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, extending the classical notion of minimal sufficient statistics, that finds concise representations for an 'input' random variable that are as relevant as possible for an 'output' 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 non-uniform finite sample bounds that show that it can provide concise representations with good generalization based on smaller sample sizes than needed to estimate the underlying distribution. Based on these results, we can analyze 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

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 in 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 (Fisher, 1922) as function(s) of a sample that capture all the information about the parameter(s). The notion of *minimal sufficient statistics* was introduced by Lehmann and Scheffé (Lehmann and Scheffé, 1950) as the simplest sufficient statistics, or the coarsest *sufficient partition* of the sample space which captures the relevant components of the sample with respect to the parameter. However, this important concept was not pursued much further mainly due to the Pitman-Koopman-Darmois theorem, which showed that exact sufficient statistics with bounded dimensionality exist only for distributions of exponential form (Koompan, 1936).

Kullback and Leibler (Kullback and Leibler) 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 (Cover and Thomas, 1991). The Information Bottleneck (IB) method, introduced in (Tishby, Pereira and Bialek, 1999), is an information theoretic generalization of the minimal-sufficient-statistic concept to general distributions of two variables, X and Y. It also provides a converging algorithm for extracting minimal relevant components of the variable X with respect to the variable Y, by finding a non-parametric model-independent compression of X (providing minimality), denoted by T, that is most informative about Y (providing approximate sufficiency). The 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. Further details are presented in Sec. 2.

Before turning to the topic of this paper, 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. The T variable in this case will be clusters of documents with similar word-statistics, based on "the two sample problem" (Lehmann, 1959) 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 (Tishby and Slonim, 2000). In all the applications the 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 (Pereira, Tishby and Lee, 1993), and has proven to be highly effective in data analysis of high dimensional data (Baker and McCallum, 1998; **?**).

The method has proven to be useful for a number of successful applications (see (Tishby and Slonim, 2000; Friedman, Mosenzon, Slonim and Tishby, 2001; Slonim, Atwal, Tkacik and bialek, 2005) and references therein), but its learning theoretic justification has remained unclear for a number of reasons: (i) The method assumes the knowledge of the joint distribution of X and Y, in sharp contrast to the finite-sample based machine learning algorithms. Moreover, it wasn't clear what is left to be learned if it is assumed that this distribution is known. (ii) In practice, since the joint distribution of X and Y is not known, the empirical co-occurrence distribution is used to calculate a plug-in estimate of the IB functional, without finite-sample generalization bounds or error guarantees of any kind. (iii) Finally, IB is formally related to classical information theoretic problems, such as Rate-Distortion theory and Coding with Side-Information, but it is 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 most 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 non-uniform 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. We discuss 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 notations of 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. All the proofs of our main theorems are presented in Sec. 6, and we finish by discussing our results in Sec. 7.

2 The Information Bottleneck Framework

In this section we formally describe the basic information bottleneck (IB) framework. This framework has several variants and extensions, both to multivariate variables and to continuous representations (see (Slonim, 2003; Checik, Globerson, Tishby and Weiss, 2005) for more details), but these are not the focus of this paper.

As discussed in the introduction, 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. 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 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 (Cover and Thomas, 1991). 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*,

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

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 (Tishby, Pereira and Bialek, 1999) 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 (Cover and Thomas, 1991). In (Gilad-Bachrach, Navot and Tishby, 2001) 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 $I_Y(I_X)$ 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, zero information, solution at the origin, all the way to the most complex solution where T captures all the relevant information from X and $I_X \equiv I(X;T) = H(X)$, $H(X)$ denoting the entropy of X. In addition, as β is increased, $I_Y \equiv I(Y;T)$ increases and T captures more information on Y. Due to the data-processing inequality, $I(Y;T) \le I(X;Y)$, with equality only when T becomes an exact sufficient statistic. The tradeoff 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 (Harremoes and Tishby, 2007).

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 deviate from the true values - in other words, whether this plug-in practice 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 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.

Theorem 1. *Let T be a given random mapping of X, 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.

In order to present 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, $\hat{\mathbf{H}}(T|y)$ denotes the vector $(\hat{H}(T|y_1),\ldots,\hat{H}(T|y_{|\mathcal{Y}|}))$ where $\hat{H}(T|y_i)$ is the entropy of $\hat{p}(T|y_i)$. $\mathbf{H}(T|x)$ denotes the vector $(H(T|x_1),\ldots,H(T|x_{\mathcal{X}}))$, 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:

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

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 $\phi(x) : \mathbb{R}_+ \to \mathbb{R}_+$ as

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

Note that ϕ is a continuous, monotonically increasing and concave function, and that $\lim_{x\to 0}\phi(x) = 0$.

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 any T simultaneously,*

$$|I(X;T) - \hat{I}(X;T)| \leq \sqrt{\frac{C \log(|\mathcal{Y}|/\delta) \cdot V(\mathbf{H}(T|x))}{m}} \tag{5}$$

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

and

$$|I(Y;T) - \hat{I}(Y;T)| \leq \sqrt{\frac{C \log(|\mathcal{Y}|/\delta) \cdot V(\hat{\mathbf{H}}(T|y))}{m}} \tag{6}$$

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

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 any 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.

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 (Slonim, 2003), 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, with worst-case assumptions on the statistical dependency between X and T.

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

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

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, from a learning theoretic perspective.

4 A Learning Theoretic Perspective

The IB framework optimizes a trade-off between $I(X;T)$ and $I(Y;T)$. In this section we discuss the learning theoretic properties of this tradeoff and why mutual information provide 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 (Tishby and Slonim, 2000; **?**)). 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 (Slonim and Tishby, 2001), 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.

For example, a 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 in this case - namely, 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 (Cover and Thomas, 1991), 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

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

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 $\mathrm{D}_{\mathrm{KL}}[p(t|y)\|p(t|y_1)]$ is equal for every $y \neq y_1$, we have that Eq. (8) is upper bounded by

$$(n-1)\exp\left(-\frac{n}{|\mathcal{Y}|-1} I(Y;T)\right),$$

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 or minimality 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.

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 any T simultaneously,*

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

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 and the mutual information $I(X;T)$. This provides motivation for the use of $I(X;Y)$ as a regularization term, beyond its obvious description length, or coding, interpretation.

5 Relationship with Sufficient Statistics

As we discussed in the introduction, there is a natural relationship between the IB framework and the fundamental statistical concept of *minimal sufficient statistics*, which captures the notion of relevance in the context of parametric distributions. In this section we elaborate on this connection.

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.

Fisher (Fisher, 1922) introduced the concept of sufficient statistic that denotes the relevant part of the sample X with respect to the parameter Y, 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.

Just as X and Y may be high dimensional, so can T map X to a multidimensional space. If X denotes and 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 (Lehmann and Scheffé, 1950) 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 (Koompan, 1936), 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.

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 (Kullback and Leibler), 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.

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

$$\begin{aligned} \min_T \quad & I(X;T) \\ \textit{s.t.} \quad & I(Y;T) = \max_{T'} I(Y;T'). \end{aligned}$$

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 random mapping of X defined by $p(t|x)$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$.

To prove the theorem, we first 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 (McDiarmid, 1989), in a manner similar to (Antos and Kontoyiannis, 2001). 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.

We use the equality $\hat{I}(X;T) = \hat{H}(T) - \hat{H}(T|X)$. First, we bound the change caused by a single replacement in $\hat{H}(T)$. We have that

$$\hat{H}(T) = -\sum_t (\sum_x p(t|x)\hat{p}(x)) \log(\sum_x p(t|x)\hat{p}(x)). \tag{10}$$

If we change a single instance in $\mathcal{S}$, then there exist two pairs (x, y) and (x', y') such that $\hat{p}(x, y)$ increases by $1/m$, and $\hat{p}(x', y')$ decreases by $1/m$. This means that $\hat{p}(x)$ and $\hat{p}(x')$ also change by at most $1/m$, while all other values in the distribution remain the same. Therefore, for each $t \in \mathcal{T}$, $\sum_x p(t|x)\hat{p}(x)$ changes by at most $1/m$.

We use the following easily-proven inequality (a special case of Lemma 2 appearing later): for any natural m and for any $a \in [0, 1 - 1/m]$ and $\Delta \leq 1/m$,

$$\Big|(a + \Delta)\log(a + \Delta) - a\log(a)\Big| \leq \frac{\log(m)}{m}. \tag{11}$$

Based on this inequality and Eq. (10), we have that $\hat{H}(T)$ changes by at most $|\mathcal{T}|\log(m)/m$. We now move to bound the change in $\hat{H}(T|X)$. We have

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

$H(T|X = x)$ is dependent only on $p(t|x)$ which is known and does not depend on the sample. Changing a single instance in $\mathcal{S}$ changes $\hat{p}(x)$ by at most $1/m$ for two values x. Since $H(T|X = x) \leq \log(|\mathcal{T}|)$, this implies that $H(T|X)$ changes by at most $\log(|\mathcal{T}|)/m$. Overall, $\hat{I}(X;T) = \hat{H}(T) - \hat{H}(T|X)$ can change by at most $(|\mathcal{T}|\log(m)+\log(|\mathcal{T}|))/m$. Invoking McDiarmid's inequality, we have that with a probability of at least $1 - \delta_1$,

$$\begin{aligned}|\hat{H}(T) - \hat{H}(T|X)) - \mathbb{E}(\hat{H}(T) - \hat{H}(T|X)| \\ \leq \frac{(|\mathcal{T}|\log(m) + \log(|\mathcal{T}|))\sqrt{\log(2/\delta_1)}}{\sqrt{2m}}.\end{aligned} \tag{12}$$

We now turn to $\hat{I}(Y;T)$ and perform a similar analysis using the fact that $\hat{I}(Y;T) = \hat{H}(Y) + \hat{H}(T) - \hat{H}(Y,T)$. First, for $\hat{H}(Y)$, we have that

$$\hat{H}(Y) = -\sum_y \hat{p}(y)\log(\hat{p}(y)).$$

Changing a single instance in $\mathcal{S}$ changes $\hat{p}(y)$ by at most $1/m$ for two values y, hence by Eq. (11), $\hat{H}(Y)$ changes by at most $2\log(m)/m$. For $\hat{H}(Y,T)$, we have

$$\hat{H}(Y,T) = -\sum_{t,y} \hat{p}(t,y)\log(\hat{p}(t,y))$$

and

$$\hat{p}(y,t) = \sum_x p(t|x)\hat{p}(x,y)$$

Since $T - X - Y$ is a Markov chain, changing a single instance in $\mathcal{S}$ changes $\sum_x p(t|x)\hat{p}(x,y)$ by at most $1/m$ for two values y. Using Eq. (11), we have that $\hat{H}(Y,T)$ can change by at most $2|\mathcal{T}|\log(m)/m$. Finally, as we saw above, by replacing a single instance $\hat{H}(T)$ can change by at most $|\mathcal{T}|\log(m)/m$. Overall, we have that $\hat{I}(Y;T)$ can change by at most $(3|\mathcal{T}| + 2)\log(m)/m$. By McDiarmid's inequality, we have that with a probability of at least $1 - \delta_2$,

$$\begin{aligned}|(\hat{H}(Y)+\hat{H}(T)-\hat{H}(Y,T))-\mathbb{E}(\hat{H}(Y)+\hat{H}(T)-\hat{H}(Y,T))| \\ \leq \frac{(3|\mathcal{T}| + 2)\log(m)\sqrt{\log(2/\delta_2)}}{\sqrt{2m}}.\end{aligned} \tag{13}$$

Eq. (12) and Eq. (13) provide 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 (Paninski, 2003).

Lemma 1 (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 this lemma, we have that

$$|\mathbb{E}H(T) - H(T)| \leq \frac{|\mathcal{T}| - 1}{m},$$
$$|\mathbb{E}H(Y) - H(Y)| \leq \frac{|\mathcal{Y}| - 1}{m},$$
$$|\mathbb{E}H(Y,T) - H(Y,T)| \leq \frac{|\mathcal{Y}||\mathcal{T}| - 1}{m}.$$

Combining these with Eq. (12) and Eq. (13), and setting $\delta_1 = \delta_2 = \delta/2$, we get the bounds in Thm. 1.

6.2 Proof of Thm. 2

We start the proof by bounding $|I(X;T) - \hat{I}(X;T)|$ and $|I(Y;T) - \hat{I}(Y;T)|$ with deterministic bounds that depend on $p(x)$. These bounds are then factorized, such 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.

Starting with $|I(X;T) - \hat{I}(X;T)|$, we use the fact that

$$|I(X;T)-\hat{I}(X;T)| \leq |H(T|X)-\hat{H}(T|X)|+|H(T)-\hat{H}(T)|$$

and bound each of the summands on the right separately. For the first summand, since $\sum_x p(x) = \sum_x \hat{p}(x) = 1$, we have that for any scalar a,

$$\begin{aligned}|H(T|X) - \hat{H}(T|X)| &= \Big|\sum_x (p(x) - \hat{p}(x))H(T|x)\Big| \\ &= \Big|\sum_x (p(x) - \hat{p}(x))(H(T|x) - a)\Big| \\ &\leq \|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\|\|\mathbf{H}(T|x) - a\|,\end{aligned} \tag{14}$$

where $\mathbf{p}$ and $\mathbf{H}$ stand for vectors indexed by the values of X, and we subtract a from all entries of the vector. Setting $a = \frac{1}{|\mathcal{X}|}\sum_x H(T|x)$ we get

$$\begin{aligned} &|H(T|X) - \hat{H}(T|X)| \\ &\leq \|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\| \sqrt{V(\mathbf{H}(T|x))}, \end{aligned} \tag{15}$$

Where $V(\cdot)$ is defined in Eq. (3).

We now turn to bound the second summand. For the rest of the proof, we use the following easily proven lemma.

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

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

where $\phi(\cdot)$ is defined in Eq. (4).

From this lemma we have that

$$\begin{aligned} &|H(T) - \hat{H}(T)| = \Big|\sum_t p(t)\log(p(t)) - \hat{p}(t)\log(\hat{p}(t))\Big| \\ &\leq \sum_t \phi(p(t) - \hat{p}(t)) \\ &= \sum_t \phi\left(\sum_x p(t|x)(p(x) - \hat{p}(x))\right) \\ &\leq \sum_t \phi\left(\sqrt{V(\mathbf{p}(T=t|x))}\|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\|\right), \end{aligned} \tag{16}$$

where the last inequality is derived as in Eq. (14), by setting $a \triangleq \frac{1}{|\mathcal{X}|}\sum_x p(T=t|x)$.

From Eq. (15) and Eq. (16) we conclude the following deterministic bound:

$$\begin{aligned} &|I(X;T) - \hat{I}(X;T)| \leq \\ &\sum_t \|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\| \cdot \phi\left(\sqrt{V(\mathbf{p}(T=t|x))}\right) \\ &\quad + \|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\| \cdot \sqrt{V(\mathbf{H}(T|x))}. \end{aligned} \tag{17}$$

Turning now to $|I(Y;T) - \hat{I}(Y;T)|$, we similarly use the inequality

$$|I(Y;T) - \hat{I}(Y;T)| \leq |H(T|Y) - \hat{H}(T|Y)| + |H(T) - \hat{H}(T)|.$$

It remains to bound the first summand, as the second summand was already bounded above. We have

$$\begin{aligned} &|H(T|Y) - \hat{H}(T|Y)| \\ &= \Big|\sum_y \left(p(y)H(T|y) - \hat{p}(y)\hat{H}(T|y)\right)\Big| \\ &\leq \Big|\sum_y p(y)\left(H(T|y) - \hat{H}(T|y)\right)\Big| \\ &\quad + \Big|\sum_y (p(y) - \hat{p}(y))\hat{H}(T|y)\Big|. \end{aligned} \tag{18}$$

For the first summand in this bound we have

$$\begin{aligned} &\Big|\sum_y p(y)\left(H(T|y) - \hat{H}(T|y)\right)\Big| \\ &\leq \Big|\sum_y p(y)\sum_t \left(\hat{p}(t|y)\log(\hat{p}(t|y)) - p(t|y)\log(p(t|y))\right)\Big| \\ &\leq \sum_y p(y)\sum_t \phi\left(\hat{p}(t|y) - p(t|y)\right) \\ &= \sum_y p(y)\sum_t \phi\left(\sum_x p(t|x)\left(\hat{p}(x|y) - p(x|y)\right)\right) \\ &= \sum_y p(y)\sum_t \phi\left(\|\hat{\mathbf{p}}(x|y) - \mathbf{p}(x|y)\| \cdot \sqrt{V(\mathbf{p}(T=t|x))}\right), \end{aligned}$$

where the last inequality is again derived similarly to Eq. (14), by setting $a \triangleq \frac{1}{\mathcal{X}}\sum_x p(t|x)$. For the second summand in Eq. (18) we have

$$\Big|\sum_y (p(y) - \hat{p}(y))\hat{H}(T|y)\Big| \leq \|\mathbf{p}(y) - \hat{\mathbf{p}}(y)\| \cdot \sqrt{V(\hat{\mathbf{H}}(T|y))}.$$

Therefore,

$$\begin{aligned} &|H(T|Y) - \hat{H}(T|Y)| \leq \\ &\sum_y p(y)\sum_t \phi\left(\|\hat{\mathbf{p}}(x|y) - \mathbf{p}(x|y)\| \cdot \sqrt{V(\mathbf{p}(T=t|x))}\right) \\ &+ \|\mathbf{p}(y) - \hat{\mathbf{p}}(y)\| \cdot \sqrt{V(\hat{\mathbf{H}}(T|y))}. \end{aligned} \tag{19}$$

From Eq. (16) and Eq. (19) we conclude the following deterministic bound:

$$\begin{aligned} &|I(Y;T) - \hat{I}(Y;T)| \leq \\ &\sum_t \|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\| \cdot \phi\left(\sqrt{V(\mathbf{p}(T=t|x))}\right)) \\ &+ \sum_y p(y)\sum_t \phi\left(\|\hat{\mathbf{p}}(x|y) - \mathbf{p}(x|y)\| \cdot \sqrt{V(\mathbf{p}(T=t|x))}\right) \\ &+ \|\mathbf{p}(y) - \hat{\mathbf{p}}(y)\| \cdot \sqrt{V(\hat{\mathbf{H}}(T|y))}. \end{aligned} \tag{20}$$

In order to transform the bounds in Eq. (17) and Eq. (20) 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 (Cristianini and Shawe-Taylor, 2004): 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,

$$\|\rho - \hat{\rho}\|_2 \leq \frac{2 + \sqrt{2\log(1/\delta)}}{\sqrt{m}}. \tag{21}$$

To make sure all of our uses of this result hold simultaneously for any T with a probability of $1-\delta$, we use Eq. (21) with δ replaced by $\delta/(|\mathcal{Y}|+2)$. Applying this concentration bound to $\|\mathbf{p}(x) - \hat{\mathbf{p}}(x)\|$, $\|\mathbf{p}(y) - \hat{\mathbf{p}}(y)\|$ and $\|\hat{\mathbf{p}}(x|y) - \mathbf{p}(x|y)\|$ in Eq. (17) and Eq. (20), and using the union bound, we get that the following bounds hold simultaneously for any T with a probability of $1-\delta$:

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

and

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

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

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

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.

6.3 Proof of Thm. 3

In this proof we apply worst-case assumptions on Thm. 2 to get a bound that does not depend on $p(t|x)$ but only on the cardinality of T. The variance of any random variable bounded in $[0,1]$ is at most $1/4$. Since $\frac{1}{n}V(\mathbf{p}(T=t|x))$ is the variance of the vector $\mathbf{p}(T=t|x)$, we have that $V(\mathbf{p}(T=t|x)) \leq |\mathcal{X}|/4$ for any $p(t|x)$. Assume that

$$m \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{X}|e^2n^2(\delta), \tag{22}$$

for C as in Thm. 2, then it follows that for any $p(t|x)$,

$$\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)V(\mathbf{p}(T=t|x))}{m}} \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)|\mathcal{X}|}{4m}} \leq 1/e.$$

For readability, we define $\mathcal{V} \triangleq C\log(|\mathcal{Y}|/\delta)V(\mathbf{p}(T=t|x))$. Therefore we have that

$$\sum_t \phi\left(\sqrt{\frac{\mathcal{V}}{m}}\right) = \sum_t \left(\sqrt{\frac{\mathcal{V}}{m}}\log\left(\sqrt{\frac{m}{\mathcal{V}}}\right)\right)$$
$$\leq \sum_t \frac{\sqrt{\mathcal{V}}\log(\sqrt{m})+1/e}{\sqrt{m}},$$

where the last inequality follows from $\sqrt{\mathcal{V}}\log(\frac{1}{\sqrt{\mathcal{V}}}) \leq 1/e$. Reintroducing the definition of $\mathcal{V}$ and rearranging, we have

$$\sum_t \phi\left(\sqrt{\frac{\mathcal{V}}{m}}\right) \leq \tag{23}$$
$$\frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(\sum_t \sqrt{V(\mathbf{p}(T=t|x))}\right)+\frac{2}{e}|\mathcal{T}|}{2\sqrt{m}}.$$

To bound $\sum_t \sqrt{V(\mathbf{p}(T=t|x))}$, we note that

$$\sum_t \sqrt{V(\mathbf{p}(T=t|x))} \leq \sum_t \|\mathbf{p}(T=t|x)\|_2.$$

Finding an upper bound for the right-hand expression is equivalent to solving the following optimization problem

$$\max_{a_{i,j}} \quad \sum_t \sqrt{\sum_x a_{t,x}^2}$$
$$\text{s.t.} \quad \forall x \ \ \sum_t a_{t,x}=1 \ \ , \ \ \forall t,x \ \ a_{t,x}\geq 0.$$

It is easily seen that in this problem we are maximizing a convex function over a compact convex set. It is well known (e.g. (Rockafellar, 1970)) that the maximal values in this case are achieved on vertices of the set. In other words, we can limit ourselves to solutions $\{a_{t,x}\}$ such that for any x, $a_{t,x}=\mathbf{1}_{t=t_x^*}$ where t_x^* is a function of x. Letting $b_t=\sqrt{|\{x:t_x^*=t\}|}$, we get the following equivalent optimization problem:

$$\max_{b_t} \quad \sum_t b_t$$
$$\text{s.t.} \quad \sum_t b_t^2 = |\mathcal{X}| \ \ , \ \ \forall t \ \ b_t^2\in\mathbb{Z}_+$$

To upper bound this, we can relax the integer constraint, and get the following problem

$$\max_{\mathbf{b}=(b_1,\ldots,b_{|\mathcal{T}|})} \quad \|\mathbf{b}\|_1$$
$$\text{s.t.} \quad \|\mathbf{b}\|_2=\sqrt{|\mathcal{X}|} \ \ , \ \ \mathbf{b}\in\mathbb{R}^{|\mathcal{T}|},$$

whose optimal solution is of course $\sqrt{|\mathcal{X}||\mathcal{T}|}$ by choosing $b_t=\sqrt{|\mathcal{X}|/|\mathcal{T}|}$ for all t. We can plug this bound back into Eq. (23) to get that

$$\sum_t \phi\left(\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)V(\mathbf{p}(T=t|x))}{m}}\right)$$
$$\leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)|\mathcal{X}||\mathcal{T}|}\log(m)+\frac{2}{e}|\mathcal{T}|}{2\sqrt{m}}. \tag{24}$$

To complete the proof, note that $H(T|x)$ and $\hat{H}(T|y)$ are in $[0,\log(|\mathcal{T}|)]$. Therefore

$$V(\mathbf{H}(T|x)) \leq \frac{|\mathcal{X}|\log^2(|\mathcal{T}|)}{4}, \tag{25}$$

and

$$V(\hat{\mathbf{H}}(T|y)) \leq \frac{|\mathcal{Y}|\log^2(|\mathcal{T}|)}{4}, \tag{26}$$

Applying Eq. (24), Eq. (25) and Eq. (26) on the bounds in Thm. 2 generates the required result.

Finally, it is easy to show that the resulting bound is trivially true for m not satisfying Eq. (22), and thus this bound it true for any m.

6.4 Proof of Thm. 4

Throughout the proof we assume that our model T pertains only to values of X, Y actually observed in the sample, and therefore w.l.o.g $p(x), p(y) > 0$ for any $x \in \mathcal{X}, y \in \mathcal{Y}$ of interest.

To prove this theorem, we will find a new upper bound for Eq. (6), using the same notation as in Thm. 2. As a shorthand, We denote the two summands of Eq. (6) by S_1 for the first summand and S_2 for the second summand, so that we have $|I(Y;T) - \hat{I}(Y;T)| \leq S_1 + S_2$. We start by bounding S_2, and as first step will seek an upper bound for $\sqrt{V(\mathbf{p}(T=t|x))}$.

By definition of $V(\cdot)$ and using Bayes' formula $p(t|x) = \frac{p(x|t)p(t)}{p(x)}$, we have that

$$\sqrt{V(\mathbf{p}(T=t|x))} = \tag{27}$$
$$p(t)\sqrt{\sum_x \left(\frac{p(x|t)}{p(x)} - \frac{1}{|\mathcal{X}|}\sum_{x'}\frac{p(x'|t)}{p(x')}\right)^2}.$$

Denoting $\mathbf{1} = (1, \ldots, 1)$, we have by the triangle inequality that

$$\begin{aligned}
&\sqrt{\sum_x \left(\frac{p(x|t)}{p(x)} - \frac{1}{|\mathcal{X}|}\sum_{x'}\frac{p(x'|t)}{p(x')}\right)^2} \\
&\leq \|\frac{p(x|t)}{p(x)} - \mathbf{1}\|_2 + \sqrt{\sum_x \left(1 - \frac{1}{|\mathcal{X}|}\sum_{x'}\frac{p(x'|t)}{p(x')}\right)^2} \\
&= \|\frac{p(x|t)}{p(x)} - \mathbf{1}\|_2 + \frac{1}{\sqrt{|\mathcal{X}|}}\Big|\sum_{x'}(1 - \frac{p(x'|t)}{p(x')})\Big| \\
&= \|\frac{p(x|t)}{p(x)} - \mathbf{1}\|_2 + \frac{1}{\sqrt{|\mathcal{X}|}}\|\frac{p(x|t)}{p(x)} - \mathbf{1}\|_1 \\
&\leq \left(1 + \frac{1}{\sqrt{|\mathcal{X}|}}\right)\|\frac{p(x|t)}{p(x)} - \mathbf{1}\|_1 \\
&\leq \frac{2}{\min_x p(x)}\|p(x|t) - p(x)\|_1
\end{aligned} \tag{28}$$

From an inequality linking KL-divergence and the L_1 norm (lemma 12.6.1 in (Cover and Thomas, 1991)), we have that

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

Plugging this into Eq. (28) and using Eq. (27), we get the following bound:

$$\sqrt{V(\mathbf{p}(T=t|x))} \leq \frac{2\sqrt{2\log(2)}}{\min_x p(x)}p(t)\sqrt{\mathrm{D_{KL}}[p(x|t)\|p(x)]}. \tag{29}$$

For notational convenience, let

$$g(m) = \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)}{m}} \cdot \frac{2\sqrt{2\log(2)}}{\min_x p(x)},$$

and let $d_t = \mathrm{D_{KL}}[p(x|t)\|p(x)]$. Then, using Eq. (29), we have

$$S_2 \leq 2\sum_t \phi(g(m)p(t)\sqrt{d_t}). \tag{30}$$

At this point, let us assume that given T, m is large enough so that $g(m)p(t)\sqrt{d_t} \leq 1/e$ for any t. We will later see that this condition can be discarded. For such m, we get by definition of $\phi(\cdot)$ that

$$\begin{aligned}
S_2 &\leq 2\sum_t g(m)p(t)\sqrt{d_t}\left(\log\left(\frac{1}{g(m)}\right) + \log\left(\frac{1}{p(t)\sqrt{d_t}}\right)\right) \\
&= 2g(m)\Bigg(\log\left(\frac{1}{g(m)}\right)\sum_t p(t)\sqrt{d_t} \\
&\qquad + \sum_t p_t\sqrt{d_t}\log\left(\frac{1}{p(t)\sqrt{d_t}}\right)\Bigg).
\end{aligned}$$

It is easily verified that for any $x > 0$, $x\log(1/x) \leq \sqrt{x}$. Using this fact and thinking of $p(t)\sqrt{d_t}$ as a vector indexed by t, we have

$$S_2 \leq 2g(m)\left(\log\left(\frac{1}{g(m)}\right)\|p(t)\sqrt{d_t}\|_1 + \|\sqrt{p(t)\sqrt{d_t}}\|_1\right).$$

We use the following two inequalities:

$$\|p(t)\sqrt{d_t}\|_1 \leq \sqrt{|\mathcal{T}|}\|p(t)\sqrt{d_t}\|_2 \leq \sqrt{|\mathcal{T}|}\|\sqrt{p(t)d_t}\|_2,$$

and

$$\begin{aligned}
&\|\sqrt{p(t)\sqrt{d_t}}\|_1 \leq \sqrt{|\mathcal{T}|}\|\sqrt{p(t)\sqrt{d_t}}\|_2 \\
&\quad = \sqrt{|\mathcal{T}|}\sqrt{\|p(t)\sqrt{d_t}\|_1} \leq |\mathcal{T}|^{3/4}\sqrt{\|\sqrt{p(t)d_t}\|_2},
\end{aligned}$$

to have

$$\begin{aligned}
S_2 \leq 2g(m)\Big(&\log\left(\frac{1}{g(m)}\right)\sqrt{|\mathcal{T}|}\|\sqrt{p(t)d_t}\|_2 \\
&+ |\mathcal{T}|^{3/4}\sqrt{\|\sqrt{p(t)d_t}\|_2}\Big).
\end{aligned}$$

Using the equality

$$\|\sqrt{p(t)d_t}\|_2 = \sqrt{\mathbb{E}_t\left[\mathrm{D_{KL}}[p(x|t)\|p(x)]\right]} = \sqrt{I(X;T)},$$

we reach the following bound

$$\begin{aligned}
S_2 \leq 2g(m)\Big(&\log\left(\frac{1}{g(m)}\right)\sqrt{|\mathcal{T}|I(X;T)} \\
&+ |\mathcal{T}|^{3/4}(I(X;T))^{1/4}\Big).
\end{aligned} \tag{31}$$

By inserting the definition of $g(m)$ back into the inequality, we get our final bound for S_2,

$$\begin{aligned}
S_2 \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)}{m}}\Big(&C_1\log(m)\sqrt{|\mathcal{T}|I(X;T)} \\
&+ C_2|\mathcal{T}|^{3/4}(I(X;T))^{1/4}\Big).
\end{aligned} \tag{32}$$

with C_1 and C_2 as constants that depend only on $min_x p(x)$.

Turning now to S_1, we have to bound $\sqrt{V(\hat{\mathbf{H}}(T|y))}$. By definition of $V(\cdot)$, and using the triangle inequality, we have

$$\sqrt{V(\hat{\mathbf{H}}(T|y))} \leq \sqrt{\sum_y (\hat{H}(T|y) - \hat{H}(T))^2} + \sqrt{\sum_y \left(\hat{H}(T) - \frac{1}{|\mathcal{Y}|}\sum_{y'} \hat{H}(T|y')\right)^2}$$

For the second summand we have

$$\begin{aligned}
&\sqrt{\sum_y \left(\hat{H}(T) - \frac{1}{|\mathcal{Y}|}\sum_{y'} \hat{H}(T|y')\right)^2} \\
&= \sqrt{|\mathcal{Y}|}\left|\hat{H}(T) - \frac{1}{|\mathcal{Y}|}\sum_{y'} \hat{H}(T|y')\right| \\
&= \frac{1}{\sqrt{|\mathcal{Y}|}}\left|\sum_{y'}(\hat{H}(T) - \hat{H}(T|y'))\right| \\
&= \frac{1}{\sqrt{|\mathcal{Y}|}}\|\hat{\mathbf{H}}(T) - \hat{\mathbf{H}}(T|y)\|_1,
\end{aligned}$$

where we think of $\hat{\mathbf{H}}(T) - \hat{\mathbf{H}}(T|y)$ as a vector ranging over the values of y. Therefore, we have that

$$\sqrt{V(\hat{\mathbf{H}}(T|y))} \leq \left(1 + \frac{1}{\sqrt{|\mathcal{Y}|}}\right)\|\hat{\mathbf{H}}(T) - \hat{\mathbf{H}}(T|y)\|_1. \quad (33)$$

It is known that $\hat{H}(T) \geq \hat{H}(T|y)$ for any y, since conditioning cannot increase entropy. Therefore

$$\begin{aligned}
\|\hat{H}(T) - \hat{H}(T|y)\|_1 &\leq \sum_y \frac{p(y)}{\min_y p(y)}\left(\hat{H}(T) - \hat{H}(T|y)\right) \\
&= \frac{1}{\min_y p(y)}\left(\hat{H}(T) - \sum_y p(y)\hat{H}(T|y)\right) \\
&= \frac{1}{\min_y p(y)}\hat{I}(Y;T) \leq \frac{1}{\min_y p(y)}\hat{I}(X;T),
\end{aligned}$$

where the last inequality follows from the data processing inequality. Substituting this into Eq. (33), and since $|\mathcal{Y}| \geq 1$, we get

$$\sqrt{V(\hat{\mathbf{H}}(T|y))} \leq \frac{2}{\min_y p(y)}\hat{I}(X;T). \quad (34)$$

Setting $C_3 = \frac{2}{\min_y p(y)}$ we thus have our bound for S_1,

$$S_1 \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)}{m}} C_3 \hat{I}(X;T).$$

Plugging Eq. (32) and Eq. (34) into Eq. (6) gives us the bound in our theorem.

Lastly, recall that we derived this bound by assuming that $g(m)p(t)\sqrt{d_t} \leq 1/e$ for any t. We now show that the bound can be made trivial if this condition does not hold. If the condition does not hold, there exists a t such that $g(m)p(t)\sqrt{d_t} > 1/e$. Since

$$\sqrt{I(X;T)} = \sqrt{\sum_t p(t)d_t} \geq p(t)\sqrt{d_t}$$

for any t, we get that $\sqrt{I(X;T)} \geq \frac{1}{e \cdot g(m)}$. Since $|\mathcal{T}| \geq 1$ and $g(m) > 0$, we get that our bound in Eq. (31) is at least

$$\begin{aligned}
&2g(m)\Big(\log\left(\frac{1}{g(m)}\right)\sqrt{|\mathcal{T}|I(X;T)} + |\mathcal{T}|^{3/4}(I(X;T))^{1/4}\Big) \\
&\geq 2\sqrt{|\mathcal{T}|}\left(\frac{\log(1/g(m))}{e} + |\mathcal{T}|^{1/4}\sqrt{\frac{g(m)}{e}}\right) \\
&\geq \sqrt{|\mathcal{T}|} \geq \log(|\mathcal{T}|)
\end{aligned}$$

Therefore if indeed $g(m)p(t)\sqrt{d_t} > 1/e$ for some t, then the bound in the theorem is trivially true, since $I(Y;T), \hat{I}(Y;T)$ are both within $[0, \log(|\mathcal{T}|)]$. Hence the bound in Thm. 4 holds for any m.

6.5 Proof of Thm. 5

Similar formulations of Thm. 5 can be gleaned from (Kullback and Leibler) and (Cover and Thomas, 1991). We present here the full proof of our formulation for completeness. Thm. 5 follows directly from the following two lemmas.

We denote by $\mathcal{F}(X)$ the set of random mappings of X, and by $\mathcal{S}(Y)$ the set of sufficient statistics for Y.

Lemma 3. *Let T be a random mapping of X. Then T is a sufficient statistic for Y if and only if*

$$I(Y;T) = \max_{T' \in \mathcal{F}(X)} I(Y;T')$$

Proof. First, assume that T is a sufficient statistic for Y. For every T' which is a random mapping of X, we have the Markov chain $Y - X - T'$. Therefore, by the data processing inequality, $I(Y;X) \geq I(Y;T')$. In addition, $X \in \mathcal{F}(X)$. Therefore

$$I(Y;X) = \max_{T' \in \mathcal{F}(X)} I(Y;T').$$

Since T is a sufficient statistic, $Y - T - X$ is also a Markov chain, hence $I(Y;X) \leq I(Y;T)$. It follows that

$$I(Y;T) = I(Y;X) = \max_{T' \in \mathcal{F}(X)} I(Y;T').$$

This completes one direction of the claim. For the other direction, assume that

$$I(Y;T) = \max_{T' \in \mathcal{F}(X)} I(Y;T').$$

Then $I(Y;T) = I(Y;X)$. Since $Y - X - T$ is a Markov chain, it follows that Y and X are conditionally independent given T (see (Cover and Thomas, 1991)), hence T is a sufficient statistic. $\square$

Lemma 4. *Let T be a sufficient statistic for Y. Then T is a minimal sufficient statistic for Y if and only if*

$$I(X;T) = \min_{T' \in \mathcal{S}(Y)} I(X;T'). \tag{35}$$

Proof. First, let T be a minimal sufficient statistic, and let T' be some sufficient statistic. By the definition of a minimal sufficient statistic, there is a function f such that $T = f(T')$. Therefore, $X - T' - T$ is a Markov chain. Therefore, $I(X;T) \leq I(X;T')$. This holds for any sufficient statistic T', hence indeed Eq. (35) holds. This completes the first direction of the proof.

For the second direction, we show that if T is not minimal, then there exists a sufficient statistic V such that $I(X;T) > I(X;V)$, thus Eq. (35) does not hold. We will use the Fisher-Neyman factorization theorem (Fisher, 1922) which states that T is a sufficient statistic for Y if and only if there exist functions h_T and g_T such that

$$\forall x, y \quad p(x|y) = h_T(x) g_T(T(x), y). \tag{36}$$

Since T is not minimal, there exists a sufficient statistic T' such that T is not a function of T'. Define the equivalence relation $\sim$ by

$$t_1 \sim t_2 \iff \frac{g_T(t_1, y)}{g_T(t_2, y)} \text{ is a constant function of } Y,$$

where g_T is a function satisfying Eq. (36) with some h_T. Let $V : \mathcal{X} \to \mathcal{T}$ be a function such that

$$\forall x, \quad V(x) \in \{t \mid t \sim T(x)\}.$$

V is thus a function of T. We use Fisher-Neyman's theorem to show that V is a sufficient statistic: Define

$$h_V(x) \triangleq h_T(x) \frac{g_T(T(x), y)}{g_T(V(x), y)}$$
$$g_V(V(x), y) \triangleq g_T(V(x), y).$$

Then

$$\begin{aligned} p(x|y) &= h_T(x) g_T(T(x), y) \\ &= h_T(x) \frac{g_T(T(x), y)}{g_T(V(x), y)} g_T(V(x), y) \\ &= h_V(x) g_V(V(x), y). \end{aligned}$$

Therefore V has a factorization; hence it is a sufficient statistic. It is left to show that $I(X;T) > I(X;V)$. V is a function of T', for let x_1, x_2 such that $T'(x_1) = T'(x_2)$, then

$$\begin{aligned} \frac{g_T(T(x_1), y)}{g_T(T(x_2), y)} &= \frac{p(x_1|y) h_T(x_2)}{p(x_2|y) h_T(x_1)} \\ &= \frac{h_{T'}(x_1) g_{T'}(T'(x_1), y) h_T(x_2)}{h_T(x_1) g_{T'}(T'(x_1), y) h_{T'}(x_2)} \\ &= \frac{h_{T'}(x_1) h_T(x_2)}{h_T(x_1) h_{T'}(x_2)}. \end{aligned}$$

Hence $T(x_1) \sim T(x_2)$, therefore $V(x_1) = V(x_2)$ for any x_1, x_2 such that $T'(x_1) = T'(x_2)$.

Since $X - T - V$ is a Markov chain, we have

$$\begin{aligned} I(X;T) =& I(X;V) + I(X;T \mid V) \\ &\geq I(X;V) + I(X;T \mid T', V) \\ &= I(X;V) + I(X;T \mid T'). \end{aligned}$$

since T is a function of X but is not a function of T', we have that $I(X;T \mid T') > 0$. Therefore $I(X;T) > I(X;V)$, hence Eq. (35) does not hold. □

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 bounds without resorting to VC-theory or similar complexity measures common in learning theory.

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 analyze 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 provides a theoretical justification for many applications of interest, and in fact characterizes the learning scenarios for which this method is best suited for. Finally, we discuss how this framework extends the classical statistical concept of minimal sufficient statistics.

Although we have focused here on a setting involving two discrete variables, our results may also be relevant for a more complete learning theoretic analysis of information theoretic based algorithms.

References

A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19(3–4):163–193, 2001.

L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In W. Bruce Croft, Alistair Moffat, Cornelis J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, AU, 1998. ACM Press, New York, US.

R. Gilad-Bachrach, A. Navot, and N. Tishby. An Information Theoretic Tradeoff between Complexity and Accuracy. In *Proceedings of Computational Learning Theory (COLT)*, pages 595–609, 2003.

G. Chechik, A. Globerson, N. Tishby, Y. Weiss. An Information Bottleneck for Gaussian Variables. In *Journal of Machine Learning Research (JMLR)*, 6(Jan):165–188, 2005.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

R.A. Fisher. On the mathematical foundation of theoretical statistics. *Philos. Trans. Roy. Soc., London*, Sec. A, 222:309–368, 1922.

N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proceedings in Uncertainty in Artificial Intelligence 17*, pages 152–161, 2001.

GuyonGuNiZa06 M. Nikravesh I. Guyon, S. Gunn and L. A. Zadeh, editors. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer, 2006.

P. Harremöes 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.

S. Kullback and R.A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

E.L. Lehmann and H. Scheffé. Completeness, similar regions and unbiased estimation. *Sankhya*, 10:305–340, 1950.

E.L. Lehmann. Testing Statistical Hypotheses. John Wiley and sons, New-York, 1959.

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.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

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.

B. Koopman. On distributions admitting a sufficient statistic. *Trans. Amer. math. Soc.*, 39:399–409, 1936.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, Hebrew University, Jerusalem, 2003.

N. Slonim, G. Singh S. Atwal, G. Tkacik, and W. Bialek. Information-based clustering. *Proc. Natl. Acad. Sci. U.S.A*, December 2005.

N. Slonim and N. Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*, 2001.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *The 37'th Allerton Conference on Communication, Control, and Computing*, 1999.

N. Tishby and N. Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *Advances in Neural Information Processing Systems 13*, pages 640–646, 2000.