The information bottleneck method

Naftali Tishby,[1,2] Fernando C. Pereira,[3] and William Bialek[1]

[1]NEC Research Institute, 4 Independence Way
Princeton, New Jersey 08540
[2]Institute for Computer Science, and
Center for Neural Computation
Hebrew University
Jerusalem 91904, Israel
[3]AT&T Shannon Laboratory
180 Park Avenue
Florham Park, New Jersey 07932

30 September 1999

We define the relevant information in a signal $x \in X$ as being the information that this signal provides about another signal $y \in Y$. Examples include the information that face images provide about the names of the people portrayed, or the information that speech sounds provide about the words spoken. Understanding the signal x requires more than just predicting y, it also requires specifying which features of X play a role in the prediction. We formalize this problem as that of finding a short code for X that preserves the maximum information about Y. That is, we squeeze the information that X provides about Y through a 'bottleneck' formed by a limited set of codewords $\tilde{X}$. This constrained optimization problem can be seen as a generalization of rate distortion theory in which the distortion measure $d(x, \tilde{x})$ emerges from the joint statistics of X and Y. This approach yields an exact set of self consistent equations for the coding rules $X \to \tilde{X}$ and $\tilde{X} \to Y$. Solutions to these equations can be found by a convergent re–estimation method that generalizes the Blahut–Arimoto algorithm. Our variational principle provides a surprisingly rich framework for discussing a variety of problems in signal processing and learning, as will be described in detail elsewhere.

1 Introduction

A fundamental problem in formalizing our intuitive ideas about information is to provide a quantitative notion of "meaningful" or "relevant" information. These issues were intentionally left out of information theory in its original formulation by Shannon, who focused attention on the problem of transmitting information rather than judging its value to the recipient. Correspondingly, information theory has often been viewed as being strictly a theory of communication, and this view has become so accepted that many people consider statistical and information theoretic principles as almost irrelevant for the question of meaning. In contrast, we argue here that information theory, in particular lossy source compression, provides a natural quantitative approach to the question of "relevant information." Specifically, we formulate a variational principle for the extraction or efficient representation of relevant information. In related work [1] we argue that this single information theoretic principle contains as special cases a wide variety of problems, including prediction, filtering, and learning in its various forms.

The problem of extracting a relevant summary of data, a compressed description that captures only the relevant or meaningful information, is not well posed without a suitable definition of relevance. A typical example is that of speech compression. One can consider lossless compression, but in any compression beyond the entropy of speech some components of the signal cannot be reconstructed. On the other hand, a transcript of the spoken words has much lower entropy (by orders of magnitude) than the acoustic waveform, which means that it is possible to compress (much) further without losing any information about the words and their meaning.

The standard analysis of lossy source compression is "rate distortion theory," which characterizes the tradeoff between the rate, or signal representation size, and the average distortion of the reconstructed signal. Rate distortion theory determines the level of inevitable expected distortion, D, given the desired information rate, R, in terms of the *rate distortion function* $R(D)$. The main problem with rate distortion theory is in the need to specify the distortion function first, which in turn determines the relevant features of the signal. Those features, however, are often not explicitly known and

an arbitrary choice of the distortion function is in fact an arbitrary feature selection.

In the speech example, we have at best very partial knowledge of what precise components of the signal are perceived by our (neural) speech recognition system. Those relevant components depend not only on the complex structure of the auditory nervous system, but also on the sounds and utterances to which we are exposed during our early life. It therefore is virtually impossible to come up with the "correct" distortion function for acoustic signals. The same type of difficulty exists in many similar problems, such as natural language processing, bioinformatics (for example, what features of protein sequences determine their structure) or neural coding (what information is encoded by spike trains and how). This is the fundamental problem of feature selection in pattern recognition. Rate distortion theory does not provide a full answer to this problem since the choice of the distortion function, which determines the relevant features, is not part of the theory. It is, however, a step in the right direction.

A possible solution comes from the fact that in many interesting cases we have access to an additional variable that determines what is relevant. In the speech case it might be the transcription of the signal, if we are interested in the speech recognition problem, or it might be the speaker's identity if speaker identification is our goal. For natural language processing, it might be the part of speech labels for words in grammar checking, but the dictionary senses of ambiguous words in information retrieval. Similarly, for the protein folding problem we have a joint database of sequences and three dimensional structures, and for neural coding a simultaneous recording of sensory stimuli and neural responses defines implicitly the relevant variables in each domain. All of these problems have the same formal underlying structure: extract the information from one variable that is relevant for the prediction of another one. The choice of additional variable determines the relevant components or features of the signal in each case.

In this short paper we formalize this intuitive idea using an information theoretic approach which extends elements of rate distortion theory. We derive self consistent equations and an iterative algorithm for finding representations of the signal that capture its relevant structure, and prove

convergence of this algorithm.

2 Relevant quantization

Let X denote the signal (message) space with a fixed probability measure $p(x)$, and let $\tilde{X}$ denote its quantized codebook or compressed representation. For ease of exposition we assume here that both of these sets are finite, that is, a continuous space should first be quantized.

For each value $x \in X$ we seek a possibly stochastic mapping to a representative, or codeword in a codebook, $\tilde{x} \in \tilde{X}$, characterized by a conditional p.d.f. $p(\tilde{x}|x)$. The mapping $p(\tilde{x}|x)$ induces a soft partitioning of X in which each block is associated with one of the codebook elements $\tilde{x} \in \tilde{X}$, with probability given by

$$p(\tilde{x}) = \sum_x p(x)p(\tilde{x}|x) \ . \tag{1}$$

The average volume of the elements of X that are mapped to the same codeword is $2^{H(X|\tilde{X})}$, where

$$H(X|\tilde{X}) = -\sum_{x\in X} p(x) \sum_{\tilde{x}\in\tilde{X}} p(\tilde{x}|x) \log p(\tilde{x}|x) \tag{2}$$

is the conditional entropy of X given $\tilde{X}$.

What determines the quality of a quantization? The first factor is of course the rate, or the average number of bits per message needed to specify an element in the codebook without confusion. This number *per element of* X is bounded from below by the mutual information

$$I(X;\tilde{X}) = \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x}) \log\left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right], \tag{3}$$

since the average cardinality of the partitioning of X is given by the ratio of the volume of X to that of the mean partition, $2^{H(X)}/2^{H(X|\tilde{X})} = 2^{I(X;\tilde{X})}$, via the standard asymptotic arguments. Notice that this quantity is different from the entropy of the codebook, $H(\tilde{X})$, and this entropy normally is not what we want to minimize.

However, information rate alone is not enough to characterize good quantization since the rate can always be reduced by throwing away details of the original signal x. We need therefore some additional constraints.

2.1 Relevance through distortion: Rate distortion theory

In rate distortion theory such a constraint is provided through a distortion function, $d : X \times \tilde{X} \to \mathcal{R}^+$, which is presumed to be small for good representations. Thus the distortion function specifies implicitly what are the most relevant aspects of values in X.

The partitioning of X induced by the mapping $p(\tilde{x}|x)$ has an expected distortion

$$\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x})d(x,\tilde{x}) \ . \tag{4}$$

There is a monotonic tradeoff between the rate of the quantization and the expected distortion: the larger the rate, the smaller is the achievable distortion.

The celebrated rate distortion theorem of Shannon and Kolmogorov (see, for example Ref. [2]) characterizes this tradeoff through the rate distortion function, $R(D)$, defined as the minimal achievable rate under a given constraint on the expected distortion:

$$R(D) \equiv \min_{\{p(\tilde{x}|x):\langle d(x,\tilde{x})\rangle\le D\}} I(X;\tilde{X}) \ . \tag{5}$$

Finding the rate distortion function is a variational problem that can be solved by introducing a Lagrange multiplier, β, for the constrained expected distortion. One then needs to minimize the functional

$$\mathcal{F}[p(\tilde{x}|x)] = I(X;\tilde{X}) + \beta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} \tag{6}$$

over all normalized distributions $p(\tilde{x}|x)$. This variational formulation has the following well known consequences:

Theorem 1 *The solution of the variational problem,*

$$\frac{\delta \mathcal{F}}{\delta p(\tilde{x}|x)} = 0, \tag{7}$$

for normalized distributions $p(\tilde{x}|x)$, is given by the exponential form

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left[-\beta d(x,\tilde{x})\right] \ , \tag{8}$$

where $Z(x,\beta)$ is a normalization (partition) function. Moreover, the Lagrange multiplier β, determined by the value of the expected distortion, D, is positive and satisfies

$$\frac{\delta R}{\delta D} = -\beta \ . \tag{9}$$

Proof. Taking the derivative w.r.t. $p(\tilde{x}|x)$, for given x and $\tilde{x}$, one obtains

$$\begin{aligned} \frac{\delta \mathcal{F}}{\delta p(\tilde{x}|x)} &= p(x)\Bigg[\log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + 1 \\ &\qquad - \frac{1}{p(\tilde{x})} \sum_{x'} p(x')p(\tilde{x}|x') + \beta d(x,\tilde{x}) + \frac{\lambda(x)}{p(x)}\Bigg], \end{aligned} \tag{10}$$

since the marginal distribution satisfies $p(\tilde{x}) = \sum_{x'} p(x')p(\tilde{x}|x')$. $\lambda(x)$ are the normalization Lagrange multipliers for each x. Setting the derivatives to zero and writing $\log Z(x,\beta) = \lambda(x)/p(x)$, we obtain Eq. (8). When varying the normalized $p(\tilde{x}|x)$, the variations $\delta I(X;\tilde{X})$ and $\delta \langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}$ are linked through

$$\delta \mathcal{F} = \delta I(X;\tilde{X}) + \beta \delta \langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = 0, \tag{11}$$

from which Eq. (9) follows. The positivity of β is then a consequence of the concavity of the rate distortion function (see, for example, Chapter 13 of Ref. [2]). □

2.2 The Blahut–Arimoto algorithm

An important practical consequence of the above variational formulation is that it provides a converging iterative algorithm for self consistent determination of the distributions $p(\tilde{x}|x)$ and $p(\tilde{x})$.

Equations (8) and (1) must be satisfied simultaneously for consistent probability assignment. A natural approach to solve these equations is to alternately iterate between them until reaching convergence. The following lemma, due to Csiszár and Tusnády [3], assures global convergence in this case.

Lemma 2 *Let $p(x,y) = p(x)p(y|x)$ be a given joint distribution. Then the distribution $q(y)$ that minimizes the relative entropy or Kullback–Leibler divergence, $D_{KL}[p(x,y)|p(x)q(y)]$, is the marginal distribution*

$$p(y) = \sum_x p(x)p(y|x).$$

Namely,

$$I(X;Y) = D_{KL}[p(x,y)|p(x)p(y)] = \min_{q(y)} D_{KL}[p(x,y)|p(x)q(y)] \ .$$

Equivalently, the distribution $q(y)$ which minimizes the expected relative entropy,

$$\sum_x p(x) D_{KL}[p(y|x)|q(y)],$$

is also the marginal distribution $p(y) = \sum_x p(x)p(y|x)$.

The proof follows directly from the non–negativity of the relative entropy.

This lemma guarantees the marginal condition Eq. (1) through the same variational principle that leads to Eq. (8):

Theorem 3 *Equations (1) and (8) are satisfied simultaneously at the minimum of the functional,*

$$\mathcal{F} = -\langle \log Z(x,\beta) \rangle_{p(x)} = I(X;\tilde{X}) + \beta \langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})} \ , \tag{12}$$

where the minimization is done independently *over the convex sets of the normalized distributions, $\{p(\tilde{x})\}$ and $\{p(\tilde{x}|x)\}$,*

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \mathcal{F}\left[p(\tilde{x}); p(\tilde{x}|x)\right] \ .$$

These independent conditions correspond precisely to alternating iterations of Eq. (1) and Eq. (8). Denoting by t the iteration step,

$$\begin{cases} p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta d(x,\tilde{x})) \end{cases} \tag{13}$$

where the normalization function $Z_t(x,\beta)$ is evaluated for every t in Eq. (13). Furthermore, these iterations converge to a unique minimum of $\mathcal{F}$ in the convex sets of the two distributions.

For the proof, see references [2, 4]. This alternating iteration is the well known Blauht-Arimoto (BA) algorithm for calculation of the rate distortion function.

It is important to notice that the BA algorithm deals only with the optimal partitioning of the set X given the set of representatives $\tilde{X}$, and not with an optimal choice of the representation $\tilde{X}$. In practice, for finite data, it is also important to find the optimal representatives which minimize the expected distortion, *given* the partitioning. This joint optimization is similar to the EM procedure in statistical estimation and does not in general have a unique solution.

3 Relevance through another variable: The Information Bottleneck

Since the "right" distortion measure is rarely available, the problem of relevant quantization has to be addressed directly, by preserving the *relevant information* about another variable. The relevance variable, denoted here by Y, must not be independent from the original signal X, namely they have positive mutual information $I(X;Y)$. It is assumed here that we have access to the joint distribution $p(x,y)$, which is part of the setup of the problem, similarly to $p(x)$ in the rate distortion case.[1]

[1]The problem of actually obtaining a good enough sample of this distribution is an interesting issue in learning theory, but is beyond the scope of this paper. For a start on this problem see Ref. [1].

3.1 A new variational principle

As before, we would like our relevant quantization $\tilde{X}$ to compress X as much as possible. In contrast to the rate distortion problem, however, we now want this quantization to capture as much of the information about Y as possible. The amount of information about Y in $\tilde{X}$ is given by

$$I(\tilde{X};Y) = \sum_y \sum_{\tilde{x}} p(y,\tilde{x}) \log \frac{p(y,\tilde{x})}{p(y)p(\tilde{x})} \leq I(X;Y). \tag{14}$$

Obviously lossy compression cannot convey more information than the original data. As with rate and distortion, there is a tradeoff between compressing the representation and preserving meaningful information, and there is no single right solution for the tradeoff. The assignment we are looking for is the one that keeps a fixed amount of meaningful information about the relevant signal Y while minimizing the number of bits from the original signal X (maximizing the compression). [2] In effect we pass the information that X provides about Y through a "bottleneck" formed by the compact summaries in $\tilde{X}$.

We can find the optimal assignment by minimizing the functional

$$\mathcal{L}[p(\tilde{x}|x)] = I(\tilde{X};X) - \beta I(\tilde{X};Y), \tag{15}$$

where β is the Lagrange multiplier attached to the constrained meaningful information, while maintaining the normalization of the mapping $p(\tilde{x}|x)$ for every x. At $\beta = 0$ our quantization is the most sketchy possible—everything is assigned to a single point—while as $\beta \to \infty$ we are pushed toward arbitrarily detailed quantization. By varying the (only) parameter β one can explore the tradeoff between the preserved meaningful information and compression at various resolutions. As we show elsewhere [1, 5], for interesting special cases (where there exist sufficient statistics) it is possible to preserve almost all the meaningful information at finite β with a significant compression of the original data.

[2] It is completely equivalent to maximize the meaningful information for a fixed compression of the original variable.

3.2 Self-consistent equations

Unlike the case of rate distortion theory, here the constraint on the meaningful information is *nonlinear* in the desired mapping $p(\tilde{x}|x)$ and this is a much harder variational problem. Perhaps surprisingly, this general problem of extracting the meaningful information—minimizing the functional $\mathcal{L}[p(\tilde{x}|x)]$ in Eq. (15)—can be given an exact formal solution.

Theorem 4 *The optimal assignment, that minimizes Eq. (15), satisfies the equation*

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left[-\beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})}\right], \tag{16}$$

where the distribution $p(y|\tilde{x})$ in the exponent is given via Bayes' rule and the Markov chain condition $\tilde{X} \leftarrow X \leftarrow Y$, as,

$$p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(y|x)p(\tilde{x}|x)p(x). \tag{17}$$

This solution has a number of interesting features, but we must emphasize that it is a *formal* solution since $p(y|\tilde{x})$ in the exponential is defined implicitly in terms of the assignment mapping $p(\tilde{x}|x)$. Just as before, the marginal distribution $p(\tilde{x})$ must satisfy the marginal condition Eq. (1) for consistency.

Proof. First we note that the conditional distribution of y on $\tilde{x}$

$$p(y|\tilde{x}) = \sum_{x \in X} p(y|x)p(x|\tilde{x}) \ , \tag{18}$$

follows from the Markov chain condition $Y \leftarrow X \leftarrow \tilde{X}$.[3] The only variational variables in this scheme are the conditional distributions, $p(\tilde{x}|x)$, since other unknown distributions are determined from it through Bayes' rule and consistency. Thus we have

$$p(\tilde{x}) = \sum_x p(\tilde{x}|x)p(x) \ , \tag{19}$$

[3]It is important to notice that this not a modeling assumption and the quantization $\tilde{X}$ is *not* used as a hidden variable in a model of the data. In the latter, the Markov condition would have been different: $Y \leftarrow \tilde{X} \leftarrow X$.

and

$$p(\tilde{x}|y) = \sum_x p(\tilde{x}|x)p(x|y) \ . \tag{20}$$

The above equations imply the following derivatives w.r.t. $p(\tilde{x}|x)$,

$$\frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} = p(x) \tag{21}$$

and

$$\frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} = p(x|y) \ . \tag{22}$$

Introducing Lagrange multipliers, β for the information constraint and $\lambda(x)$ for the normalization of the conditional distributions $p(\tilde{x}|x)$ at each x, the Lagrangian, Eq. (15), becomes

$$\begin{aligned} \mathcal{L} &= I(X,\tilde{X}) - \beta I(\tilde{X},Y) - \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) & (23) \\ &= \sum_{x,\tilde{x}} p(\tilde{x}|x)p(x) \log\left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right] - \beta \sum_{\tilde{x},y} p(\tilde{x},y) \log\left[\frac{p(\tilde{x}|y)}{p(\tilde{x})}\right] \\ &\qquad - \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) \ . & (24) \end{aligned}$$

Taking derivatives with respect to $p(\tilde{x}|x)$ for given x and $\tilde{x}$, one obtains

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta p(\tilde{x}|x)} &= p(x)\left[1 + \log p(\tilde{x}|x)\right] - \frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)}\left[1 + \log p(\tilde{x})\right] \\ &\qquad -\beta \sum_y \frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} p(y)\left[1 + \log p(\tilde{x}|y)\right] \\ &\qquad -\beta \frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)}\left[1 + \log p(\tilde{x})\right] - \lambda(x) \ . & (25) \end{aligned}$$

Substituting the derivatives from Eq's. (21) and (22) and rearranging,

$$\frac{\delta \mathcal{L}}{\delta p(\tilde{x}|x)} = p(x)\left\{\log\left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right] - \beta \sum_y p(y|x) \log\left[\frac{p(y|\tilde{x})}{p(y)}\right] - \frac{\lambda(x)}{p(x)}\right\} \ . \tag{26}$$

Notice that $\sum_y p(y|x) \log \frac{p(y|x)}{p(y)} = I(x,Y)$ is a function of x only (independent of $\tilde{x}$) and thus can be absorbed into the multiplier $\lambda(x)$. Introducing

$$\tilde{\lambda}(x) = \frac{\lambda(x)}{p(x)} - \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y)} \right] \ ,$$

we finally obtain the variational condition:

$$\frac{\delta \mathcal{L}}{\delta p(\tilde{x}|x)} = p(x) \left[\log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})} - \tilde{\lambda}(x) \right] = 0 \ , \quad (27)$$

which is equivalent to equation (16) for $p(\tilde{x}|x)$,

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left(-\beta D_{KL}\left[p(y|x)|p(y|\tilde{x})\right]\right) \ , \quad (28)$$

with

$$Z(x,\beta) = \exp[\beta \tilde{\lambda}(x)] = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(-\beta D_{KL}\left[p(y|x)|p(y|\tilde{x})\right]\right) \ ,$$

the normalization (partition) function. □

Comments:

1. The Kullback–Leibler divergence, $D_{KL}[p(y|x)|p(y|\tilde{x})]$, *emerged* as the relevant "effective distortion measure" from our variational principle but is not assumed otherwise anywhere! It is therefore natural to consider it as the "correct" distortion $d(x,\tilde{x}) = D_{KL}[p(y|x)|p(y|\tilde{x})]$ for quantization in the information bottleneck setting.

2. Equation (28), together with equations (18) and (19), determine self consistently the desired conditional distributions $p(\tilde{x}|x)$ and $p(\tilde{x})$. The crucial quantization is here performed through the conditional distributions $p(y|\tilde{x})$, and the self consistent equations include also the optimization over the representatives, in contrast to rate distortion theory, where the selection of representatives is a separate problem.

3.3 The information bottleneck iterative algorithm

As for the BA algorithm, the self consistent equations (16) and (17) suggest a natural method for finding the unknown distributions, at every value of β. Indeed, these equations can be turned into converging, alternating iterations among the three convex distribution sets, $\{p(\tilde{x}|x)\}$, $\{p(\tilde{x})\}$, and $\{p(y|\tilde{x})\}$, as stated in the following theorem.

Theorem 5 *The self consistent equations (18), (19), and (28), are satisfied simultaneously at the minima of the functional,*

$$\begin{aligned} \mathcal{F}\left[p(\tilde{x}|x); p(\tilde{x}); p(y|\tilde{x})\right] &= -\langle \log Z(x,\beta) \rangle_{p(x)} & (29) \\ &= I(X;\tilde{X}) + \beta \langle D_{KL}[p(y|x)|p(y|\tilde{x})] \rangle_{p(x,\tilde{x})} & (30) \end{aligned}$$

where the minimization is done independently *over the convex sets of the normalized distributions, $\{p(\tilde{x})\}$ and $\{p(\tilde{x}|x)\}$ and $\{p(y|\tilde{x})\}$. Namely,*

$$\min_{p(y|\tilde{x})} \min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \mathcal{F}\left[p(\tilde{x}|x); p(\tilde{x}); p(y|\tilde{x})\right] \ .$$

This minimization is performed by the converging alternating iterations. Denoting by t the iteration step,

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta d(x,\tilde{x})) \\ p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_{t+1}(y|\tilde{x}) = \sum_y p(y|x) p_t(x|\tilde{x}) \end{cases} \tag{31}$$

and the normalization (partition function) $Z_t(\beta,\tilde{x})$ is evaluated for every t in Eq. (31).

Proof. For lack of space we can only outline the proof. First we show that the equations indeed are satisfied at the minima of the functional $\mathcal{F}$ (known for physicists as the "free energy"). This follows from lemma (2) when applied to $I(X;\tilde{X})$ with the convex sets of $p(\tilde{x})$ and $p(\tilde{x}|x)$, as for the BA algorithm. Then the second part of the lemma is applied to $\langle D_{KL}[p(y|x)|p(y|\tilde{x})] \rangle_{p(x,\tilde{x})}$ which is an expected relative entropy. Equation (28) minimizes the expected relative entropy w.r.t. to variations in the convex set of the normalized

$\{p(y|\tilde{x})\}$. Denoting by $d(x,\tilde{x}) = D_{KL}[p(y|x)|p(y|\tilde{x})]$ and by $\lambda(\tilde{x})$ the normalization Lagrange multipliers, we obtain

$$\delta d(x,\tilde{x}) = \delta\left(-\sum_y p(y|x)\log p(y|\tilde{x}) + \lambda(\tilde{x})(\sum_y p(y|\tilde{x}) - 1)\right) \quad (32)$$

$$= \sum_y \left(-\frac{p(y|x)}{p(y|\tilde{x})} + \lambda(\tilde{x})\right)\delta p(y|\tilde{x}) \ . \quad (33)$$

The expected relative entropy becomes,

$$\sum_x \sum_y \left(-\frac{p(y|x)p(x|\tilde{x})}{p(y|\tilde{x})} + \lambda(\tilde{x})\right)\delta p(y|\tilde{x}) = 0 \ , \quad (34)$$

which gives Eq. (28), since $\delta p(y|\tilde{x})$ are independent for each $\tilde{x}$. Equation (28) also have the interpretation of a weighted average of the data conditional distributions that contribute to the representative $\tilde{x}$.

To prove the convergence of the iterations it is enough to verify that each of the iteration steps minimizes the same functional, independently, and that this functional is bounded from below as a sum of two non–negative terms. The only point to notice is that when $p(y|\tilde{x})$ is fixed we are back to the rate distortion case with fixed distortion matrix $d(x,\tilde{x})$. The argument in [3] for the BA algorithm applies here as well. On the other hand we have just shown that the third equation minimizes the expected relative entropy without affecting the mutual information $I(X;\tilde{X})$. This proves the convergence of the alternating iterations. However, the situation here is similar to the EM algorithm and the functional $\mathcal{F}[p(\tilde{x}|x);p(\tilde{x});p(y|\tilde{x})]$ is convex in each of the distribution independently but *not* in the product space of these distributions. Thus our convergence proof does not imply uniqueness of the solution. □

3.4 The structure of the solutions

The formal solution of the self consistent equations, described above, still requires a specification of the structure and cardinality of $\tilde{X}$, as in rate distortion theory. For every value of the Lagrange multiplier β there are corresponding values of the mutual information $I_X \equiv I(X,\tilde{X})$, and $I_Y \equiv$

$I(\tilde{X}, Y)$ for every choice of the cardinality of $\tilde{X}$. The variational principle implies that

$$\frac{\delta I(\tilde{X}, Y)}{\delta I(X, \tilde{X})} = \beta^{-1} > 0 \ , \tag{35}$$

which suggests a *deterministic annealing* approach. By increasing the value of β one can move along *convex* curves in the "information plane" (I_X, I_Y). These curves, analogous to the rate distortion curves, exists for every choice of the cardinality of $\tilde{X}$. The solutions of the self consistent equations thus correspond to a family of such annealing curves, all starting from the (trivial) point $(0, 0)$ in the information plane with infinite slope and parameterized by β. Interestingly, every two curves in this family separate (bifurcate) at some finite (critical) β through a second order phase transition. These transitions form a hierarchy of relevant quantizations for different cardinalities of $\tilde{X}$, as described in [1, 5, 6].

Further work

The most fascinating aspect of the information bottleneck principle is that it provides a unified framework for different information processing problems, including prediction, filtering and learning [1]. There are already several successful applications of this method to various "real" problems, such as semantic clustering of English words [6], document classification [5], neural coding, and spectral analysis.

Acknowledgements

Helpful discussions and insights on rate distortion theory with Joachim Buhmann and Shai Fine are greatly appreciated. Our collaboration was facilitated in part by a grant from the US–Israel Binational Science Foundation (BSF).

References

[1] W. Bialek and N. Tishby, "Extracting relevant information," in preparation.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[3] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and Decisions* **Suppl. 1,** 205–237 (1984).

[4] R. E. Blahut, "Computation of channel capacity and rate distortion function," *IEEE Trans. Inform. Theory* **IT-18,** 460–473 (1972).

[5] N. Slonim and N. Tishby, "Agglomerative information bottleneck," To appear in *Advances in Neural Information Processing systems (NIPS-12)* 1999.

[6] F. C. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *30th Annual Mtg. of the Association for Computational Linguistics*, pp. 183–190 (1993).