Extracting Meaningful Information: from Rate Distortion Theory to the Information Bottleneck Method

30592 Research Project

Samuele Nicolò Straccialini

Abstract

When dealing with signals, we can define the relevant information of a signal $x \in X$ as the information provided about a second signal $y \in Y$. Understanding what x really is means not only correctly predicting y, but also specifying the features of X that played a role in the prediction. In this sense, we can aim to find a short code for X which preserves the maximum information about Y.

The Information Bottleneck (IB) Method, proposed by Tishby et al. (2000) [1], is a general framework for extracting relevant information from an input variable x to obtain a target variable y by balancing compression and prediction. The underlying idea is that we can "squeeze" the information that X provides to Y, reducing it to a limited set of codewords $\tilde{X}$. This approach extends concepts from information theory, and in particular can be viewed as an extension of the rate-distortion theory.

This method has broad applications in machine learning, clustering, and neural network analysis, offering a principled way to trade off information preservation and reduction. In this report, we provide an overview of the IB method, discussing its theoretical foundations, key formulations, and practical applications.

1 Introduction

Extracting a meaningful representation of data is a key challenge in machine learning and statistical inference. While Shannon, in the original formulation of information theory concepts, focuses more on the problem of transmitting information rather than assessing its value, Tishby et al. (2000) [1] suggest that information theory provides a natural approach to the question of identifying the "relevant" information, especially when dealing with lossy compression. This is a technique to reduce data size using approximations, to achieve a given fidelity measure: the reconstructed data will be equal to the original signal up to a certain threshold.

Determining "relevant" information is strictly related to the actual definition of relevance. For instance, consider an audio recording containing speech over background music. Depending on the task, different aspects of the signal can be relevant: linguistic content is sufficient for speech recognition; if the goal is to analyze the background music, the spoken words become irrelevant, and only the musical features should be retained; for waveform reconstruction, preserving the full high-fidelity signal might be necessary.

The standard approach to lossy compression is through rate-distortion theory, which balances compression and fidelity. Mathematically, this tradeoff is characterized by the rate-distortion function $R(D)$, which measures the lowest possible bit rate R needed to encode a signal, while ensuring that the average distortion D remains below a given threshold level of distortion. The main problem with this approach is that the distortion function, which determines the relevant features of the signal, must be specified in advance. These features, however, are often unknown a priori. A potential solution is accessing some additional information (for example, a transcription of the input in speech recognition, or a database of sequences and 3D structures in the protein folding problem); this is, however, not always possible.

In the following paragraphs, we will use an information-theoretic approach to extend rate-distortion theory and derive an iterative algorithm for finding representations of the signal that capture its essential structure. Then, we will also explore some possible applications in neural networks.

2 Quantization and Rate Distortion Theory

In class, we introduced the problem of quantization in the case of a continuous random source whose signals need to be reproduced using a finite-rate code. In general, quantization refers to mapping some input signal x from the space X with a fixed probability measure $p(x)$ to a signal $\tilde{x}$ from a smaller space $\tilde{X}$ characterized by the conditional p.d.f. $p(\tilde{x}|x)$. Thus, we have

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

The average volume of elements of X which are mapped to the same codeword $\tilde{x}$ is

$$2^{H(X|\tilde{X})}, \tag{2}$$

where $H(X|\tilde{X})$ is the conditional entropy of X given $\tilde{X}$:

$$H(X|\tilde{X}) = \sum_{x\in X} p(x) \sum_{\tilde{x}\in\tilde{X}} p(\tilde{x}|x) \log p(\tilde{x}|x) \tag{3}$$

By the standard asymptotic arguments, the average cardinality of partitioning X is given by the ratio of its volume to that of the mean partition:

$$\frac{2^{H(X)}}{2^{H(X|\tilde{X})}} = 2^{I(X;\tilde{X})} \tag{4}$$

where $I(X;\tilde{X})$ is the mutual information between X and $\tilde{X}$:

$$I(X;\tilde{X}) = \sum_{x\in X} \sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x}) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} \tag{5}$$

However, information rate alone is not enough to define a good quantization: we need another constraint that, in classical rate-distortion theory, is provided by the distortion function.

Definition 1. *A distortion function or distortion measure $d : X \times \tilde{X} \to \mathbb{R}^+$ is a mapping from the set of source alphabet-reproduction alphabet pairs into the set of nonnegative real numbers.*

The distortion function measures the cost of compressing the source signal x with its representation $\tilde{x}$. The expected distortion is then:

$$\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})] = \langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = \sum_{x\in X} \sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x})d(x,\tilde{x}). \tag{6}$$

The tradeoff between the rate of quantization and the expected distortion is monotonic: the larger the rate, the smaller the achievable distortion. This is formalized by the following theorem about the rate distortion function $R(D)$, which is the minimum number of bits per data unit needed to represent the data so that the expected distortion doesn't exceed some threshold D.

Theorem 1 (Shannon-Kolmogorov)**.** *The rate distortion function $R(D)$ for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x,\tilde{x})$ is equal to the associated information rate distortion function:*

$$R(D) = \min_{p(\tilde{x}|x):\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})]<D} I(X;\tilde{X}) \tag{7}$$

This is a variational problem. To solve it, we can introduce a Lagrange multiplier β and minimize the functional

$$\mathcal{F}[p(\tilde{x}|x)] = I(X;\tilde{X}) + \beta\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})] \tag{8}$$

Theorem 2. *The solution of the variational problem $\frac{\delta\mathcal{F}[p(\tilde{x}|x)]}{\delta p(\tilde{x}|x)}$ for normalized distributions $p(\tilde{x}|x)$ is given by:*

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta d(x,\tilde{x})}, \tag{9}$$

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

A complete proof of the above theorem can be found in Appendix A.

Notice that equations (1) and (9) must hold simultaneously. A natural approach to solve these is to alternate between them until reaching convergence, which is assured by the following Lemma.

Lemma 1 (Csiszár-Tusnády)**.** *For a given joint distribution $p(x,y) = p(y|x)p(x)$, the distribution $q(y)$ which minimizes the relative entropy $D_{KL}[p(x,y)||p(x)q(y)]$ is the marginal distribution*

$$p(y) = \sum_x p(x)p(y|x). \tag{10}$$

This means that

$$\min_{q(y)} D_{KL}[p(x,y)||p(x)q(y)] = D_{KL}[p(x,y)||p(x)p(y)] = I(X;Y) \tag{11}$$

This Lemma follows from the non-negativity of the relative entropy. A detailed proof is in Appendix A.

The alternative iteration for calculating the rate distortion function is known as Blauht-Arimoto (BA).

Theorem 3 (BA)**.** *Equations (1) and (9) are satisfied simultaneously at the minimum of*

$$\mathcal{F} = -\mathbb{E}_x[\log Z(x,\beta)] = I(X;\tilde{X}) + \beta \mathbb{E}_{x,\tilde{x}} d(x,\tilde{x}), \tag{12}$$

where minimization is done independently over $\{p(\tilde{x})\}$ and $\{p(\tilde{x}|x)\}$:

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \mathcal{F}[p(\tilde{x}); p(\tilde{x}|x)] \tag{13}$$

This corresponds to alternating iterations over the two equations. Denoting t the iteration step,

$$\begin{cases} p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta d(x,\tilde{x})) \end{cases} \tag{14}$$

Notice that the BA algorithm optimizes the partitioning of X given the set $\tilde{X}$ of representatives, and not the choice of the representation $\tilde{X}$ itself. In practice, it is also crucial to find the optimal representatives that minimize the expected distortion during partitioning.

3 The Information Bottleneck

Finding the "right" distortion measure to appropriately capture the cost of compression is a challenging task. Instead, we can preserve the relevant information about another variable Y, which should not be independent from the original signal X (ensuring that $I(X;Y) \geq 0$). Analogously as before, where we assumed (1), we now assume the joint distribution $p(x,y)$ is known.

Our goal is again to compress X in $\tilde{X}$ as much as possible. This time, however, we capture as much of the information about Y as possible:

$$I(\tilde{X};Y) = \sum_{y \in Y} \sum_{\tilde{x} \in \tilde{X}} p(y,\tilde{x}) \log \frac{p(y,\tilde{x})}{p(y)p(\tilde{x})} \leq I(X;Y), \tag{15}$$

where the inequality arises because a lossy compression cannot give more information about Y than the original data. Similarly to rate and distortion, there's a tradeoff between compression and preserving meaningful information. We now aim to find the assignment that retains a fixed amount of relevant information about the signal Y while minimizing the bits of the original X (maximizing the compression). Essentially, we are passing the information that X provides about Y through a "bottleneck" created by the compressed representation $\tilde{X}$. Similarly to the previous situation, the optimal assignment is found by minimizing the functional

$$\mathcal{L}[p(\tilde{x}|x)] = I(\tilde{X};X) - \beta I(\tilde{X};Y), \tag{16}$$

where again, β is the Lagrange multiplier. Notice that at $\beta = 0$, we have the most extreme quantization, with all information shrunk to a single point. For $\beta > 0$ we have increasingly detailed quantization (so varying β allows us to explore the tradeoff between preserved information and compression).

Differently to the rate distortion case, equation (16) is now nonlinear, due to the nonlinearity of the constraint. Surprisingly, it is still possible to obtain an exact solution.

Theorem 4. *The optimal assignment minimizing equation (16) satisfies*

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left[-\beta \sum_{y\in Y} p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})}\right] \tag{17}$$

where $p(y|\tilde{x})$ in the exponent can be found as

$$p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(y|x)p(\tilde{x}|x)p(x). \tag{18}$$

More details about the proof of the above theorem are in Appendix A.

Remark. In Theorem 4, the optimal "distortion measure" in the bottleneck setting is revealed to be the following relative entropy (have a look at equation A3):

$$d(x,\tilde{x}) = D_{KL}[p(y|x)||p(y|\tilde{x})]. \tag{19}$$

This KL divergence is not assumed anywhere else, and is therefore natural to be considered as the "correct" distortion for quantization in the information bottleneck setting.

As for the BA algorithm, equations (17) and (18), together with the fact that the optimal distortion metric in this setting is (19), can be used to find the unknown distributions over $\{p(\tilde{x}|x)\}$, $\{p(\tilde{x})\}$, and $\{p(y|\tilde{x})\}$, as stated in the following theorem.

Theorem 5. *Equations (17) and (18) are satisfied simultaneously at the minimum of*

$$\mathcal{F} = -\mathbb{E}_x[\log Z(x,\beta)] = I(X;\tilde{X}) + \beta \mathbb{E}_{x,\tilde{x}}\Big[D_{KL}[p(y|x)||p(y|\tilde{x})]\Big], \tag{20}$$

where minimization is done independently over $\{p(\tilde{x})\}$, $\{p(\tilde{x}|x)\}$, and $p(y|\tilde{x})$:

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \min_{p(y|\tilde{x})} \mathcal{F}[p(\tilde{x}); p(\tilde{x}|x); p(y|\tilde{x})] \tag{21}$$

This corresponds to alternating iterations over three equations. Denoting t the iteration step,

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta D_{KL}[p(y|x)||p(y|\tilde{x})]) \\ p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_{t+1}(y|\tilde{x}) = \sum_y p(y|x) p_t(x|\tilde{x}) \end{cases} \tag{22}$$

The solution described above still requires specifying the structure and the cardinality of $\tilde{X}$. In particular, for every value of the Lagrange multiplier β, there are corresponding values of the mutual information $I(X;\tilde{X})$ and $I(\tilde{X},Y)$, which satisfy $\frac{\delta I(\tilde{X};Y)}{\delta I(X,\tilde{X})} = \beta^{-1} > 0$. These can be found using a deterministic annealing approach. Thus, solutions to equations correspond to a family of annealing curves starting from the point $(0,0)$ and parameterized by β. An outline for the proof of Theorem 5 is again in Appendix A.

4 Applications

The information bottleneck principle provides a unified framework for addressing various problems, from clustering to document classification or spectral analysis. A particularly intriguing application is perhaps to Deep Neural Networks (DNNs). In a paper by Tishby and Zaslavsky (2015) [2], the authors analyze DNNs in the framework of the IB principle to determine the optimal information-theoretic limits of a DNN and derive finite-sample bounds for the generalization error. This is a measure of the difference between the performance of the DNN on training data and unseen data, indicating how well the network can generalize to new samples.

The goal of DNNs (and supervised learning technique in general) is to capture and efficiently represent the relevant information in the input variable about the output variable. This is nothing more than finding a maximally compressed mapping of the input that preserves the information on the output as much as possible, which aligns precisely with the IB method's goal.

4.1 Information of layers

Consider the simplest DNN model as shown in Figure 1. Each layer processes only the information received from the previous one. So, for $i \geq j$, the following holds

$$I(Y;X) \geq I(Y;\mathbf{h}_j) \geq I(Y;\mathbf{h}_i) \geq I(Y;\hat{Y}). \tag{23}$$

This makes sense since it means that the mutual information between the input and the true label is greater than or equal to the mutual information between the true label and a deeper layer in the network, which in turn is greater than or equal to the mutual information between the true and the predicted label. This means we are progressively losing information. Each layer should aim to provide the most concise and relevant input representation by maximizing $I(Y;\mathbf{h}_i)$ while minimizing $I(\mathbf{h}_{i-1};\mathbf{h}_i)$ as much as possible. Reducing $I(\mathbf{h}_{i+1},\mathbf{h}_i)$ can be interpreted as obtaining the minimal description length of the layer, while $I(Y;\hat{Y})$ serves as a natural quantifier for the quality of the DNN.

In this framework, we can use the IB principle as a measure of optimality not only for the output layer but also for evaluating the optimality of each hidden layer. Namely, each layer can be compared to the optimal IB limit for some β:

$$I(\mathbf{h}_{i+1};\mathbf{h}_i) + \beta I(Y;\mathbf{h}_{i+1}|\mathbf{h}_i), \tag{24}$$

where $\mathbf{h}_0 = X$ and $\mathbf{h}_{m+1} = \hat{Y}$. This framework is more advantageous than other cost functions, which are generally not applicable for evaluating the optimality of the hidden layers. The input level clearly has the least IB distortion and requires the longest description. Then, each subsequent layer can only increase the IB distortion level, but at the same time, it will compress its inputs, hopefully eliminating only irrelevant information.

4.2 Generalization Bounds

In machine learning, we are interested in finding a joint distribution $p(X,Y)$, which we know only for a limited number of (training) samples, and we try to generalize it. However, representations that encode more information than necessary for prediction may end up fitting noise or irrelevant structure in the input data, causing overfitting.

It has been shown that IB principle can serve as a learning objective from finite samples. In particular, it introduces a tradeoff between the amount of information the representation $\hat{X}$ retains about the input – $I(X;\hat{X})$ – and the amount of relevant information it retains for predicting the target – $I(\hat{X};Y)$. The objective is to find a stochastic encoder $p(\hat{x}|x)$ that minimizes the IB Lagrangian:

$$\mathcal{L}_{IB} = I(X;\hat{X}) - \beta I(\hat{X};Y) \tag{25}$$

The parameter $\beta > 0$ in (25) balances between compression and prediction performance. This interpretation leads to the insight that compression acts as a form of regularization and has similar effects to classical regularization

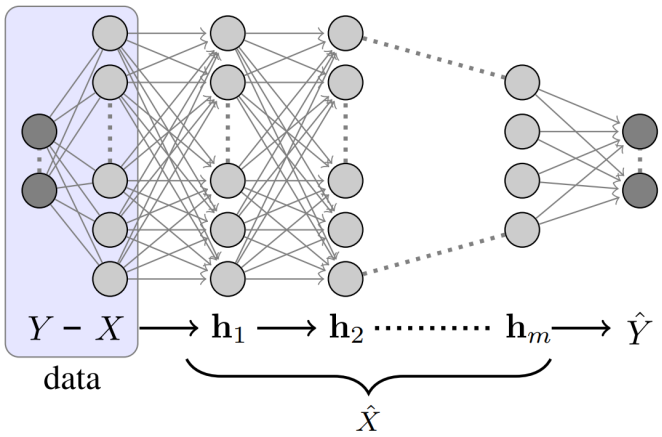

Figure 1: Simple Neural Network model. An n-dimensional input $x \in X$, where each x is associated to a true label $y \in Y$, is sent through m layers $\mathbf{h}_i$ ($i \in \{1,\ldots,m\}$). The network outputs a predicted label $\hat{y} \in \hat{Y}$, aiming to obtain the true label.

techniques, such as weight decay or early stopping, but emerges naturally from the IB objective. The complexity of the representation is measured by $I(X;\hat{X})$, and the higher this value is, the higher the risk of overfitting.

The generalization bounds found are as follows:

$$I(\hat{X};Y) \leq \hat{I}(X,Y) + \mathcal{O}\left(\frac{K|Y|}{\sqrt{n}}\right), \tag{26}$$

$$I(X;\hat{X}) \leq \hat{I}(X,\hat{X}) + \mathcal{O}\left(\frac{K}{\sqrt{n}}\right), \tag{27}$$

where $\hat{I}$ the empirical estimate of the mutual information based on the finite sample distribution $\hat{p}(x,y)$, n is the sample size, and $K := |\hat{X}|$. Notice that these bounds get worse with K, but do not depend on the cardinality of X. This means that the mutual information is well estimated for learning compressed representations (small K), and is badly estimated for learning complex models (large K). The complexity of the representation $K = |\hat{X}|$, i.e. the cardinality of the support of $\hat{X}$ is about its effective description length: $K \approx 2^{I(X;\hat{X})}$. This provides a continuous worst-case upper bound on the true $I(\hat{X};Y)$ for any sample size n. The take-home message we obtain is that compression helps generalization, and inserting more layers or neurons without appropriate compression won't help if one has limited data.

References

[1] Naftali Tishby, Fernando C. Pereira, and William Bialek. *The information bottleneck method*. 2000. DOI: 10.48550/ARXIV.PHYSICS/0004057. URL: https://arxiv.org/abs/physics/0004057.

[2] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*. Jerusalem, Israel: IEEE, Apr. 2015, pp. 1–5. ISBN: 9781479955244 9781479955268. DOI: 10.1109/ITW.2015.7133169. URL: http://ieeexplore.ieee.org/document/7133169/.

A Appendix: Proofs

In this appendix, I will provide some proofs of the theorems stated in the report for the sake of completeness.

Proof of Theorem 2.

We can expand (8) and add the normalization constraint on $p(\tilde{x}|x)$ as:

$$\begin{aligned}\mathcal{F} &= \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x}) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + \beta \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x}) d(x,\tilde{x}) + \sum_{x\in X}\lambda(x)\left(\sum_{\tilde{x}\in\tilde{X}} p(\tilde{x}|x) - 1\right) \\ &= \sum_{x,\tilde{x}}\left[p(x)p(\tilde{x}|x)\log\frac{p(\tilde{x}|x)}{p(\tilde{x})} + \beta p(x)p(\tilde{x}|x)d(x,\tilde{x})\right] + \sum_{x\in X}\lambda(x)\left(\sum_{\tilde{x}\in\tilde{X}} p(\tilde{x}|x) - 1\right).\end{aligned}$$

Taking the derivative of with respect to $p(\tilde{x}|x)$ we find:

$$\frac{\delta\mathcal{F}}{\delta p(\tilde{x}|x)} = p(x)\log\frac{p(\tilde{x}|x)}{p(\tilde{x})} + p(x) + \beta p(x) d(x,\tilde{x}) + \lambda(x) = 0,$$

implying

$$\log p(\tilde{x}|x) = \log p(\tilde{x}) - 1 - \beta d(x,\tilde{x}) - \frac{\lambda(x)}{p(x)} \implies p(\tilde{x}|x) = p(\tilde{x}) e^{-1-\beta d(x,\tilde{x}) - \frac{\lambda(x)}{p(x)}}.$$

Normalizing,

$$\sum_{\tilde{x}\in\tilde{X}} p(\tilde{x}|x) = 1 \implies e^{-\frac{\lambda(x)}{p(x)}} = \frac{1}{\sum_{\tilde{x}} p(\tilde{x}) e^{-1-\beta d(x,\tilde{x})}} = \frac{1}{e^{-1}}\frac{1}{\sum_{\tilde{x}} p(\tilde{x}) e^{-\beta d(x,\tilde{x})}}$$

Finally,

$$p(\tilde{x}|x) = \frac{e^{-1}p(\tilde{x})e^{-\beta d(x,\tilde{x})}}{e^{-1}\sum_{\tilde{x}} p(\tilde{x})e^{-\beta d(x,\tilde{x})}} = \frac{p(\tilde{x})e^{-\beta d(x,\tilde{x})}}{\sum_{\tilde{x}} p(\tilde{x})e^{-\beta d(x,\tilde{x})}} =: \frac{p(\tilde{x})}{Z(x,\beta)}e^{-\beta d(x,\tilde{x})}$$

□

Proof of Lemma 1.

We want to show that:

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

Notice that

$$\begin{aligned} D_{KL}[p(x,y)||p(x)q(y)] &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)q(y)} = \sum_{x,y} p(x,y)\log\frac{p(x)p(y|x)}{p(x)q(y)} = \sum_{x,y} p(x,y)\log\frac{p(y|x)}{q(y)} \\ &= \sum_{y} p(y)\sum_{x}\frac{p(x,y)}{p(y)}\log\frac{p(y|x)}{q(y)} = \sum_{y} p(y)\sum_{x} p(x|y)\log\frac{p(y|x)}{q(y)} \end{aligned}$$

Recall Jensen inequality for convex functions f: $\sum_i p_i f(x_i) \geq f\left(\sum_i p_i x_i\right)$. Thus,

$$\begin{aligned} \sum_{y} p(y)\sum_{x} p(x|y)\log\frac{p(y|x)}{q(y)} &\geq \sum_{y} p(y)\log\left(\sum_{x} p(x|y)\frac{p(y|x)}{q(y)}\right) = \sum_{y} p(y)\log\left(\sum_{x} p(x|y)\frac{p(y|x)}{q(y)}\right) \\ &= \sum_{y} p(y)\log\left(\sum_{x}\frac{p(x,y)^2}{p(x)p(y)}\frac{1}{q(y)}\right) = \sum_{y} p(y)\log\left(\sum_{x}\frac{p(x,y)}{q(y)}\right) \\ &= \sum_{y} p(y)\log\left(\frac{p(y)}{q(y)}\right) = D_{KL}(p(y)||q(y)) \end{aligned}$$

Recall that KL divergence $D_{KL}(p(y)||q(y))$ is always non-negative, and equals zero if and only if $q(y) = p(y)$. So, the minimum of $D_{KL}[p(x,y)||p(x)q(y)] \geq D_{KL}(p(y)||q(y))$ is obtained for $q^*(y) = p(y)$. □

Proof of Theorem 4.

First, we can rewrite the distributions as follows:

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

$$p(\tilde{x}) = \sum_{x} p(\tilde{x}|x)p(x), \tag{A1}$$

$$p(\tilde{x}|y) = \sum_{x} p(\tilde{x}|x)p(x|y). \tag{A2}$$

In particular, equations (A1) and (A2) imply the following derivatives with respect to $p(\tilde{x}|x)$:

$$\frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} = p(x), \qquad \frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} = p(x|y).$$

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, we get

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

Now, taking derivatives with respect to $p(\tilde{x}|x)$:

$$\begin{aligned}
\frac{\delta\mathcal{L}}{\delta p(\tilde{x}|x)} =& p(x)\left[1+\log p(\tilde{x}|x)\right] - \frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)}\left[1+\log p(\tilde{x})\right] \\
&- \beta\sum_y \frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} p(y)\left[1+\log p(\tilde{x}|y)\right] - \beta\frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)}\left[1+\log p(\tilde{x})\right] - \lambda(x) \\
=& p(x)\log\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right) - \beta\left(\sum_y p(x|y)p(y)[1+\log p(\tilde{x}|y)] + p(x)[1+\log p(\tilde{x})]\right) - \lambda(x) \\
=& p(x)\log\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right) - \beta\left(\sum_y p(x|y)p(y)[1+\log p(\tilde{x}|y)] + \sum_y p(x|y)p(y)[1+\log p(\tilde{x})]\right) - \lambda(x) \\
=& p(x)\left[\log\left(\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right) - \beta\sum_y p(y|x)\log\left(\frac{p(y|x)}{p(y|\tilde{x})}\right) - \frac{\lambda(x)}{p(x)}\right] = 0.
\end{aligned}$$

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

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

we finally obtain the variational condition:

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

which is equivalent to

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)}\exp\left(-\beta D_{KL}[p(y|x)||p(y|\tilde{x})]\right), \tag{A3}$$

with

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

the normalization (partition) function. □

Outline of the proof of Theorem 5.

First, we can show that the equations are indeed satisfied at the minima of the functional $\mathcal{F}$. This follows from Lemma 1 when applied to $I(X;\tilde{X})$ with the convex sets of $p(\tilde{x})$ and $\tilde{p}(\tilde{x}|x)$, as for the BA algorithm. The second part of the Lemma is then applied to $\langle D_{KL}[p(y|x)p(y|\tilde{x})]\rangle_{p(x,\tilde{x})}$, which is an expected relative entropy. 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

$$\begin{aligned}
\delta d(x,\tilde{x}) &= \delta\left(-\sum_y p(y|x)\log p(y|x) + \lambda(\tilde{x})\left(\sum_y p(y|x) - 1\right)\right) \\
&= \sum_y\left(\frac{p(y|x)}{p(y|\tilde{x})} + \lambda(\tilde{x})\right)\delta p(y|\tilde{x})
\end{aligned}$$

The expected relative entropy becomes,

$$\sum_x\sum_y\left(-\frac{p(y|x)p(x|\tilde{x})}{p(y|\tilde{x})} + \lambda(\tilde{x})\right)\delta p(y|\tilde{x}) = 0, \tag{A4}$$

which gives Eq. (A3), since $\delta p(y|\tilde{x})$ are independent for each $\tilde{x}$. Equation (A3) also has 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. Notice that when $p(y|\tilde{x})$ is fixed, we are back to the rate distortion case with fixed distortion matrix $d(x, \tilde{x})$, so it follows from the convergence of the BA algorithm. 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. Note that $\mathcal{F}$ is convex in each of the distributions independently, but not in the product space of these distributions. Thus, our convergence proof does not imply uniqueness of the solution. □