An Information Theoretic Tradeoff between Complexity and Accuracy

Ran Gilad-Bachrach Amir Navot Naftali Tishby

School of Computer Science and Engineering and
Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem, Israel
ranb,anavot,tishby@cs.huji.ac.il

Abstract. A fundamental question in learning theory is the quantification of the basic tradeoff between the complexity of a model and its predictive accuracy. One valid way of quantifying this tradeoff, known as the "*Information Bottleneck*", is to measure both the complexity of the model and its prediction accuracy by using Shannon's mutual information. In this paper we show that the Information Bottleneck framework answers a well defined and known coding problem and at same time it provides a general relationship between complexity and prediction accuarcy, measured by mutual information. We study the nature of this complexity-accuracy tradeoff and discuss some of its theoretical properties. Furthermore, we present relations to classical information theoretic problems, such as *rate-distortion theory*, *cost-capacity tradeoff* and *source coding with side information*.

1 Introduction

Learning, in human and machine, is known to be related to the ability to find compact representations of data. In supervised learning this is done by choosing an hypothesis which somehow summarizes the training data, where as in unsupervised learning - clusters or low dimensional features play the same role. In both cases we are interested in a concise description that preserves the *relevant* essence of the data. Therefore any learning process has to deal with the basic tradeoff between the complexity (conciseness) of the data representation available and the best accuracy (goodness of fit) that this complexity enables. The measures of complexity and accuracy may change from one task to another. In learning theory complexity is commonly measured by the VC-dimension, covering numebrs, metric entropies of the class, or by the description (coding) length of the representation. Accuracy can be measured by generalization error, mistake bound, clusters purity, feature efficiency, and various other ways.

In this paper we choose study the nature of the tradeoff between complexity and accuracy using information theoretic concepts. The main advantage of this choice is in its model independece and its powerful asymthotic properties. We therefore measure the representation complexity by the minimal number of bits needed to describe the data per sample - known as rate. We choose to measure the accuracy by the amount of information our data representation preserves on the target variable. While the precise nature of the target variable depends on the task, labels in supervised learning, it can be categories, noisy data, or other weakly dependent random variable. Statistical dependence with the data seems to be the only general property of the target - thus universally quantified by mutual information. The fact that both complexity and accuarcy can be quantifed by mutual information - as proposed in the *Information Bottleneck* (IB) framework - enables us to quantify their traedoff in a general yet very precise way, as shown in this study.

The Information-Bottleneck (IB) method was first introduced by Tishby, Pereira and Bialek [12] about 5 years ago. The *relevant information* in a one variable (signal) - with respect to another one - is defined as the (mutual) information that this variable provides about the other. One can think about it as the minimal number of binary questions, on average, that should be asked about the values of one variable, in order to reduce as much as possible the uncertainty in the value of the other variable. Examples include the relationship between document category and its words statistics, face features and person's identity, speech sounds and spoken words, gene expression levels and tissue samples, etc. In all such cases, the problem is how to map one of the variables, considered as the source signal, X into a more concise reproduction signal $\hat{X}$ while preserving the information about the relevant (predicted) signal Y.

The Information-Bottleneck was found useful in various learning applications. Slonim et. al. [10, 11] used it for clustering data. Note that since IB uses information theoretic point of view it does not suffer from basic flaws of geometric based algorithms as presented by Kleinberg [5]. In [9] the IB was used for feature selection. Poupart et. al. [6] used it while studying POMDPs, and Baram et. al. [2] used it for evaluating the expected performance of active learning algorithms. For a comprehensive study of the IB see [8].

The IB is related to several classic problems in information theory such as *rate-distortion* and *cost-capacity* [7]. In rate-distortion problem the goal is to encode a source in a way that minimizes the code length under a constraint on the average distortion. In cost-capacity problem a cost is assigned to each symbol of the channel alphabet. The task is to minimize the ambiguity at the receiver under a constraint on the average cost. The IB can be presented in "rate-distortion like" formulation, but with non-fixed distortion measure and at the same time it can take the form of "cost capacity like" problem with non-fixed cost function. Moreover it combines these two problems in a way that is free of the arbitrary nature of both the distortion measure and channel cost. Another formally related problem is *source coding with side information*[14, 1]. The setting of this problem is very different but the solution happen to be similar. See section 3 for a detailed discussion of this relationship.

1.1 Summary of Results

- We define the IB coding problem and the IB optimization problem in section 2.

– We discuss the relationship between the IB and the problem of source coding with side information in section 3.
– We show that the IB optimization problem provides a tight lower bound for the IB coding problem in section 4.
– We define the IB-tradeoff (information-curve) which assigns to any restriction on $I(\hat{X};Y)$ the minimal possible value of $I(X;\hat{X})$, where the free variables are the conditional distributions $p(\hat{x}|x)$. In section 5 we study this functional optimization and utilize its formal relation to source coding with side information to prove that the IB-curve is a smooth, monotone and convex function. Furthermore we show that $|\hat{\mathcal{X}}| = |\mathcal{X}| + 2$ is sufficient to achieve the best encoding.
– In section 6, we show that the Information Bottleneck optimization is locally equivalent to a Rate-Distortion problem with an adequate distortion measure, and thus can be considered as a rate-distortion problem with a variable distortion measure. We also show that the information curve is the envelope of all such "locally equivalent" rate-distortion functions.
– In section 7, a dual representation of the IB problem is presented. In the dual representation the IB takes the form of a cost-capacity problem with non fixed cost function. However the optimum of the primal (rate-distortion like) and the dual (cost-capacity like) is equivalent.

1.2 Preliminaries and Notation

We use the following notation: $X, Y, \hat{X}$ are random variables over a finite alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\hat{\mathcal{X}}$ respectively. $x, y, \hat{x}$ are instances of these variables. $H(X)$ denote the Entropy of the random variable X, the Mutual Information between X and Y is denoted by $I(X;Y)$ and $D_{KL}[p(x)||q(x)]$ is the Kullback and Liebler (KL)-Divergence between the two distributions $p(x)$ and $q(x)$. We also assume that $|\hat{\mathcal{X}}| \geq |\mathcal{X}| + 2$ in this paper unless specified otherwise[1] All logarithms are base 2.

2 Problem Setting

Assume that we have two random variables X and Y, and their joint distribution $p(x, y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. We would like to encode X using reproduction alphabet $\hat{\mathcal{X}}$ in a way that keeps the maximum information on Y for a given rate or alternatively, use the minimal rate for a given value I_Y of information on Y).

As usually done in Information Theory we use block encoding and thus discuss average rate and *average information*. We will show that the tradeoff between these two values can be discovered using the optimization problem presented in definition 3. Before going any further we introduce the definitions of encoding and how do we measure its average information. Note that the definition is similar to the one of rate-distortion code [4]. The only difference is that the average distortion is replaced by the Y-information.

[1] This last assumption is used for proving the convexity of the information curve (lemma 5), which by itself is used in some other proofs.

Definition 1 (Rate Information Code). *A* $(2^{nR}, n)$ *rate information code consists of:*

$$\begin{aligned} &\textit{an encoding function:} && f_n : \mathcal{X}^n \longrightarrow \{1, 2, \cdots, 2^{nR}\} \\ &\textit{a decoding function:} && g_n : \{1, 2, \cdots, 2^{nR}\} \longrightarrow \hat{\mathcal{X}}^n \end{aligned}$$

The Y*-information associated with the* $(2^{nR}, n)$ *is defined as*

$$Y_{info}(f_n, g_n) = \frac{1}{n} \sum_{i=1}^{n} I\Big(\big(g_n \circ f_n(X^n)\big)_i \; ; \; \big(Y^n\big)_i\Big)$$

where the information of the i*'s element is calculated with respect to the distribution defined by the code for this coordinate as follows:*

$$\bar{p}_i(x, \hat{x}) = \sum_{x^n \; : \; (x^n)_i = x \; \wedge \; (g_n \circ f_n(x^n))_i = \hat{x}} p(x^n) \tag{1}$$

A rate information pair (R, I_Y) is said to be *achievable* if there exists a sequence of rate information codes (f_n, g_n) with asymptotic rate R and asymptotic Y-information larger than or equal to I_Y, i.e. with $\lim_{n\to\infty} Y_{info}(f_n, g_n) \geq I_Y$

The *rate information region* is the closure of the set of achievable rate information pairs (R, I_Y).

Definition 2 (Rate Information Function). *The* rate information function $R(I_Y) : [0, I(X,Y)] \to [0, h(X)]$ *is the infimum of rates* R *such that* (R, I_Y) *is in the rate information region for a given constraint on the information on* Y*,* I_Y*.*

Definition 3 (IB-Function). *The* IB-function $R^{(I)}(I_Y)$ *for two random variables* X *and* Y *is defined as*

$$R^{(I)}(I_Y) = \min_{p(\hat{x}|x) : I(\hat{X};Y) \geq I_Y} I(X; \hat{X})$$

Where the minimization is over all the normalized distributions $p(\hat{x}|x)$*.*

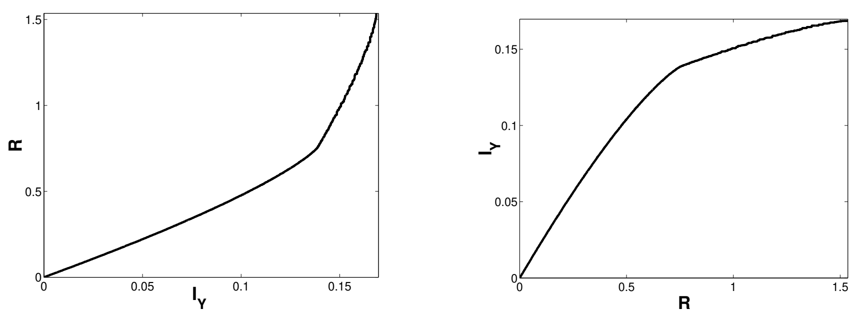
Note that the constraint depends on $p(y|\hat{x})$ which does not appear explicitly in the minimization problem, but is given by $p(y|\hat{x}) = \sum_x p(y|x)p(x|\hat{x})$, which follows from the Markov chain $\hat{X} \to X \to Y$. The minimum exists as $I(\hat{X}, X)$ is a continuous function of $p(\hat{x}|x)$ and the minimization is over a compact set.

It is also possible to define the dual function

$$I_Y^{(I)}(R) = \max_{p(\hat{x}|x) \; : \; I(\hat{X};X) \leq R} I(\hat{X}; Y) \tag{2}$$

later on we will show that the two functions are indeed equivalent, i.e. that they defines the same curve. Theorem 2 shows that the curve of $R(I_Y)$ is also the same. We will refer to this curve later on as the *IB-curve*. See figure 1 for an illustration of this curve.

Tishby et al. [12] used Lagrange multipliers to analyze the IB optimization problem. They proved that the conditional distribution $p(\hat{x}|x)$ which achieves the minimum has an exponential form, as stated in theorem 1.



(a) $R^{(I)}(I_Y)$ (b) $I_Y^{(I)}(R)$

Fig. 1. A typical IB-curve. The graphs of $R^{(I)}(I_Y)$ (left) and of $I_Y^{(I)}(R)$ (right). The curve was computed empirically for a joint distribution of size 3×3.

Theorem 1 (Tishby et al. 1999). *The optimal assignment, that minimizes the IB minimization problem given in definition 3, satisfies the equation*

$$p(\hat{x}|x) = \frac{p(\hat{x})}{Z(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p(y|\hat{x})]} \tag{3}$$

where β is the Lagrange multiplier corresponds to the constraint $I(\hat{X};Y) \geq I_Y$ and $Z(x,\beta) = \sum_{\hat{x}} p(\hat{x}) e^{-\beta D_{KL}[p(y|x)||p(y|\hat{x})]}$ is a normalization function. and the distribution $p(y|\hat{x})$ in the exponent is given via Bayes' rule and the Markov chain condition $\hat{X} \leftarrow X \leftarrow Y$, as,

$$p(y|\hat{x}) = \frac{1}{p(\hat{x})} \sum_x p(y|x) p(\hat{x}|x) p(x) \tag{4}$$

Note that this solution is a *formal* solution since $p(y|\hat{x})$ in the exponent is defined implicitly in the terms of the assignment mapping $p(\hat{x}|x)$.

3 Relation to Source Coding with Side Information

The problem of source coding with side information at the decoder is being studied in the Information Theory community since the mid seventies [14, 1]. It is also known as the Wyner-Ahlswede-Koroner (WAK) problem. Lately it was discovered [3] that it is closely related to the Information Bottleneck. In order to explore the relations between the two frameworks we first give here a short description of the WAK problem.

The WAK framework study the situation where one would like to encode information about one variable in a way that allow to reconstruct it in the presence of some information about another variable. More formally, let X and Y be two (non independent) random variables. Each of them is encoded separately with rates R_0 and R_1 accordingly. Both codes are available at the decoder. A pair of rates R_0, R_1 is *achievable* if it allows exact reconstruction of Y in the usual Shannon sense. X is referred as *side information*. See figure 2 for an illustration.

[14, 1] found independently, at the same time, that the minimal achievable rate R_1 for a given constraint $R_0 \leq r_0$ is given by

$$F(r_0) = \min_{p(\hat{x}|x):I(\hat{X};X)\leq r_0} H(Y|\hat{X})$$

By adding the constant $H(Y)$, we get

$$F(r_0) = H(Y) - I_Y^{(I)}(r_0) \tag{5}$$

where $I_Y^{(I)}(\cdot)$ is as defined in (2).

Although surprising, this equivalence (5) can be explained as follows: The value $F(r_0)$ measures the rate one should add on top of the side information in order to fully reconstruct Y. The compliment of this quantity is the amount of information known about Y from the side information $\hat{X}$. The last measures the quality of the quantization (accuracy) in the IB framework.

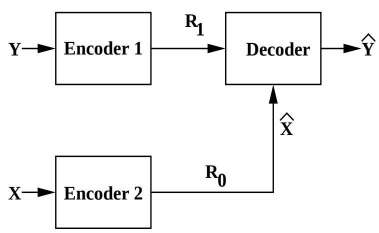


Fig. 2. The network correspond to the WAK problem

Using the similarity between the two problems it is possible to share the knowledge about the optimization problem. For an example, any algorithms that was developed for the IB is also valid for WAK. Anyway, despite the technical similarity, the motivation and the applications are very different and therefore the kind of questions arises are different. For an example, the coding theorems for IB and for WAK are related but different.

4 The Coding Theorem

In this section we state and prove our main results, that the IB-function is the solution of the coding problem presented in definition 2.

Theorem 2. *The rate information function for i.i.d sampling of (X, Y) is equal to the associated IB-function. Thus*

$$R(I_Y) = R^{(I)}(I_Y) \triangleq \min_{p(\hat{x}|x):I(\hat{X};Y)\geq I_Y} I(X;\hat{X}) \tag{6}$$

4.1 The IB-function Lower Bounds $\mathbf{R(I_Y)}$

First we show that any code which achieves Y-information larger than I_Y has rate of $R^{(I)}(I_Y)$ at least. The proof is very similar to the converse proof of the *rate-distortion theorem* (see [4]) and is given here for the sake of completeness. The proof uses some properties of $R^{(I)}(I_Y)$, that will be presented later on in section 5. We complete the proof of the theorem in section 4.2.

Proof (Lower bound in theorem 2). Consider any $(2^{nR}, R)$ code defined by functions f_n and g_n. Let $\hat{X}^n = \hat{X}^n(X^n) = g_n \circ f_n(X^n)$ be the reproducing sequence corresponding to X^n. The joint distribution induced by the code is:

$$\bar{p}(x^n, \hat{x}^n) = \begin{cases} p(x^n) \text{ if } \hat{x} = g_n \circ f_n(x^n) \\ 0 \qquad \text{otherwise} \end{cases}$$

Since at most 2^{nR} elements of $\hat{\mathcal{X}}^n$ are in use, the elements of X^n are independent and the fact that conditioning reduces entropy, we have that

$$\begin{aligned} nR &\geq H(\hat{X}^n) = H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1\cdots i-1}) \\ &\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \end{aligned} \tag{7}$$

using the definition and convexity of $R^{(I)}(\cdot)$ we have that

$$\begin{aligned} \sum_{i=1}^{n} I(X_i; \hat{X}_i) &\geq \sum_{i=1}^{n} R^{(I)}(I(\hat{X}_i; Y_i)) = n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(I(\hat{X}_i; Y_i)) \\ &\geq nR^{(I)} \left(\sum_{i=1}^{n} \frac{1}{n} I(\hat{X}_i; Y_i) \right) \end{aligned} \tag{8}$$

and since the marginals of $\bar{p}$ are exactly the $\bar{p}_i$'s defined in (1), and the fact that $R^{(I)}(\cdot)$ is non-decreasing,

$$nR^{(I)} \left(\sum_{i=1}^{n} \frac{1}{n} I(\hat{X}_i; Y_i) \right) = nR^{(I)} \left(Y_{info}(f_n, g_n) \right) \geq nR^{(I)} \left(I_Y \right) \tag{9}$$

Combining (7), (8) and (9) we get the stated result. $\square$

4.2 Achievability of The IB-function

We now prove the achievability of the IB-function $R^{(I)}(I_Y)$, or in other words that this function is a tight bound. Given any $p(\hat{x}|x)$, a series of codes with $R_n \to I(\hat{X}; X)$ and $Y_{info_n} \to I(\hat{X}; Y)$ is constructed. We enhance the standard construction of the rate-distortion in two ways: first we encode with respect to multiple distortion measures at the same time. Moreover, we require that the distortion for each coordinate of the block will be close to the distortion induced by $p(\hat{x}|x)$, while in rate-distortion only the average distortion counts. By selecting the appropriate distortion measures we complete the proof. First we introduce a few definitions and lemmas.

Definition 4 (multi distortion jointly typical). *Let $p(x,\hat{x})$ be a joint probability distribution on $\mathcal{X} \times \hat{\mathcal{X}}$. Let $d_1, \ldots, d_k$ be a set of distortion measures on $\mathcal{X} \times \hat{\mathcal{X}}$. For any $\epsilon > 0$, a pair of sequences $(x^n, \hat{x}^n)$ is said to be* multi distortion jointly ϵ-typical *if*

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon$$
$$\left| -\frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon$$
$$\left| -\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon$$
$$\left| d_j(x^n, \hat{x}^n) - E d_j(X, \hat{X}) \right| < \epsilon \qquad \forall j,\ 1 \leq j \leq k$$

where $d_j(x^n, \hat{x}^n)$ is defined as $\frac{1}{n}\sum_{i=1}^{n} d_j((x^n)_i, (\hat{x}^n)_i)$.
This set is denoted $A_\epsilon^{(n)}$

Lemma 1. *Let $(X_i, \hat{X}_i)$ be drawn i.i.d $\sim\ p(x,\hat{x})$. Then $Pr(A_\epsilon^{(n)}) \to 1$ as $n \to \infty$*

This result follows from the *central limit theorem.*

Lemma 2. *Given $p(x,\hat{x}) = p(\hat{x}|x)p(x)$ and a finite set of bounded distortion measures $d_1 \ldots d_k$ there exists a series of codes (f_n, g_n) with asymptotic rate $I(\hat{X};X)$, and such that for any d_j the average distortion of the code converges uniformly to $E_{p(\hat{x},x)} d_j(X,\hat{X})$. i.e. for any $\epsilon > 0$ there exist N such that for any d_j and any $n > N$ we have that*

$$\left| E_{x^n} d_j(X^n, g_n \circ f_n(X^n)) - E_{p(\hat{x},x)} d_j(X, \hat{X}) \right| < \epsilon$$

Proof. We use random codebook and mapping by multi distortion joint typicality as follows:
Generation of codebook Randomly generate a codebook $\mathcal{C}$ consisting of $2^{nI(X;\hat{X})}$ sequences $\hat{x}^n$ drawn i.i.d from $p(\hat{x})$. Index these codewords by $w \in \{1, 2, \ldots, 2^{nI(X;\hat{X})}\}$.
Encoding Encode x^n by w if there exist a w s.t. $(x^n, \hat{x}^n(w)) \in A_\epsilon^{(n)}$. If there is more than one such w send the least. If There is no such w let $w = 1$.

The rest of the proof, i.e. showing that with probability greater than zero this construction achieves the required distortion, is the same as the one used in the rate-distortion theorem that can be found for example in [4] (pp.350). □

Definition 5 (The distortion of a coordinate). *Given source $p(x)$, a distortion measure d and a code (f_n, g_n).* The distortion of a coordinate *i is*

$$E_{p^n(x^n)} d((X^n)_i, (g_n \circ f_n(X^n))_i)$$

Note that the total distortion of the code (f_n, g_n) is the average of its coordinate distortions.

Lemma 3. *In the setting of lemma 2, It is possible to add the requirement that for each distortion measure d_j, the distortion of all the coordinates is the same.*

Proof. We will achieve the additional demand by making the code symmetric. Start with the code that satisfies the requirements of lemma 2. For each $\hat{x}^n$ in the code add all its cyclic permutations to the codebook. This will enlarge the codebook by factor n at most, and thus does not change the asymptotic rate. Let σ be any cyclic permutation, note that $(x^n, \hat{x}^n) \in A_\epsilon^{(n)}$ implies $(\sigma(x^n), \sigma(\hat{x}^n)) \in A_\epsilon^{(n)}$. Hence it is possible to change the encoding such that the following hold:

$$\sigma(g_n \circ f_n(x^n)) = g_n \circ f_n(\sigma(x^n)) \tag{10}$$

without sacrificing the average distortion. Fix a distortion measure d_j and let i be one of the coordinates of the code. For any cyclic permutation σ the following holds:

$$\begin{aligned}
E_{p^n(x^n)} d((X^n)_i, (g_n \circ f_n(X^n))_i) &= \sum_{x^n} p(x^n) d((X^n)_i, (g_n \circ f_n(X^n))_i) \\
&= \sum_{x^n} p(\sigma(x^n)) d((\sigma(X^n))_i, (g_n \circ f_n(\sigma(X^n)))_i) \\
&= \sum_{x^n} p(x^n) d((X^n)_{\sigma(i)}, (g_n \circ f_n(X^n))_{\sigma(i)}) \\
&= E_{p^n(x^n)} d((X^n)_{\sigma(i)}, (g_n \circ f_n(X^n))_{\sigma(i)})
\end{aligned}$$

These equalities follows from equation (10) and since $p(x^n) = p(\sigma(x^n))$. □

We are now ready to complete the proof of theorem 2.

Proof (Achievability in theorem 2). It suffices to show that for any joint distribution $p(\hat{x}, x) = p(\hat{x}|x)p(x)$ it is possible to construct a series of codes (f_n, g_n) with asymptotic rate $I(X; \hat{X})$ and asymptotic Y-information $I(Y; \hat{X})$.

Let define for any pair $(x_0, \hat{x}_0)$ a distortion measure as follows:

$$d = d_{x_0,\hat{x}_0}(x, \hat{x}) = \begin{cases} 1 \text{ if } (x_0, \hat{x}_0) = (x, \hat{x}) \\ 0 \text{ otherwise} \end{cases}$$

Then we have:

$$E_{p(x,\hat{x})} d_{x_0,\hat{x}_0}(X, \hat{X}) = p(x_0, \hat{x}_0) \tag{11}$$

and

$$E_{p(x^n)} d_{x_0,\hat{x}_0}(X^n, g_n \circ f_n(X^n)) = \frac{1}{n} \sum_{i=1}^{n} \bar{p}_i(x_0, \hat{x}_0) \tag{12}$$

where $\bar{p}_i$ are as defined in (1).

Using these distortion measures, the construction in lemma 3 and equations (11) and (12) we build a series of codes (f_n, g_n) with asymptotic rate $I(X; \hat{X})$ such that for large enough n, any $(x_0, \hat{x}_0)$ and any i

$$|\bar{p}_i(x_0, \hat{x}_0) - p(x_0, \hat{x}_0)| < \epsilon$$

Since the Y-information of the code is continuous function with respect to the $\bar{p}_i$ it follows that the Y-information convergence to $I(Y; \hat{X})$. □

5 Properties of the IB-curve

In this section we study the IB-function $R^{(I)}(I_Y)$ and present some properties it poses. We use the similarity to the WAK problem (see section 3) to adopt results from [13].

First note that $R^{(I)}(I_Y)$ is the minimum of $I(X;\hat{X})$ over decreasingly smaller sets as I_Y increases. Thus $R^{(I)}(I_Y)$ is non-decreasing function of I_Y.

Second, note that in contrary to the rate-distortion scenario, where the elements of $\hat{\mathcal{X}}$ get meaning from the distortion measure, in our scenario only the size of $\hat{\mathcal{X}}$ matter. It is clear that if $|\hat{\mathcal{X}}| < |\mathcal{X}|$ the solution may not be optimal. However the following lemma shows that $|\hat{\mathcal{X}}|$ does not have to be much bigger.

Lemma 4. *$\hat{\mathcal{X}}$ of cardinality $|\mathcal{X}|+2$ is sufficient to achieve the optimal $I(X;\hat{X})$ for any constraint I_Y on $I(\hat{X};Y)$*

And for this case we also have convexity:

Lemma 5. *For $|\hat{\mathcal{X}}| \geq |\mathcal{X}|+2$, the IB-function $R^{(I)}(I_Y)$ is a convex function of I_Y and $I_Y^{(I)}(R)$ is concave function of R.*

The above two lemmas were proved in [13]. Note that [13] prove it for a slightly different setting, but the modifications are straightforward.

The curve is continuous in the interior since it is monotonic and convex. Moreover it is smooth under mild conditions as stated in the following lemma.

Lemma 6. *For a $p(x,y) > 0$ the slope of $R^{(I)}(I_Y)$ is continuous and approaches ∞ as I_Y approaches $I(X;Y)$*

Proof. Let (I_Y^*, R^*) be a point on the curve and let $p^*(\hat{x}|x)$ be a distribution that achieves this point. From convexity we know that there is a straight line that pass through (I_Y^*, R^*) such that all the curve lies in the upper half space defined by this line. Denote by β the slope of this line. Then

$$R - \beta I_Y \geq R^* - \beta I_Y^* \qquad \forall\ (I_Y, R) \tag{13}$$

Thus p^* is optimal and has the following exponential form:

$$p^*(\hat{x}|x) = \frac{p^*(\hat{x})}{Z(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p^*(y|\hat{x})]} \tag{14}$$

Assume that there is a line with slope $\beta' \neq \beta$ with the same properties (i.e. that the curve is not smooth at this point). Then it is also true that:

$$p^*(\hat{x}|x) = \frac{p^*(\hat{x})}{Z(x,\beta')} e^{-\beta' D_{KL}[p(y|x)||p^*(y|\hat{x})]} \tag{15}$$

From (14) and (15) and the fact that D_{KL} is finite for $p(x,y) > 0$ we have that for every $\hat{x}$ with $p(\hat{x}) > 0$:

$$\frac{Z(x,\beta)}{Z(x,\beta')} = e^{(\beta-\beta') D_{KL}[p(y|x)||p^*(y|\hat{x})]}$$

The left-hand side is independent of $\hat{x}$ and thus the right-hand side must as well. It follow then that $p(\hat{x}|x) = p(\hat{x})$ and therefore $I(X;\hat{X}) = 0$. □

It is easy to see that the assumption that there are no zeros in $p(x, y)$ in the above lemma is necessary, for example the curve correspond to the following block matrix has a constant finite slope.

$$p(x,y) = \frac{1}{8}\begin{pmatrix} 1\,1\,0\,0 \\ 1\,1\,0\,0 \\ 0\,0\,1\,1 \\ 0\,0\,1\,1 \end{pmatrix}$$

Corollary 1. *From equation (13) it follows that any optimal solution is a global minimum*[2] *of the Lagrangian $I(X;\hat{X}) - \beta I(\hat{X};Y)$ and β is the slope of $R^{(I)}(I_Y)$ in that point.*

Lemma 7. *The IB function $R^{(I)}(I_Y)$ is continuous as a function of $p(x, y)$.*

The proof follows from the continuity of both mutual information and $R^{(I)}(I_Y)$ (as a function of I_Y).

6 Information Bottleneck and Rate Distortion

In this section we show that the Information Bottleneck is locally equivalent to a rate-distortion problem (RDT) [7] with adequate distortion measure and can be considered as a RDT with a non fixed distortion measure. We also show that the information curve is the envelope of all these "locally equivalent" RDT curves.

Definition 6. *For given $p(y|x)$ and $p(y|\hat{x})$ we define the following distortion measure on $\mathcal{X} \times \hat{\mathcal{X}}$:*

$$d_{IB}(x, \hat{x}) = D_{KL}[p(y|x)||p(y|\hat{x})]$$

Lemma 8. *For fixed $p(\hat{x}|x)$ and $p(y|\hat{x}) = \sum_x p(y|x)p(x|\hat{x})$*

$$\langle d_{IB} \rangle_{x,\hat{x}} = I(X;Y) - I(\hat{X};Y)$$

The proof of lemma 8 follows from the Markov chain $\hat{X} \to X \to Y$ and simple algebric manipulation.
Now define the "rate-distortion like" minimization problem

$$R_{DT}(D) = \min_{p(\hat{x}|x)\ :\ \langle d_{IB} \rangle \leq D} I(\hat{X};X) \tag{16}$$

and from the lemma we have that $R_{DT}(D) = R^{(I)}(I(X;Y) - D)$. Note that $R_{DT}(D)$ is not a rate-distortion problem as the "distortion measure" d_{IB} is not

[2] note that generally when using Lagrange multipliers, the optimum can be any point where the gradient of the Lagrangian vanishes.

fixed. Where "not fixed" means that it depends on the minimization parameter $p(\hat{x}|x)$. The dependency is as follows:

$$d_{IB}(x,\hat{x}) = D_{KL}[p(y|x)||p(y|\hat{x})] \tag{17}$$
$$= D_{KL}\left[p(y|x)||\frac{1}{p(\hat{x})}\sum_{x'} p(y|x')p(\hat{x}|x')p(x')\right] \tag{18}$$

Lemma 9. *Let $D \geq 0$. For a conditional distribution $q(y|\hat{x})$ define the rate-distortion problem:*

$$R(q,D) = \min_{p(\hat{x}|x)\ :\ \sum_{x,\hat{x}} p(\hat{x}|x)p(x)D_{KL}[p(y|x)||q(y|\hat{x})]\leq D} I(X;\hat{X}) \tag{19}$$

Then

$$R^{(I)}(I(X;Y)-D) = \min_{q(y|\hat{x})} R(q,D)$$

The proof of lemma 9 is omitted due to space limitation.

Corollary 2. *For any $p(x|\hat{x})$, let $p(y|\hat{x}) = \sum_{x'} p(y|x')p(x'|\hat{x})$ and consider the rate-distortion problems given by the source X and the distortion measure $d(x,\hat{x}) = D_{KL}[p(y|x)|p(y|\hat{x})]$. Then the following hold:*

- *The IB-curve (with switched x-axis) is below the rate-distortion curve.*
- *If the $p(x|\hat{x})$ is correspond to a point on the IB-curve, the rate-distortion curve is tangent to the IB-curve at this point.*
- *The IB-curve is a tight lower bound ("envelope") for all the rate-distortion curves of the above form.*

A demonstration of corollary 2 is given in figure 3.

7 Information Bottleneck and Cost Capacity

In the previous section we have shown that the information bottleneck can be considered as a rate-distortion problem with a non-fixed distortion measure. In this section we consider a dual representation. This representation takes the form of a cost-capacity problem with a non-fixed cost function. We show that these two dual problems are indeed equivalent.

Definition 7. *Given a noisy channel $p(b|a)$ and a cost function $c : A \to \mathbb{R}^+$, let $e(p) = \sum_a c(a)p(a)$ then the* cost-capacity function *is defined as*

$$C(E) = \max_{p(a)\ :\ e(p)\leq E} I(A;B)$$

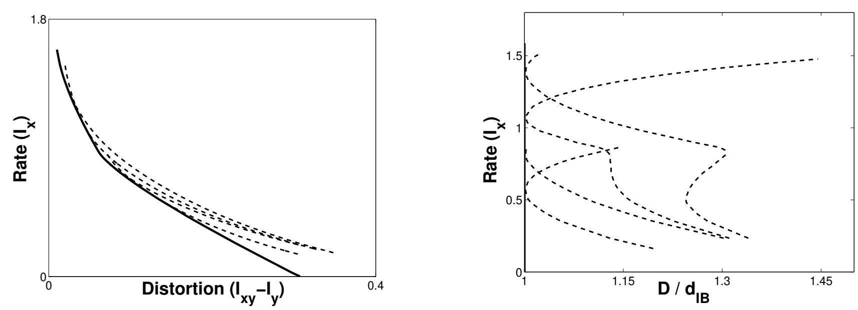


Fig. 3. IB-curve as an envelope of rate-distortion curves for a random 4x4 matrix. The bold line is the IB curve with switched x-axis. The dashed lines are the rate distortion curves for the same source, with different distortion measures induced from substituting different $p(y|\hat{x})$ in the D_{KL}. The right figure shows the normalized graph, i.e. each curve was divided by the IB curve.

Consider $p(y|\hat{x})$ as defining a noisy channel and define a cost function

$$c(\hat{x}) = \sum_x p(x|\hat{x}) \log \frac{p(x|\hat{x})}{p(x)}$$

then $$e(p) = \sum_{\hat{x}} c(\hat{x}) p(\hat{x}) = \sum_{x,\hat{x}} p(x|\hat{x}) p(\hat{x}) \log \frac{p(x|\hat{x})}{p(x)} = I_p(X;\hat{X})$$

and $$C(R) = \max_{p(\hat{x}|x)\,:\,e(p)\leq R} I(\hat{X};Y) = \max_{p(\hat{x}|x)\,:\,I(\hat{X};X)\leq R} I(\hat{X};Y) = I_Y^{(I)}(R)$$

Note that $p(\hat{x}|x)$ defines $p(\hat{x})$ as $p(x)$ is fixed and thus it is sound to maximize over $p(\hat{x}|x)$ instead of over $p(\hat{x})$.

Next we show that this maximization problem is really dual to our original minimization problem, i.e. that the two optimization problems defines the same curve.

Lemma 10. *The two optimization problems $R^{(I)}(I_Y)$ and $I_Y^{(I)}(R)$ defines the same curve with switched axes for $0 \leq I_Y \leq I(X;Y)$ and $0 \leq R \leq R^{(I)}(I(X;Y))$*

Proof. We have to prove that for any I_Y, $R^{(I)}(I_Y) = I_Y^{(I)}(R)$ for some R and vice-verse, i.e. for any R there exists I_Y such that $I_Y^{(I)}(R) = R^{(I)}(I_Y)$. For this purpose it suffices to show that the following two properties hold:

- for any relevant I_Y, it is not possible to achieve $I(\hat{X};Y)$ larger than I_Y with $I(X;\hat{X})$ equal to or smaller than $R^{(I)}(I_Y)$
- for any relevant R, it is not possible to achieve $I(\hat{X};X)$ smaller than R with $I(\hat{X};Y)$ equal to or larger than $I_Y^{(I)}(R)$

The first property is clear from the convexity of $R^{(I)}(I_Y)$ and the obvious fact that it is impossible to achieve $I(\hat{X};Y) > 0$ with $I(X;\hat{X}) = 0$.

The second property follows from the concavity of $I_Y^{(I)}(R)$ as follows: if it is possible to move down from any $(I_Y^{(I)}(R), R)$ point (i.e. to achieve a smaller $I(X,\hat{X})$ with $I(\hat{X};Y) = I_Y^{(I)}(R)$) then concavity is possible only if $I_Y^{(I)}(R)$ is the maximal possible value of $I(\hat{X};Y)$, and this possible only for $R \geq R^{(I)}(I(X;Y))$. $\square$

8 Conclusions and Further Research

In this paper we provided a rigorous formulation of the Information Bottleneck method as a fumdamental information theoretic question: what is the lower bound on the rate of a code that preserves mutual information on another variable. This method was successfully applied for finding efficient representations for numerous learning problems for which co-occurance distribution of two variables can be estimated, so-far without proper information theoretic justification. We showed that the IB method indeed solves a natural information theoretic problem, formally related to source coding with side information. In a well defined sense this problem unifies aspects of Rate-Distortion Theory and Cost-Capacity tradeoff, but in both cases the effective distortion and the channel cost function are simultaneously determined from the joint statistics of the data, given a single tradeoff parameter. We proved that given the joint distribution, there is a tight achievable convex and smooth bound on the representation length of one variable, for a given mutual information that this representation maintains on the other variable. Since the problem is continuous as a function of the joint distribution (lemma 7), we expect that calculating the bound using a sampled version of the joint distribution can give a good approximation. We also showed that the representation cardinality should not exceed - by more than 2 - the cardinality of the original variable.

Finding a compact representation is a crucial known component in learning (Occam razor, MDL, etc.). In this work we used information theoretic tools in order to quantify the quality of representations, when the accuracy is measured by mutual information as well. A natural extention to these ideas should connect our bounds to generalization error bounds, more common in learning theory.

Many related interesting issues are left outside of this paper, such as analytic properties of the IB-curve for specific joint distributions, or simpler conditions that ensure its convexity and smoothness. We know that for some interesting classes of joint distributions there are analytic expressions for this curve, but for general distributions finding this curve can be computationally very difficult. Some most intriguing remaining questions are: (i) What is the optimal compexity-accuarcy tradeoff for bounded computational complexity? (ii) What is the nature of the deviations from the optimal when given only a finite sample from the joint distribution $p(x,y)$ (the sample complexity - over-fitting - probelm)? (iii) Is there a similar coding theoretic formulation for the multivariate

IB, which is in fact a network information theoretic tradeoff?

Acknowledgment: We would like to thank Eyal Krupka and Noam Slonim for many invaluable discussions and ideas. RB is supported by the Clore foundation. AN is supported by the Horowitz foundation. This work is partly supported by a grant from the Israel Science Foundation.

References

1. R. F. Ahlswede and J. Korner. Source coding with side information and a converse for degraded broadcast channels. *IEEE transaction on information theory*, 21(6):629–637, November 1975.
2. Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. Submitted for publication.
3. J. Cardinal. Compression of side information. In *IEEE International Conference on Multimedia and Expo*, 2003.
4. T. M. Cover and J. A. Thomas. *Elements Of Information Thory*. Wiley Interscience, 1991.
5. J. Kleinberg. An impossibility theorem for clustering. In *Proc. of the 16th conference on Neural Information Processing Systems*, 2002.
6. P. Poupart and C. Boutilier. Value-directed compression of pomdps. In *Proc. of the 16th conference on Neural Information Processing Systems*, 2002.
7. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, July and October 1948.
8. N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, The Hebrew University, 2002.
9. N. Slonim and Tishby N. The power of word clustering for text classification. In *Proc. of the 23rd European Colloquium on Information Retrieval Research*, 2001.
10. N. Slonim, R. Somerville, N. Tishby, and O. Lahav. Objective classification of galaxy spectra using the information bottleneck method. *Monthly Notes of the Royal Astronomical Society*, 323:270–284, 2001.
11. N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *Proc. of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2000.
12. N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Comunnication, Control and Computing*, pages 368–377, 1999.
13. H. S. Witsenhausen and A. D. Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE transaction on information theory*, 21(5):493–501, September 1975.
14. A. D. Wyner. On source coding with side information at the decoder. *IEEE transaction on information theory*, 21(3):294–300, May 1975.