

The Information Bottleneck Revisited or How to Choose a Good Distortion Measure

Peter Harremoës

Centrum voor Wiskunde en Informatica
P.O. 94079, 1090 GB Amsterdam
The Netherlands
P.Harremoës@cwi.nl

Naftali Tishby

School of Engineering and Computer Science
The Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

Abstract—It is well-known that the information bottleneck method and rate distortion theory are related. Here it is described how the information bottleneck can be considered as rate distortion theory for a family of probability measures where information divergence is used as distortion measure. It is shown that the information bottleneck method has some properties that are not shared with rate distortion theory based on any other divergence measure. In this sense the information bottleneck method is unique.

I. INTRODUCTION

For many problems in the theory of lossy compression it is difficult to specify a distortion measure. In some cases we want to compress a variable X but what we are really interested in is not the value of X but the value of a variable Y correlated or coupled with X . We are only interested in retrieving the value of X to the extent that it gives information about Y . Thus, each value of X gives a distribution on Y . The information bottleneck method was introduced in [1] to solve this problem. It has always been known that the information bottleneck method is related to rate distortion theory. In this paper we shall explore the relation in detail. The information bottleneck method has found natural interpretations and a number of applications as described in [2], [3], [4], [5], [6], [7], [8]. The results in these papers do not rule out the possibility that similar results could have been obtained by other means (other distortion measure). Our approach will be via rate distortion theory and as we shall see we are lead directly to the information bottleneck via a number of assumptions or simplifications that efficiently rule out all other distortion or divergence measures than information divergence.

Let \mathbb{A} and \mathbb{B} be alphabets and let X and Y be random variables with values in \mathbb{A} and \mathbb{B} : For simplicity we shall assume that \mathbb{A} and \mathbb{B} are finite but most of the results in this paper holds for infinite sets as well. The joint distribution of X and Y is given and given by a distribution on \mathbb{A} and a Markov kernel $\Phi : \mathbb{A} \rightarrow M_+^1(\mathbb{B})$, where $M_+^1(\mathbb{B})$ denotes the set of probability measures on \mathbb{B} . As reconstruction alphabet we use $\hat{\mathbb{A}} = M_+^1(\mathbb{B})$.

$$X \longrightarrow Y$$

As distortion measure $d : \mathbb{A} \times \hat{\mathbb{A}} \rightarrow \mathbb{R}$ we use

$$d(x, \hat{x}) = D(\Phi(x), \hat{x}) \quad (1)$$

where D denotes some divergence measure on $M_+^1(\mathbb{B})$. Our goal is to minimize both the rate $I(X, \hat{X})$ and the distortion $E(d(X, \hat{X}))$ over all joint distribution of $(X, \hat{X}) \in \mathbb{A} \times \hat{\mathbb{A}}$ with prescribed marginal distribution of X . The trade-off between rate and distortion is given by the rate distortion curve. To find the point on the rate distortion curve with slope $-\beta$ one should minimize

$$I(X, \hat{X}) + \beta \cdot E[d(X, \hat{X})].$$

The most important divergence measure is *information divergence* (or Kullback-Leibler information or relative entropy) defined by

$$D(P\|Q) = \sum_{y \in \mathbb{B}} \log \left(\frac{p_i}{q_i} \right) p_i.$$

We know from Sanov's Theorem [9] that the difficulty in distinguishing a distribution P from a distribution Q by a statistical test is given by the information divergence. This suggests to use information divergence as distortion measure. We shall now formalize these ideas.

Information divergence belong to a class of divergence measures called *Csiszár f -divergences* defined by

$$D_f(P, Q) = \sum_{i \in \mathbb{B}} f \left(\frac{p_i}{q_i} \right) p_i$$

where f is denotes a convex function satisfying $f(1) = 0$ [10], [11]. For $f(x) = x \log x$ we get information divergence. For $f(x) = (x-1)^2$ we get χ^2 -divergence. For $f(x) = (x^{1/2} - 1)^2$ we get Hellinger divergence. for $f(x) = |x-1|$ we get variational distance.

Information divergence also belongs to the class of *Bregman divergences* [12]. For a finite output alphabet a Bregman divergence on $M_+(\mathbb{B})$ is defined by

$$B_f(P, Q) = f(P) - (f(Q) + (P - Q) \cdot \nabla f(Q))$$

where $f : M_+(\mathbb{B}) \rightarrow \mathbb{R}$ is some convex function. Information divergence is obtained when

$$f(P) = \sum_{i \in \mathbb{B}} p_i \log p_i.$$

Rate distortion theory with Bregman divergences is well studied in [13].

Sometimes so-called *Burbea-Rao* divergences are used. The Burbea-Rao divergence between P and Q is defined by

$$BR_f(P, Q) = \sum_{i \in \mathbb{B}} \frac{f(p_i) + f(q_i)}{2} - f\left(\frac{p_i + q_i}{2}\right)$$

for some convex function f [14].

An important class of divergence measures are the separable divergences introduced in [15]. These are divergences defined by

$$D(P, Q) = \sum_{i \in \mathbb{B}} \delta(p_i, q_i)$$

for some function $\delta : [0; 1]^2 \rightarrow \mathbb{R}$. We note that Csiszár f -divergences and Burbea-Rao divergences are separable. A separable Bregman divergence is given by

$$B_f(P, Q) = \sum_{i \in \mathbb{B}} g(p_i) - (g(q_i) + (p_i - q_i) g'(q_i))$$

for some convex function $g : [0; 1] \rightarrow \mathbb{R}$. A divergence measure is said to be reflexive if $D(P, Q) \geq 0$ with equality for $P = Q$.

For distributions P and Q close to each other all these divergence measures are approximately proportional [16] if the functions used to define them are sufficiently smooth.

II. ADDITIVITY

Rate distortion theory is most interesting when you have a rate distortion theorem. In order to get a rate distortion theorem one has to consider sequences instead of single events and one has to extend the definition of distortion from a sequence of inputs x^n in A^n to a sequence of reconstruction points \hat{x}^n by

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i),$$

see [9]. The factor $\frac{1}{n}$ is just a matter of normalization but for our problem it is essential that our divergence measure is *additive*, i.e.

$$D(P_1 \times P_2, Q_1 \times Q_2) = D(P_1, Q_2) + D(P_2, Q_2). \quad (2)$$

As the divergence on the left hand side lives on a different space than the divergences on the right hand side Equation 2 could be used to *define* the divergence on a product space and in this sense Equation 2 cannot be used to characterize divergence measures suitable for rate distortion theory. If we require that the divergence measures in the left and the right hand side of Equation 2 belong to special classes of divergence measures one can obtain non-trivial characterizations.

Theorem 1: A separable reflexive divergence measure that is additive is a linear combination of information divergence and reversed information divergence.

Proof: Let the Divergence measure D be based on $\delta :]0; 1]^2 \rightarrow \mathbb{R}$. Take $P_1 = (p_i)_{i \in \mathbb{B}}, Q_1 = (q_i)_{i \in \mathbb{B}}, P_2 =$

$(s, 1 - s)$ and $Q_2 = (1/2, 1/2)$. Then

$$\begin{aligned} \sum_{i \in \mathbb{B}} \delta(p_i, q_i) + \delta\left(s, \frac{1}{2}\right) + \delta\left(1 - s, \frac{1}{2}\right) \\ = \sum_{i \in \mathbb{B}} \delta\left(sp_i, \frac{q_i}{2}\right) + \delta\left((1 - s)p_i, \frac{q_i}{2}\right). \end{aligned}$$

By taking the second derivative at both sides we get

$$\begin{aligned} \delta''_{11}\left(s, \frac{1}{2}\right) + \delta''_{11}\left(1 - s, \frac{1}{2}\right) = \\ \sum_{i \in \mathbb{B}} \delta''_{11}\left(sp_i, \frac{q_i}{2}\right) p_i^2 + \delta''_{11}\left((1 - s)p_i, \frac{q_i}{2}\right) p_i^2. \end{aligned}$$

For $s = 1/2$ we get

$$\delta''_{11}\left(\frac{1}{2}, \frac{1}{2}\right) = \sum_{i \in \mathbb{B}} \delta''_{11}\left(\frac{p_i}{2}, \frac{q_i}{2}\right) p_i^2.$$

This equation should hold for all probability vectors P and Q which implies that $(x, y) \rightarrow \delta''_{11}(x, y) x^2$ is linear in x and y . Thus there exists constants c_1, c_2 and such that $\delta''_{11}(x, y) x^2 = c_1 x + c_2 y$ having solutions of the form

$$\delta(x, y) = c_1 x \log \frac{x}{y} + c_2 y \log \frac{y}{x} + f(y) x + g(y)$$

for some functions f, g . The first two terms gives information divergence and reversed information divergence so we just have to check additivity of a divergence based on $\tilde{\delta}(x, y) = f(y) x + g(y)$.

Taking the second derivative with respect to y leads to

$$(f''(y) x + g''(y)) y^2 = c_4 x + c_5 y.$$

in the same way as above. This should hold for all x leading to $f''(y) = c_4 y^{-2}$ and $g''(y) = c_5 y^{-1}$. The solutions are

$$f(y) = -c_4 \log y + c_6 y + c_7$$

$$g(y) = c_5 y \log y + c_8 y + c_9$$

for some constants c_6, c_7, c_8 and c_9 . Thus

$$\tilde{\delta}(x, y) = (-c_4 \log y + c_6 y + c_7) x + c_5 y \log y + c_8 y + c_9.$$

The terms that are linear in x or y may be replaced by constants without changing the divergence so we may assume that

$$\tilde{\delta}(x, y) = -c_4 x \log y + c_5 y \log y + c_7 x y + c_{10}.$$

One easily checks that the first two terms satisfies additivity and the second ones do not except for $c_7 = c_{10} = 0$. For $x = y$ we have

$$0 = \tilde{\delta}(x, x) = (c_5 - c_4) x \log x$$

which implies that $c_5 = c_4$ and $c_7 = c_{10} = 0$. Thus $\tilde{\delta}(x, y) = c_4 (y \log y - x \log y)$ and positivity is only obtained for $c = 0$. ■

III. SUFFICIENCY

The number of parameters in our rate distortion problem can often be reduced by replacing the original variables with *sufficient variables*. This idea is actually used already in formulating the model. In principle any model can always be extended by including less relevant or even irrelevant variables. Normally one would leave out these less relevant variables at an early stage, but as we shall see in the information bottle neck method one can also get rid of irrelevant variables *within* the model.

First we shall consider sufficiency on the input side. Assume that $X = (X_1, X_2)$ and that X_1 is independent of Y given X_2 . Equivalently,

$$X_1 \longrightarrow X_2 \longrightarrow Y$$

is assumed to be a Markov chain. As Y only depend on X_1 via X_2 we would like to leave out X_1 from the analysis. Thus we should compare the bottleneck problem $X \rightarrow Y$ with the bottleneck problem $X_2 \rightarrow Y$ and show that they have the same rate distortion function.

Obviously any joint distribution on (\hat{X}, X, Y) gives a joint distribution on (\hat{X}, X_1, Y) with the same mean distortion and a smaller (or equal) rate. Let a joint distribution on (\hat{X}, X_1, Y) be given where \hat{X} and Y are independent given X_1 . The joint distribution on (X_1, X_2, Y) defines a Markov kernel from X_2 to X_1 . Now consider the joint distribution on (\hat{X}, X_1, X_2, Y) where X_1, \hat{X} and Y are independent given X_2 . For this joint distribution the mean distortion is equal to the mean distortion of (\hat{X}, X_1, Y) and the rate equals

$$\begin{aligned} I(\hat{X}, X) &= I(\hat{X}, X_2) + I(\hat{X}, X_1 | X_2) \\ &= I(\hat{X}, X_2). \end{aligned}$$

We note that this sufficiency result on the input side holds for any distortion measure.

Next we shall consider sufficiency on the output side. Assume that $Y = (Y_1, Y_2)$ and that X is independent of Y_2 given Y_1 . Equivalently,

$$X \longrightarrow Y_1 \longrightarrow Y_2$$

is assumed to be a Markov chain. As Y_2 only depend on X via Y_1 we would like to leave out Y_2 from the analysis. Thus we should compare the bottleneck problem $X \rightarrow Y$ with the bottleneck problem $X \rightarrow Y_1$ and show that they have the same rate distortion function. We have

$$\begin{aligned} D(P_Y(\cdot | X) \| P_Y(\cdot | \hat{X})) \\ = \left(\begin{array}{c} D(P_{Y_1}(\cdot | X) \| P_{Y_1}(\cdot | \hat{X})) \\ + D(P_{Y_2}(\cdot | Y_1, X) \| P_{Y_2}(\cdot | Y_1, \hat{X})) \end{array} \right). \end{aligned}$$

Therefore

$$\begin{aligned} E \left[D \left(P_Y(\cdot | X) \| P_Y(\cdot | \hat{X}) \right) \right] = \\ E \left[D \left(P_{Y_1}(\cdot | X) \| P_{Y_1}(\cdot | \hat{X}) \right) \right]. \end{aligned} \quad (3)$$

Note that Equation (3) holds for any f -divergence. We shall show that it essentially only holds for f -divergences. It is easy to find examples of pairs of general divergence measures that fulfill Equation (3) for some specific joint distribution on (X, Y) but it is natural to require that the distortion measure on $M_+^1(\mathbb{B})$ should *not* depend on the joint distribution on (X, Y) .

Theorem 2: A separable divergence measure that satisfies the sufficiency condition has the form

$$D(P, Q) = \sum_{i \in \mathbb{B}} f \left(\frac{p_i}{q_i} \right) q_i. \quad (4)$$

Remark 3: The function f in Equation 4 is not necessarily convex. If f is convex and $f(1) = 0$ the divergence is a Csiszár f -divergence.

Proof: The proof of this result was essentially given as part of the proof of Theorem 1 in [17]. Their theorem states that a separable divergence satisfying a data processing inequality must be a Csiszár f -divergence. ■

Under some conditions this result can even be extended to Bregman divergences that are not separable.

IV. FINITENESS

Assume that the number of elements in \mathbb{A} is n . Let $P \in M_+^1(\mathbb{A})$ denote the marginal distribution of X . Consider distributions on $\mathbb{A} \times M_+^1(\mathbb{B})$, i.e. a joint distribution on (X, \hat{X}) . The joint distribution can be specified by the distribution Q of \hat{X} and the conditional distribution of X given \hat{X} given by a Markov kernel $\mathbb{E} : M_+^1(\mathbb{B}) \rightarrow M_+^1(\mathbb{A})$. The set of distributions on $M_+^1(\mathbb{B})$ is an infinite simplex. Consider the convex set C of distribution Q such that Q and \mathbb{E} determines joint distribution of (X, \hat{X}) has P as the marginal distribution of X . The condition that the marginal distribution of X has P as prescribed distribution gives $n - 1$ linearly independent conditions. Therefore the extreme points of C are mixtures of at most n points.

We have

$$\begin{aligned} I(X, \hat{X}) + \beta \cdot E \left[d(X, \hat{X}) \right] = \\ H(X) - H(X | \hat{X}) + \beta \cdot E \left[\mathbb{E} \left(d(X, \hat{X}) | \hat{X} \right) \right] \\ = H(X) + E \left[\beta \cdot \mathbb{E} \left(d(X, \hat{X}) | \hat{X} \right) - H(X | \hat{X}) \right]. \end{aligned}$$

Note that

$$\beta \cdot \mathbb{E} \left(d(X, \hat{X}) | \hat{X} \right) - H(X | \hat{X})$$

is a function of \hat{X} , i.e. a function $M_+^1(\mathbb{B}) \rightarrow \mathbb{R}$. Thus the minimum of $I(X, \hat{X}) + \beta \cdot E \left[d(X, \hat{X}) \right]$ is attained for an extreme point of C , i.e. the minimum is attained for a distribution on \hat{A} with support on at most n points. In the

information bottleneck literature one will normally find the result that the support has at most $n + 2$ points. Note that no particular properties of the distortion measure d have been used.

Let $\widehat{\mathbb{A}}$ denote a set with n elements. From now on we shall identify a coupling between \mathbb{B} and $M_+^1(\mathbb{A})$ with a coupling between \mathbb{B} and $\widehat{\mathbb{A}}$ together with a map $i : \widehat{\mathbb{A}} \rightarrow M_+^1(\mathbb{B})$. We know that this is no restriction for calculating the rate distortion function.

V. THE BOTTLENECK EQUATIONS

Let a joint distribution on $(X, \hat{X}) \in \mathbb{A} \times \widehat{\mathbb{A}}$ be given. This gives Markov kernel $\Psi : \widehat{\mathbb{A}} \rightarrow M_+^1(\mathbb{A})$. It can be composed with Φ to give the map $i_\Psi : \widehat{\mathbb{A}} \rightarrow M_+^1(\mathbb{B})$ defined by

$$i_\Psi(\hat{x}) = \sum_{x \in \mathbb{A}} \Psi(x | \hat{x}) \cdot \Phi(x).$$

We shall use that information divergence satisfies the so-called compensation equality first recognized in [18]. Let (s_1, s_2, \dots, s_k) denote a probability vector and let Q and P_1, P_2, \dots, P_k denote probability measures on the same set. Then

$$\sum_{j=1}^k s_j D(P_j \| Q) = \sum_{j=1}^k s_j D(P_j \| \bar{P}) + D(\bar{P} \| Q)$$

where $\bar{P} = \sum_j s_j P_j$. In particular

$$\sum_{j=1}^k s_j D(P_j \| Q) \geq \sum_{j=1}^k s_j D(P_j \| \bar{P}). \quad (6)$$

This leads to

$$\begin{aligned} E \left[d(X, i(\hat{x})) \mid \hat{X} = \hat{x} \right] &= E \left[D(\Phi(X) \| i(\hat{x})) \mid \hat{X} = \hat{x} \right] \\ &\geq E \left[D(\Phi(X) \| i_\Psi(\hat{x})) \mid \hat{X} = \hat{x} \right]. \end{aligned}$$

Therefore

$$\begin{aligned} E \left[d(X, i(\hat{X})) \right] &\geq E \left[D(\Phi(X) \| i_\Psi(\hat{X})) \right] \\ &= E \left[d(X, i_\Psi(\hat{X})) \right]. \end{aligned}$$

We also have that $I(X, i(\hat{X})) \geq I(X, i_\Psi(\hat{X}))$ so instead of minimizing over all possible maps i and all possible joint distributions on $\mathbb{A} \times \widehat{\mathbb{A}}$ we just have to minimize over joint distributions and put $i = i_\Psi$.

$$\hat{X} \longrightarrow X \longrightarrow Y$$

For a joint distribution of (X, \hat{X}) we have that $i_\Psi(\hat{x})$ is the distribution of Y given $\hat{X} = \hat{x}$ if \hat{X} and Y are independent given X , i.e. $\hat{X} \rightarrow X \rightarrow Y$ form a Markov chain. Now

$$\begin{aligned} I(X; Y) &= I(\hat{X}; Y) + I(X; Y \mid \hat{X}) \\ &= I(\hat{X}; Y) + E \left[D(\Phi(X) \| i_\Psi(\hat{X})) \right] \\ &= I(\hat{X}; Y) + E \left[d(X; i_\Psi(\hat{X})) \right]. \end{aligned}$$

Thus

$$E \left[d(X, i_\Psi(\hat{X})) \right] = I(X; Y) - I(\hat{X}; Y).$$

We note that $I(X, Y)$ is a constant and get the following theorem.

Theorem 4: When information divergence is used as distortion measure

$$\begin{aligned} \inf \left[I(X, \hat{X}) + \beta \cdot E \left(d(X, \hat{X}) \right) \right] \\ = \beta \cdot I(X; Y) + \inf \left[I(X, \hat{X}) - \beta \cdot I(\hat{X}; Y) \right]. \end{aligned}$$

The solution satisfies the so-called bottleneck equations, i.e. the minimum is attained for reconstruction points satisfying $i = i_\Psi$ and joint distribution satisfies the Kuhn-Tucker conditions for rate distortion with these reconstructions points.

The last term is essentially the one that shall be minimized in the information bottleneck where one wants to minimize $I(X, \hat{X})$ and at the same time maximize $I(\hat{X}; Y)$. We have seen that inequality (6) is essentially in deriving Theorem 4.

Theorem 5: If $d : M_+^1(A) \times M_+^1(A) \rightarrow \mathbb{R}$ is a Csiszár f -divergence for some differentiable function f and

$$\sum_{j=1}^k s_j d(P_j, Q) \geq \sum_{j=1}^k s_j d(P_j, \bar{P})$$

for all mixtures $\bar{P} = \sum_{j=1}^k s_j P_j$ and all Q then d is proportional to information divergence.

Proof: Assume that d is the Csiszár f -divergence given by

$$d(P, Q) = D_f(P \| Q) = \sum_{i \in \mathbb{B}} f \left(\frac{P(i)}{Q(i)} \right) Q(i).$$

We shall write f as $f(x) = xg(x^{-1})$ so that

$$D_f(P \| Q) = \sum_{i \in \mathbb{B}} g \left(\frac{Q(i)}{P(i)} \right) P(i).$$

We have

$$\sum_{j=1}^k s_j \sum_{i \in \mathbb{B}} g \left(\frac{Q(i)}{P_j(i)} \right) P_j(i) \geq \sum_{j=1}^k s_j \sum_{i \in \mathbb{B}} g \left(\frac{\bar{P}(i)}{P_j(i)} \right) P_j(i).$$

The probability vector $(Q(1), Q(2), \dots)$ satisfies $\sum Q(i) = 1$. We introduce a Lagrange multiplier and calculate

$$\begin{aligned} \frac{\partial}{\partial Q(i)} \left(\sum_{j=1}^k s_j \sum_{i \in \mathbb{B}} g \left(\frac{Q(i)}{P_j(i)} \right) P_j(i) - \lambda \cdot \sum Q(i) \right) \\ = \sum_{j=1}^k s_j g' \left(\frac{Q(i)}{P_j(i)} \right) - \lambda. \end{aligned}$$

Define $h(x) = g'(x^{-1})$. Then

$$\sum_{j=1}^k s_j h \left(\frac{P_j(i)}{\bar{P}(i)} \right) = \lambda$$

holds for all i . because \bar{P} is a stationary point. In particular it should hold if there exists an i' such that $P_j(i')$ does not depend on j . In that case

$$\lambda = \sum_{j=1}^k s_j h \left(\frac{P_j(i')}{\bar{P}(i')} \right) = \sum_{j=1}^k s_j h(1) = h(1).$$

Thus for any $i \in \mathbb{B}$ we have

$$\sum_{j=1}^k s_j h \left(\frac{P_j(i)}{\bar{P}(i)} \right) = h(1).$$

That implies that $h(x) = h(1) + \alpha(x - 1)$ for some constant α . Then

$$\begin{aligned} g(x) &= \int g'(x) dx = \int (h(1) + \alpha(x^{-1} - 1)) dx \\ &= (h(1) - \alpha)x + \alpha \log x + c \end{aligned}$$

for some constant $c \in \mathbb{R}$. Then

$$\begin{aligned} f(x) &= x((h(1) - \alpha)x^{-1} + \alpha \log(x^{-1}) + c) \\ &= h(1) - \alpha + \alpha x \log x + cx. \end{aligned}$$

The condition $f(1) = 0$ implies $h(1) - \alpha = -c$

$$f(x) = \alpha x \log x + c(x - 1).$$

Hence

$$\begin{aligned} D_f(P, Q) &= \sum_{i \in \mathbb{B}} \left(\alpha \frac{P(i)}{Q(i)} \log \frac{P(i)}{Q(i)} + c \left(\frac{P(i)}{Q(i)} - 1 \right) \right) Q(i) \\ &= \alpha \sum_{i \in \mathbb{B}} P(i) \log \frac{P(i)}{Q(i)} = \alpha D(P \| Q). \end{aligned}$$

In [13] it was shown that a divergence measure satisfying (6) must be a Bregman divergence. This leads us to the following corollary.

Corollary 6: A divergence measure that is both a f -divergence and a Bregman divergence must be proportional to information divergence.

In [17, Thm. 4] the intersection of the set of Bregman and f -divergences is characterized by an equation and a concavity condition and information divergence is given as an example of an element in the intersection. Corollary 6 implies that there are essentially no other elements in the intersection.

VI. CONCLUSION

In this paper various distortion measures have been considered and their properties with respect to rate distortion theory have been studied. Some of the results are summarized in the following table.

Property	Class of divergences
Additivity	Inf. div. and reversed
Sufficiency on input side	All
Sufficiency on output side	non-convex "f-div."
Data processing inequality	Csiszár f -div.
Finiteness	All
Bottleneck equation	Bregman div.

We see that if one wants to have all the properties fulfilled the divergence is equal to or proportional to information divergence. A divergence that is proportional to information divergence is essentially information divergence measured by different units (for instance bits instead of nats). Thus it is desirable to use information divergence, that leads to the information bottleneck method.

Acknowledgement 7: The authors want to thank Peter Grünwald and Tim van Erven for useful discussions and comments.

REFERENCES

- [1] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.
- [2] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *SIGIR '00: Proceedings of the ACM SIGIR conference*, (New York, NY, USA), pp. 208–215, ACM Press, 2000.
- [3] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learn. Res.*, vol. 6, pp. 165–188, 2005.
- [4] R. Gilad-Bachrach, A. Navot, and N. Tishby, "A Study of the Information Bottleneck for Method and its Relationship to Classical Problems," HUJI Leibniz tech. rep., 2006.
- [5] C. R. Shalizi and J. P. Crutchfield, "Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction," tech. rep., Santa Fe Institute, 2002.
- [6] N. Slonim and Y. Weiss, "Maximum likelihood and the information bottleneck," in *Advances in Neural Information Processing Systems 15, Vancouver, Canada, 2002* (S. Becker, S. Thrun, and K. Obermayer, eds.), pp. 351–358, 2003.
- [7] G. Elidan and N. Friedman, "The information bottleneck EM algorithm," in *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, (San Francisco, CA), pp. 200–208, Morgan Kaufmann, 2003.
- [8] D. Gondek and T. Hofmann, "Conditional information bottleneck clustering," in *Workshop on Clustering Large Data Sets, IEEE International Conference on Data Mining, Melbourne, USA, November 19-22, 2003.*, 2003.
- [9] T. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.
- [10] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory, Now Publishers Inc., 2004.
- [11] F. Liese and I. Vajda, "On divergence and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, pp. 4394 – 4412, Oct. 2006.
- [12] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967. Translated from Russian.
- [13] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, p. 17051749, 2005.
- [14] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inform. Theory*, vol. 28, pp. 489–495, 1982.
- [15] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory," *Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1367–1433, 2004.
- [16] M. C. Pardo and I. Vajda, "On asymptotic properties of information-theoretic divergences," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1860–1867, 2003.
- [17] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1288–1293, 1997.
- [18] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 291–292, 1967.