note that $p(T|x_i)$ is known as it defines T, and thus does not need to be estimated empirically.

For any real-valued vector $\mathbf{a} = (a_1, \ldots, a_n)$, we define the function $V(\mathbf{a})$ as follows:

$$V(\mathbf{a}) = \|\mathbf{a} - \frac{1}{n}\sum_{j=1}^{n} a_j\|^2 \triangleq \sum_{i=1}^{n}\left(a_i - \frac{1}{n}\sum_{j=1}^{n} a_j\right)^2 . \tag{3}$$

Note that $\frac{1}{n}V(\mathbf{a})$ is simply the variance of the elements of $\mathbf{a}$. In addition, we define the real-valued function ϕ as follows:

$$\phi(x) = \begin{cases} 0 & x = 0 \\ x\log(1/x) & 0 < x \leq 1/e \\ 1/e & x > 1/e. \end{cases} \tag{4}$$

Note that ϕ is a continuous, monotonically increasing and concave function.

Theorem 2. *Let $\mathcal{S}$ be a sample of size m drawn from the joint probability distribution $p(X, Y)$. For any confidence parameter $\delta \in (0, 1)$, it holds with a probability of at least $1 - \delta$ over the sample $\mathcal{S}$ that for all T, $|I(X;T) - \hat{I}(X;T)|$ is upper bounded by*

$$\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\cdot V(\mathbf{H}(T|x))}{m}} + \sum_t \phi\left(\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\cdot V(\mathbf{p}(T = t|x))}{m}}\right), \tag{5}$$

and $|I(Y;T) - \hat{I}(Y;T)|$ is upper bounded by

$$\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\cdot V(\hat{\mathbf{H}}(T|y))}{m}} + 2\sum_t \phi\left(\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)\cdot V(\mathbf{p}(T = t|x))}{m}}\right), \tag{6}$$

where V and ϕ are defined in Eq. (3) and Eq. (4), and C is a small constant.

As in Thm. 1, this theorem holds for all T, not just those optimizing Eq. (1). Also, the bound enjoys the advantage of not being uniform over a hypothesis class of possible T's, but rather depending directly on the T of interest. This is achieved by avoiding standard uniform complexity tools (see the proof for further details).

Intuitively, these bounds tell us that the 'smoother' T is with respect to X, the tighter the bound. To see this, assume that for any fixed $t \in \mathcal{T}$, $p(t|x)$ is more or less the same for any choice of x. By definition, this means that $V(\mathbf{p}(T = t|x))$ is close to zero. In a similar manner, if $H(T|x)$ is more or less the same for any x, then $V(\mathbf{H}(T|x))$ is close to zero, and so is $V(\hat{\mathbf{H}}(T|y))$ if $\hat{H}(T|y)$ is more or less the same for any y. In the extreme case, if T is independent of X, then $p(t|x) = p(t)$, $H(T|x) = H(T)$ and $\hat{H}(T|y) = \hat{H}(T)$ for any choice of x, y, and the generalization bound becomes zero. This is not too surprising, since in this case $I(X;T) = I(\hat{X;T}) = 0$ and $I(Y;T) = \hat{I}(Y;T) = 0$ regardless of $p(x, y)$ or its empirical estimate $\hat{p}(x, y)$.

This theorem thus suggests that generalization becomes better as T becomes less statistically dependent on X, and so provides a more compressed probabilistic representation of X. This is exactly in line with empirical findings [20], and with the intuition that 'simpler' models should lead to better generalization.

A looser but simpler bound on Thm. 2 can be achieved by fixing the cardinality of T, and analyzing the bound with worst-case assumptions on the statistical dependency between X and T. The proof, which is rather technical, is omitted in this version and may be found in [19].

Theorem 3. *Under the conditions and notation of Thm. 2, we have that with a probability of at least $1-\delta$, for all T,*

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

and

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

where C is the same constant as in Thm. 1.

Even with this much looser bound, if $|\mathcal{Y}|$ is large and $|\mathcal{T}| \ll |\mathcal{Y}|$ the bound can be quite tight, even with sample sizes which are in general insufficient to reasonably estimate the joint distribution $p(x,y)$. One relevant setting is in unsupervised learning, when Y models the feature space.

In this section, we have shown that the quantities that make up the IB objective function can be estimated reliably from a sample of a reasonable size, depending on the characteristics of T. In the next section we investigate the motivation for using these quantities in the objective function in the first place.

4 A Learning Theoretic Perspective

The IB framework optimizes a trade-off between $I(X;T)$ and $I(Y;T)$. In this section we provide a preliminary discussion of the learning theoretic properties of this tradeoff, investigating when mutual information provides reasonable measures for both learning complexity and accuracy.

In an unsupervised setting, such as clustering, it is rather easy to see how $I(X;T)$ and $I(Y;T)$ control the complexity and granularity of the clustering by trading between homogeneity and resolution of the clusters; this has been discussed previously in the literature (such as [24], [3]). Therefore, we will focus here mainly on the use of this framework in supervised learning, where the objectives are more well defined.

Most supervised learning algorithms are based on a tradeoff between two quantities: a risk term, measuring the performance of a hypothesis on the sample data, and a regularization term, which penalizes complex hypotheses and so ensures reasonable generalization to unseen data. In the following we argue that under relevant settings it is reasonable to consider $I(Y;T)$ as a measure of risk and $I(X;T)$ as a regularization term that controls generalization.

4.1 I(Y;T) as a Measure of Performance

In this section we investigate the plausibility of $I(Y;T)$ as a measure of performance or risk in a supervised learning setting. We show that in those supervised learning settings where IB was demonstrated to be highly effective, such as document categorization [22], there is a strong connection between the classification error and the mutual information $I(Y;T)$, especially when the categories are uniformly spread. The discussion here is a first step towards a full analysis of the IB classification performance in a more general setting, which we leave for future work.

In a typical document classification task we model X as a random variable over the set of possible words, and Y as a random variable over the set of document categories or classes. Each document is treated as an i.i.d. sample of words drawn from $p(x|y)$, in accordance with the bag of words representation, where y is the class of the document. Unlike the simple supervised learning settings, where each example is described as a single data point, in this case each example (document) to be labeled is described by a sample of points (words) of variable size (usually large) and we seek the most probable class of the whole sample (document) *collectively.*

IB is used in this setting to find T, a compressed representation of the words in a document, which is as informative as possible on the categories Y. The bottleneck equations Eq. (2) provide for each class y its conditional distribution on T, via $\hat{p}(t|y) = \sum_x p(t|x)\hat{p}(x|y)$. When a new document $D = \{x_1, \ldots, x_n\}$ of size n is to be classified, the empirical distribution of T given D is $\tilde{p}(t) = \sum_{i=1}^n p(t|x_i)\hat{p}(x_i)$. Assuming that the document is sampled according to $p(t|y)$ for some class y, the most probable class y^* can be selected using the maximum likelihood principle, namely $y^* = \operatorname{argmin}_y \mathrm{D}_{\mathrm{KL}}[\tilde{p}(t)\|\hat{p}(t|y)]$.

We now show that $\hat{I}(Y;T)$ is indeed a reasonable objective function whenever we wish to collectively label an entire set of sampled instances.

Assume that the true class for document D is y_1, with its word distribution sampled via $p(t|y_1)$. The probability α_n of misclassifying this sample as y_2 for some $y_2 \neq y_1$ via the likelihood test decreases exponentially with the sample size n. The rate of exponential decrease is larger if the two distributions $p(t|y_1), p(t|y_2)$ are more distinct. Formally, by Stein's lemma [5], if $\hat{p}(t|y_1) = p(t|y_1)$ and $\hat{p}(t|y_2) = p(t|y_2)$, then

$$\lim_{n\to\infty} \frac{1}{n} \log(\alpha_n) = \mathrm{D}_{\mathrm{KL}}[p(t|y_2)\|p(t|y_1)]. \tag{7}$$

When $\hat{p}(t|y_1)$ and $\hat{p}(t|y_2)$ deviate from the true conditional distributions, Stein's Lemma still holds up to an additive constant which depends on the amount of deviation, and the exponent is still controlled mainly by $\mathrm{D}_{\mathrm{KL}}[p(t|y_2)\|p(t|y_1)]$. In the following we will assume for simplicity that Eq. (7) holds exactly.

The overall probability of misclassifying a document when there are more than two possible classes is thus upper bounded by

$$\sum_{y\neq y_1} \exp(-n\mathrm{D}_{\mathrm{KL}}[p(t|y)\|p(t|y_1)]). \tag{8}$$

On the other hand, by the definition of mutual information and the convexity of the Kullback-Leibler divergence we have that

$$I(Y;T) = \mathbb{E}_y \mathrm{D_{KL}}[p(t|y) \| p(t)] = \mathbb{E}_y \mathrm{D_{KL}}[p(t|y) \| \mathbb{E}_{y'} p(t|y')] \\ \leq \mathbb{E}_{y,y'} \mathrm{D_{KL}}[p(t|y) \| p(t|y')], \tag{9}$$

Hence $-nI(Y;T)$ is an upper bound on the expected value of the exponent in Eq. (7), assuming that y_1 and y_2 are picked according to $p(y)$. The relationship between Eq. (9) on the one hand, and Eq. (7), Eq. (8) on the other hand, is not direct. Nonetheless, these equations indicate that if the examples to classify are represented by a large sample, as in the document classification setting, higher values of $I(Y;T)$ should correspond to a reduced probability of misclassification. For example, if $\mathrm{D_{KL}}[p(t|y) \| p(t|y_1)]$ is equal for every $y \neq y_1$, we have that Eq. (8) is upper bounded by

$$(n-1) \exp\big(-nI(Y;T) / (|\mathcal{Y}| - 1) \big),$$

in which case the probability of misclassification is exponentially dominated by $I(Y;T)$. This is the case when categories are uniformly spread, which happens for many applications incidently or by design. In this case, when the bottleneck variable T captures just a fraction $\alpha = I(Y;T)/I(X;Y)$ of the relevant information, the test (document) size should increase only by a factor $1/\alpha$ in order to achieve a similar bound on the classification error.

4.2 I(X;T) as a Regularization Term

In this subsection we discuss the role of $I(X;T)$, the compression term in IB, as a regularizer when maximizing $I(Y;T)$. Note that without regularization, $I(Y;T)$ can be maximized by setting $T = X$. However, $p(x|y)$ cannot be estimated efficiently from a sample of a reasonable size; therefore the formal solution $T = X$ cannot be used to perform reliable classification. Moreover, in the context of unsupervised learning, setting $T = X$ is generally a meaningless operation, corresponding to singleton clusters.

The bottleneck variable T must therefore be restricted to allow reasonable generalization in a supervised setting and to generate a reasonable model in an unsupervised setting. In the IB framework $I(X;T)$ can be viewed as a penalty term that restricts the complexity of T. A more formal justification for this is given in the following theorem, which is derived from Thm. 2. Since the proof is quite technical, it is omitted in this version and may be found in [19].

Theorem 4. *For any probability distribution $p(x,y)$, with a probability of at least $1-\delta$ over the draw of the sample of size m from $p(x,y)$, we have that for all T,*

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

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

This bound is controlled by $I(X;T)$ and $\hat{I}(X;T)$, which are closely related as Thm. 3 shows. This is not a fully empirical bound, as it depends on the unknown quantity $I(X;T)$ and the marginal distributions of X, Y. The bound does however illustrate the relationship between the generalization error, as embodied in the difference between $I(Y;T)$ and $\hat{I}(Y;T)$, and the mutual information $I(X;T)$. This provides motivation for the use of $I(X;T)$ as a regularization term, beyond its obvious description length interpretation or coding interpretation.

5 Relationship with Sufficient Statistics

A fundamental issue in statistics, pattern recognition, and machine learning is the notion of relevance. Finding the relevant components of data is implicitly behind the problems of efficient data representation, feature selection and dimension reduction for supervised learning, and is the essence of most unsupervised learning problems. One of the earliest and more principled approaches to relevance was the concept of *sufficient statistics* for parametric distributions, introduced by Fisher [7] as function(s) of a sample that capture all the information about the parameter(s). A *sufficient statistic* is defined as follows:

Definition 1 (Sufficient Statistic). *Let Y be a parameter indexing a family of probability distributions. Let X be random variable drawn from a probability distribution determined by Y. Let T be a deterministic function of X. T is sufficient for Y if*

$$\forall x \in \mathcal{X}, t \in \mathcal{T}, y \in \mathcal{Y} \quad p(x|t,y) = p(x|t).$$

Throughout this section we assume that it suffices that the equality holds almost everywhere with respect to the probability of y and x.

In words, the sufficiency of T means that given the value of T, the distribution of X does not depend on the value of Y.

In the parametric statistics setting, Y is a random variable that parameterizes a family of probability distributions, and X is a data point drawn from $p(x|y)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. For example, the family of probability distributions may be the set of Bernoulli distributions with success probability p determined by y, with $\mathcal{Y} \subseteq [0,1]$ and some prior distribution $p(y)$. In this case, for a given y, $p(X=1|y) = y$, and $p(X=0|y) = 1-y$.

Y and X may be high dimensional. For instance, Y may determine the mean and the variance of a normal distribution, or fully parameterize a multinomial distribution. X may be a high dimensional data point. For any family of probability distributions, we can consider a sample of m i.i.d data points, all drawn from the same distribution determined by a single draw of Y. In the context of sufficient statistics, this is just a special case of a high dimensional X which is drawn from the cross-product of m identical probability distributions determined by the value of Y.

Just as X and Y may be high dimensional, so can T map X to a multidimensional space. If X denotes an i.i.d sample, the number of dimensions in T may depend on the size of the sample m. Specifically, $T = X$ is always sufficient for Y. To avoid trivial sufficient statistics such as this, Lehmann and Scheffé [12] introduced the concept of a minimal sufficient statistic, which denotes the coarsest sufficient partition of X, as follows:

Definition 2 (Minimal Sufficient Statistic). *A sufficient statistic S is minimal if and only if for any sufficient statistic T, there exists a deterministic function f such that $S = f(T)$ almost everywhere w.r.t X.*

For instance, for an i.i.d sample of size m of the Bernoulli distribution in the example above, $T = X$ is trivially a sufficient statistic, but the one-dimensional $T = \frac{1}{m}\sum_i x_i$ where $x = (x_1, \ldots x_m)$ is also sufficient. It can be shown that the latter T (and any one-to-one function of it) is a minimal sufficient statistic.

By the Pitman-Koopman-Darmois theorem [17], sufficient statistics whose dimension does not depend on the sample size exist only for families of exponential form. This makes the original concept of sufficiency rather restricted.

Kullback and Leibler [11] related sufficiency to Shannon's information theory, showing that sufficiency is equivalent to preserving mutual information on the parameter, while minimal sufficient statistics minimize the mutual information with the sample due to the data-processing inequality [5].

The IB framework allows us to naturally extend this concept of relevance to any joint distribution of X and Y, not necessarily ones of exponential form, in a constructive computational manner. In this framework, built on Kullback's information theoretic characterization of sufficiency [11], one can find compact representations T of a sample X that maximize mutual information about the parameter variable Y, corresponding to sufficiency for Y, and minimize $I(X;T)$, corresponding to the minimality of the statistic. However, unlike the original concepts of sufficient statistic and minimal sufficient statistic, the IB framework provides a soft tradeoff between these two objectives.

It can easily be seen that as β grows to infinity, if T is not restricted then $I(Y;T)$ converges to $I(X;Y)$ and T converges to a minimal sufficient statistic. The following theorem formalizes this insight. Similar formulations of this theorem can be gleaned from [11] and [5]. The full proof is presented for completeness in [19].

Theorem 5. *Let X be a sample drawn according to a distribution determined by the random variable Y. The set of solutions to*

$$\min_{T} \quad I(X;T) \quad s.t. \quad I(Y;T) = \max_{T'} I(Y;T')$$

is exactly the set of minimal sufficient statistics for Y based on the sample X.

The IB framework thus provides a natural generalization of the concept of a sufficient statistic, where by setting β to lower values, different degrees of approximate minimal sufficient statistics can be found, characterized by the fraction of mutual information they maintain on the Y. Furthermore, such approximate minimal sufficient statistics exist for any joint distribution $p(X, Y)$ in a continuous hierarchy that is fully captured by the set of optimal IB solutions for all values of β. These solutions lie on the information curve of the distribution.

6 Proofs

6.1 Proof of Thm. 1

Let $\mathcal{S}$ be a sample of size m, and let T be a probabilistic function of X into an arbitrary finite target space, defined by $p(t|x)$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$.

To prove the theorem, we bound the deviations of the information estimations from their expectation: $|\hat{I}(X;T) - \mathbb{E}(\hat{I}(X;T))|$ and $|\hat{I}(Y;T) - \mathbb{E}(\hat{I}(Y;T))|$, and then use a bound on the expected bias of entropy estimation.

To bound the deviation of the information estimates, we use McDiarmid's inequality [14], in a manner similar to [1]. For this we must bound the change in value of each of the entropy estimates when a single instance in $\mathcal{S}$ is arbitrarily changed. A useful and easily proven inequality in that regard is the following: for any natural m and for any $a \in [0, 1 - 1/m]$ and $\Delta \leq 1/m$,

$$\Big|(a + \Delta)\log(a + \Delta) - a\log(a)\Big| \leq \frac{\log(m)}{m}. \tag{10}$$

With this inequality, a careful application of McDiarmid's inequality leads to the following lemma. The proof of the lemma can be found in [19].

Lemma 1. *For any $\delta_1 > 0$, with probability of at least $1 - \delta_1$ over the sample, we have that*

$$|\hat{I}(X;T) - \mathbb{E}[\hat{I}(X;T)] \leq \frac{(|\mathcal{T}|\log(m) + \log(|\mathcal{T}|))\sqrt{\log(2/\delta_1)}}{\sqrt{2m}}. \tag{11}$$

Similarly, with a probability of at least $1 - \delta_2$,

$$|\hat{I}(Y;T) - \mathbb{E}[\hat{I}(Y;T)]| \leq \frac{(3|\mathcal{T}| + 2)\log(m)\sqrt{\log(2/\delta_2)}}{\sqrt{2m}}. \tag{12}$$

Lemma 1 provides bounds on the deviation of the $\hat{I}(X;T), \hat{I}(Y;T)$ from their expected values. In order to relate these to the true values of the mutual information $I(X;T)$ and $I(Y;T)$, we use the following bias bound from [15].

Lemma 2 (Paninski, 2003). *For a random variable X, with the plug-in estimate $\hat{H}(\cdot)$ on its entropy, based on an i.i.d sample of size m, we have that*

$$|\mathbb{E}[\hat{H}(X) - H(X)]| \leq \log\left(1 + \frac{|\mathcal{X}| - 1}{m}\right) \leq \frac{|\mathcal{X}| - 1}{m}.$$

From Lemma 2, we get that the quantities $|\mathbb{E}[H(T)-H(T)]|$, $|\mathbb{E}[H(Y)-H(Y)]|$, and $|\mathbb{E}[H(Y,T)-H(Y,T)]|$ are upper bounded by $(|\mathcal{T}|-1)/m$, $(|\mathcal{Y}|-1)/m$ and $(|\mathcal{Y}||\mathcal{T}|-1)/m$ respectively. Combining these with Eq. (11) and Eq. (12), and setting $\delta_1=\delta_2=\delta/2$, we get the bounds in Thm. 1.

6.2 Proof of Thm. 2

The idea of the proof is as follows. We bound the quantities $|I(X;T)-\hat{I}(X;T)|$ and $|I(Y;T)-\hat{I}(Y;T)|$ with deterministic bounds that depend on the empirical distribution and on the true underlying distribution. These bounds are factorized, in the sense that quantities that depend on the empirical sample are separated from quantities that depend on the characteristics of T. Quantities of the first type can be bounded by concentration of measure theorems, while quantities of the second type can be left dependent on the T we choose.

The deterministic bounds are summarized in the following lemma. The proof of this lemma is purely technical, and may be found in [19].

Lemma 3. *The following two inequalities hold:*

$$\begin{aligned}|I(X;T)-\hat{I}(X;T)| \quad &\leq \quad \sum_t \|\mathbf{p}(x)-\hat{\mathbf{p}}(x)\| \cdot \phi\left(\sqrt{V(\mathbf{p}(T=t|x))}\right) \\ &\quad + \|\mathbf{p}(x)-\hat{\mathbf{p}}(x)\| \cdot \sqrt{V(\mathbf{H}(T|x))},\end{aligned} \tag{13}$$

$$\begin{aligned}|I(Y;T)-\hat{I}(Y;T)| \quad &\leq \quad \sum_t \|\mathbf{p}(x)-\hat{\mathbf{p}}(x)\| \cdot \phi\left(\sqrt{V(\mathbf{p}(T=t|x))}\right)) \\ &\quad + \sum_y p(y) \sum_t \phi\left(\|\hat{\mathbf{p}}(x|y)-\mathbf{p}(x|y)\| \cdot \sqrt{V(\mathbf{p}(T=t|x))}\right) \\ &\quad + \|\mathbf{p}(y)-\hat{\mathbf{p}}(y)\| \cdot \sqrt{V(\hat{\mathbf{H}}(T|y))}.\end{aligned} \tag{14}$$

In order to transform the bounds in Eq. (13) and Eq. (14) to bounds that do not depend on $p(x)$, we can use concentration of measure arguments on L_2 norms of random vectors, such as the following one based on an argument in section 4.1 of [6]: Let ρ be a distribution vector of arbitrary (possible countably infinite) cardinality, and let $\hat{\rho}$ be an empirical estimation of ρ based on a sample of size m. Then with a probability of at least $1-\delta$ over the samples,

$$\|\rho-\hat{\rho}\|_2 \leq \frac{2+\sqrt{2\log(1/\delta)}}{\sqrt{m}}. \tag{15}$$

We apply this concentration bound to $\|\mathbf{p}(x)-\hat{\mathbf{p}}(x)\|$, $\|\mathbf{p}(y)-\hat{\mathbf{p}}(y)\|$, and to $\|\hat{\mathbf{p}}(x|y)-\mathbf{p}(x|y)\|$ for any y in Eq. (13) and Eq. (14). To make sure the bounds hold simultaneously over these $|\mathcal{Y}|+2$ quantities, we replace δ in Eq. (15) by $\delta/(|\mathcal{Y}|+2)$. Note that the union bound is taken with respect to the marginal distributions of $\hat{\mathbf{p}}(x)$, $\hat{\mathbf{p}}(y)$ and $\hat{\mathbf{p}}(x|y)$, which do not depend on the T chosen. Thus, the following bounds hold with a probability of $1-\delta$, for all T:

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

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

To get the bounds in Thm. 2, we note that

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

where C is a small constant.

It is interesting to note that these bounds still hold in certain cases even if $\mathcal{X}$ is infinite. Specifically, suppose that for all $t \in \mathcal{T}$, $p(t|x)$ is some constant c_t for all but a finite number of elements of $\mathcal{X}$. If the definition of $V(\cdot)$ is replaced with $V(\mathbf{p}(T=t|x)) = \sum_x (p(T=t|x) - c_t)^2$, Then $V(\mathbf{p}(T=t|x))$ is finite and the proof above remains valid. Therefore, under these restrictive assumptions the bound is valid and meaningful even though $\mathcal{X}$ is infinite.

7 Discussion

In this paper we analyzed the information bottleneck framework from a learning theoretic perspective. This framework has been used successfully for finding efficient relevant data representations in various applications, but this is its first rigorous learning theoretic analysis. Despite the fact that the information bottleneck is all about manipulating the joint input-output distribution, we show that it can generalize quite well based on plug-in empirical estimates, even with sample sizes much smaller than needed for reliable estimation of the joint distribution. In fact, it is exactly the reliance on the joint distribution that allows us to derive non-uniform and adaptive bounds.

Moreover, these bounds allow us to view the information bottleneck framework in the more familiar learning theoretic setting of a performance-complexity tradeoff. In particular, we provided a preliminary analysis of the role of mutual information as both a complexity regularization term and as a bound on the classification error for common supervised applications, such as document classification. This is the first step in providing a theoretical justification for many applications of interest, including a characterization of the learning scenarios for which this method is best suited. Finally, we showed how this framework extends the classical statistical concept of minimal sufficient statistics.

References

1. A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19(3–4):163–193, 2001.
2. L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. *Proceedings of SIGIR 1998*, pages 96–103.
3. R. Gilad-Bachrach, A. Navot, and N. Tishby. An Information Theoretic Tradeoff between Complexity and Accuracy. In *Proceedings of COLT 2003*, pages 595–609.
4. G. Chechik, A. Globerson, N. Tishby, Y. Weiss. Information Bottleneck for Gaussian Variables. In *Journal of Machine Learning Research*, 6:165–188, 2005.
5. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.
6. N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.
7. R.A. Fisher. On the Mathematical Foundation of Theoretical Statistics. *Philos. Trans. Roy. Soc., London*, Sec. A, 222:309–368, 1922.
8. N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate Information Bottleneck. In *Proceedings of UAI 2001*, pages 152–161.
9. M. Nikravesh I. Guyon, S. Gunn and L. A. Zadeh, editors. *Feature Extraction: Foundations and Applications.* Springer, 2006.
10. P. Harremoës and N. Tishby. The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. In *Proceedings of the IEEE Int. Symp. on Information Theory*, pages 566–571, 2007.
11. S. Kullback and R.A. Leibler. On Information and Sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.
12. E.L. Lehmann and H. Scheffé. Completeness, Similar Regions and Unbiased Estimation. *Sankhya*, 10:305–340, 1950.
13. E.L. Lehmann. Testing Statistical Hypotheses. Wiley, New-York, 1959.
14. C. McDiarmid. On the Method of Bounded Differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989.
15. L. Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.
16. F. C. Pereira, N. Tishby, and L. Lee. Distributional Clustering of English Words. In *Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.
17. B. Koopman. On Distributions Admitting a Sufficient Statistic. *Trans. Amer. math. Soc.*, 39:399–409, 1936.
18. R.T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.
19. O. Shamir, S. Sabato and N. Tishby. *Learning and Generalization with the Information Bottleneck.* Full version available at *www.cs.huji.ac.il/∼ohads03/ShamirSabatoTishbyALT2008_full.pdf*
20. N. Slonim. *The Information Bottleneck: Theory and Applications.* PhD thesis, Hebrew University, Jerusalem, 2003.
21. N. Slonim, G. Singh S. Atwal, G. Tkacik, and W. Bialek. Information-based Clustering. *Proc. Natl. Acad. Sci. U.S.A*, December 2005.
22. N. Slonim and N. Tishby. The Power of Word Clusters for Text Classification. In *23rd European Colloquium on Information Retrieval Research*, 2001.
23. N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. In *The 37'th Allerton Conference on Communication, Control, and Computing*, 1999.
24. N. Tishby and N. Slonim. Data clustering by Markovian relaxation and the information bottleneck method. In *Proceedings of NIPS 2000*, pages 640–646.