

# Estimation-Theoretic Representation of Mutual Information

Daniel P. Palomar and Sergio Verdú

Department of Electrical Engineering  
Princeton University  
Engineering Quadrangle, Princeton, NJ 08544, USA  
{danielp,verdu}@princeton.edu

## Abstract

A fundamental relationship between information theory and estimation theory was recently unveiled for the Gaussian channel, relating the derivative of mutual information with the minimum mean-square error.

This paper generalizes this fundamental link between information theory and estimation theory to arbitrary channels and in particular encompasses the discrete memoryless channel (DMC). In addition to the intrinsic theoretical interest of such a result, it naturally leads to an efficient numerical computation of mutual information for cases in which it was previously infeasible such as with LDPC codes.

## 1 Introduction and Motivation

A fundamental relationship between estimation theory and information theory was recently unveiled in [1] for Gaussian channels; in particular, it was shown that, for the scalar Gaussian channel

$$Y = \sqrt{\text{snr}} X + N \quad (1)$$

and regardless of the input distribution, the mutual information and the minimum mean-square error (MMSE) are related (assuming complex-valued inputs/outputs) by

$$\frac{d}{d\text{snr}} I(X; \sqrt{\text{snr}} X + N) = \mathbb{E} \left[ \left| X - \mathbb{E} [X | \sqrt{\text{snr}} X + N] \right|^2 \right] \quad (2)$$

where the right-hand side is the MMSE corresponding to the best estimation of  $X$  upon the observation  $Y$  for a given signal-to-noise ratio (SNR)  $\text{snr}$ . The same relation was shown to hold for the linear vector Gaussian channel

$$\mathbf{Y} = \sqrt{\text{snr}} \mathbf{H} \mathbf{X} + \mathbf{N}. \quad (3)$$

---

This work was supported in part by the Fulbright Program and the Ministry of Education and Science of Spain; the U.S. National Science Foundation under Grant NCR-0074277; and through collaborative participation in the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

Similar results hold in a continuous-time setting, i.e., the derivative of the mutual information is equal to the noncausal MMSE. Other generalizations were also obtained in [1] such as when the input undergoes an arbitrary random transformation before contamination by additive Gaussian noise.

The previous results on the derivative of the mutual information with respect to the SNR for Gaussian channels were later generalized in [2] to embrace derivatives with respect to arbitrary parameters; in particular, the relation was compactly expressed for the linear vector Gaussian channel in terms of the gradient of the mutual information with respect to the channel matrix  $\mathbf{H}$  as

$$\nabla_{\mathbf{H}} I(\mathbf{X}; \mathbf{H}\mathbf{X} + \mathbf{N}) = \mathbf{H}\mathbf{E} \quad (4)$$

where

$$\mathbf{E} \triangleq \mathbb{E} \left[ (\mathbf{X} - \mathbb{E}[\mathbf{X} | \mathbf{Y}]) (\mathbf{X} - \mathbb{E}[\mathbf{X} | \mathbf{Y}])^\dagger \right] \quad (5)$$

is the covariance matrix of the estimation error vector, also known as the MMSE matrix. The derivative with respect to an arbitrary parameter can be readily obtained from this gradient via a chain rule for differentiation. In addition to their intrinsic theoretical interest, these fundamental relations between mutual information and MMSE have already found several applications.

Counterparts of the fundamental relation that express the derivative of mutual information have been explored for other types of channels; namely, for Poisson channels [3], for additive non-Gaussian channels [4], and for the discrete memoryless channel (DMC) [5]. As should be expected the MMSE does not play a role in the representation of mutual information for these channels.

The goal of this paper is to generalize the link between information theory and estimation theory to arbitrary channels. Thus, completing the connection between information theory and estimation theory found in [1]. Generalizing the abovementioned approaches, our main result gives the derivative of mutual information with respect to a channel parameter  $\theta$  in terms of the input estimate given by the posterior distribution  $P_{X|Y}^\theta$  as<sup>1</sup>

$$\frac{\partial}{\partial \theta} I(X; Y) = \mathbb{E} \left[ \frac{\partial \log_e P_{Y|X}^\theta(Y | X)}{\partial \theta} \log P_{X|Y}^\theta(X | Y) \right]. \quad (6)$$

For the particular case of a memoryless channel, the derivative is expressed in terms of the individual input estimates given by the posterior marginals  $P_{X_i|Y^n}^\theta$  as

$$\frac{\partial}{\partial \theta} I(X^n; Y^n) = \sum_{i=1}^n \mathbb{E} \left[ \frac{\partial \log_e P_{Y_i|X_i}^\theta(Y_i | X_i)}{\partial \theta} \log P_{X_i|Y^n}^\theta(X_i | Y^n) \right]. \quad (7)$$

Observe that in this more general setup that embraces any arbitrary channel, the role of the conditional estimator  $\mathbb{E}[X_i|Y^n]$  (which arises in the Gaussian channel) has been generalized to the corresponding conditional distribution  $P_{X_i|Y^n}^\theta$ .

In addition to the theoretical interest of this characterization, one practical application is the computation of the mutual information  $I(X^n; Y^n)$  achieved by a given code (as

---

<sup>1</sup>The base of the second logarithm in (6) agrees with the units of the mutual information.

opposed to an ensemble of codes) when input to a memoryless channel. In such a case,  $P_{X^n}$  is a distribution that puts equal mass on the codewords and zero mass elsewhere. Indeed, the mutual information achieved by a given code over a channel is a useful information that finds several applications, for example, in studying the concatenation of two or more coding schemes, in bounding the size of a code to achieve a desired block error rate or bit error rate, and in the analysis and design of systems via EXIT charts.

In [6, 7, 8], the computation of the information rate for finite-state Markov sources over channels with finite memory was efficiently obtained with a Monte Carlo algorithm, based on the fact that  $P_{Y^n}$  can be computed very efficiently in practice with the forward recursion of the BCJR algorithm. However, in a more general setting where the source is not a Markov process, the existing approaches cannot be used. Indeed, for an arbitrary source, a direct computation of the mutual information is a notoriously difficult task and infeasible in most realistic cases since it requires an enumeration of the whole codebook for the computation of the output probability  $P_{Y^n}$  or of the posterior probability of the input conditioned on the output  $P_{X^n|Y^n}$  [9].

Based on (7), it is possible to obtain a numerical method to compute the mutual information via its derivative which requires the posterior marginals  $P_{X_i|Y^n}$  (instead of the joint posterior  $P_{X^n|Y^n}$  or joint output  $P_{Y^n}$ ) or, equivalently, the symbolwise a posteriori probabilities (APP) obtained by an optimum soft decoder. As is well known, in some notable cases of interest, the APPs can be computed or approximated very efficiently in practice by message-passing algorithms. For example, for Markov sources (e.g., convolutional codes or trellis codes) the forward-backward algorithm (also known as BCJR algorithm) computes the exact posterior marginals. In other cases, the posterior marginals can only be approximated such as in the turbo decoding for concatenated codes, the soft decoding of Reed-Solomon codes, the *sum-product* algorithm for factor graphs (e.g., sparse codes such as low-density parity-check (LDPC) codes) [10, 11], and Pearl's *belief propagation* algorithm for Bayesian networks [9].

In addition to its computational applications, the representation of the derivative the mutual information has some interesting analytical applications. Indeed, it has been recently shown in [5] that the derivative of the conditional entropy of the input given the output (or, equivalently, the mutual information) with respect to a channel parameter can be used as a generalization of the EXIT chart (called GEXIT chart) which has very appealing properties as a tool for analyzing the behavior of ensembles of codes using iterative decoding.

## 2 Derivative of Mutual Information

We first give a general representation for arbitrary random transformations and memoryless channels. Then, we particularize the results for specific types of channels such as the BSC, BEC, DMC, and Gaussian channel.

### 2.1 General Representation

The following result characterizes the derivative of the mutual information for an arbitrary

bitrary random transformation with arbitrary input and output alphabets.<sup>2</sup>

**Theorem 1** Consider a random transformation  $P_{Y|X}^\theta$ , which is differentiable as a function of the parameter  $\theta$ , and a random input with distribution  $P_X$  (independent of  $\theta$ ). Then, the derivative of the mutual information  $I(X; Y)$  with respect to the parameter  $\theta$  can be written in terms of the posterior distribution  $P_{X|Y}^\theta$  as<sup>3</sup>

$$\frac{\partial}{\partial \theta} I(X; Y) = \mathbb{E} \left[ \frac{\partial \log_e P_{Y|X}^\theta(Y | X)}{\partial \theta} \log P_{X|Y}^\theta(X | Y) \right], \quad (8)$$

where the expectation is with respect to the joint distribution  $P_X P_{Y|X}^\theta$ .

Observe that when  $P_{Y|X}^\theta$  and  $P_{X|Y}^\theta$  are not pdf's or pmf's, Theorem 1 similarly holds using instead Radon-Nikodym derivatives.

The result in Theorem 1 particularizes to the formulas found for the cases previously solved such as the Gaussian channel [1], the additive-noise (not necessarily Gaussian) channel [4], and the Poisson channel [3].

Theorem 1 can be readily particularized to the case of an arbitrary channel with transition probability  $P_{Y^n|X^n}^\theta$ , where  $n$  denotes the number of uses of the channel and the input and output alphabets are  $n$ -dimensional Cartesian products, and input distribution  $P_{X^n}$ . For the case of a memoryless channel (with possibly dependent inputs), Theorem 1 simplifies as follows.

**Theorem 2** Consider a memoryless channel with transition probability  $P_{Y^n|X^n}^\theta = \prod_{i=1}^n P_{Y_i|X_i}^{\theta_i}$ , where  $P_{Y_i|X_i}^{\theta_i}$  is differentiable as a function of the parameter  $\theta_i$  (and independent of  $\theta_j$  for  $j \neq i$ ), and a random input with distribution  $P_{X^n}$  (independent of  $\theta_i$  for all  $i$ ). Then, the derivative of the mutual information  $I(X^n; Y^n)$  with respect to the parameter  $\theta_i$  can be written in terms of the posterior marginal distribution  $P_{X_i|Y^n}^\theta$  as

$$\frac{\partial}{\partial \theta_i} I(X^n; Y^n) = \mathbb{E} \left[ \frac{\partial \log_e P_{Y_i|X_i}^{\theta_i}(Y_i | X_i)}{\partial \theta_i} \log P_{X_i|Y^n}^\theta(X_i | Y^n) \right] \quad (9)$$

where the expectation is with respect to the joint distribution  $P_{X_i} P_{Y^n|X_i}^\theta$ .

Observe that if the channel is time-invariant (i.e., if  $P_{Y_i|X_i}^{\theta_i} = P_{Y|X}^\theta$  for all  $i$ ), then, by simply applying the chain rule for differentiation with  $\theta_i = \theta$  for all  $i$ , we get (7).

The result in Theorem 2 for a memoryless channel can be easily extended to a finite-state Markov channel.

An interesting application of Theorem 2 is the computation of the derivative of the mutual information of a given and fixed  $(2^{nR}, n)$  code used over a memoryless channel, where  $n$  and  $R$  are the blocklength and the rate of the code, respectively. This is easily done by defining the input distribution  $P_{X^n}$  as the one induced by the code (typically

---

<sup>2</sup>The results in this paper require some mild ‘‘regularity conditions’’ about the interchange of the order of differentiation and integration (expectation) which are satisfied in most cases of interest and implicitly assumed hereinafter.

<sup>3</sup>Unless the logarithm basis is indicated, it can be chosen arbitrarily as long as both sides of the equation have the same units.

under an equiprobable choice of codewords). Indeed, the practical relevance of Theorem 2 for numerical computation is remarkable since, as already mentioned, the symbolwise APP  $P_{X_i|Y^n}$  obtained by an optimum soft decoder can be efficiently computed in practice with a message-passing algorithm such as the BCJR, sum-product, or belief-propagation algorithms [9, 10, 11]. The expectation over  $X_i$  and  $Y^n$  can be numerically approximated with a Monte Carlo approach by averaging over many realizations of  $X_i$  and  $Y^n$ . Alternatively, one can consider the numerical approximation of the expectation only over  $Y^n$  and then obtain the an inner expectation over  $X_i$  conditioned on  $Y^n$  through  $P_{X_i|Y^n}$ ; then, for a finite input alphabet, (7) becomes

$$\frac{\partial}{\partial \theta} I(X^n; Y^n) = \sum_{i=1}^n \mathbb{E} \left[ \sum_{x_i} P_{X_i|Y^n}^\theta(x_i | Y^n) \frac{\partial \log_e P_{Y|X}^\theta(Y_i | x_i)}{\partial \theta} \log P_{X_i|Y^n}^\theta(x_i | Y^n) \right]. \quad (10)$$

## 2.2 Binary Symmetric Channel (BSC)

Theorem 2 can be readily particularized for the BSC with

$$\frac{d \log_e P_{Y|X}^\delta(y_i | x_i)}{d\delta} = \frac{x_i \oplus y_i}{\delta} - \frac{1 - x_i \oplus y_i}{1 - \delta} \quad (11)$$

where  $\oplus$  denotes the xor operation or sum in modulo 2.

The following result carries out the expectation over  $X_i$  and  $Y_i$  analytically and provides an alternative formula in terms of extrinsic information (summarized by the distribution  $P_{X_i|Y^{n \setminus i}}^\delta$ , where  $y^{n \setminus i}$  denotes the sequence  $y^n$  except the  $i$ th element  $y_i$ ) which is frequently more useful.

**Theorem 3** *Consider a BSC with crossover probability  $\delta \in (0, 1)$  and input distribution  $P_{X^n}$ . Then,*

$$\begin{aligned} \frac{d}{d\delta} I(X^n; Y^n) &= \sum_{i=1}^n \left( \mathbb{E} \left[ \tanh \left( \frac{\lambda_i(Y^{n \setminus i})}{2 \log e} \right) \log \left( \frac{\exp(\lambda_i(Y^{n \setminus i})) + \exp(-\gamma)}{\exp(\lambda_i(Y^{n \setminus i})) + \exp(\gamma)} \right) \right] - 2\gamma P_{X_i}(1) \right) \end{aligned} \quad (12)$$

where

$$\lambda_i(y^{n \setminus i}) \triangleq \log \frac{P_{X_i|Y^{n \setminus i}}^\delta(0 | y^{n \setminus i})}{P_{X_i|Y^{n \setminus i}}^\delta(1 | y^{n \setminus i})} \quad (13)$$

and

$$\gamma = \log \frac{1 - \delta}{\delta}. \quad (14)$$

## 2.3 Binary Erasure Channel (BEC)

The following result refines Theorem 2 for the BEC and provides an alternative formula in terms of extrinsic information.

**Theorem 4** Consider a BEC with erasure probability  $\epsilon \in (0, 1)$  and input distribution  $P_{X^n}$ . Then,

$$\frac{d}{d\epsilon} I(X^n; Y^n) = - \sum_{i=1}^n \mathbb{E} \left[ \frac{\log(1 + \exp(\lambda_i(Y^{n \setminus i})))}{1 + \exp(\lambda_i(Y^{n \setminus i}))} + \frac{\log(1 + \exp(-\lambda_i(Y^{n \setminus i})))}{1 + \exp(-\lambda_i(Y^{n \setminus i}))} \right]$$

where

$$\lambda_i(y^{n \setminus i}) \triangleq \log \frac{P_{X_i|Y^{n \setminus i}}^\epsilon(0 | y^{n \setminus i})}{P_{X_i|Y^{n \setminus i}}^\epsilon(1 | y^{n \setminus i})}. \quad (15)$$

## 2.4 Discrete Memoryless Channel (DMC)

Consider a DMC with arbitrary finite input alphabet  $\mathcal{X} = \{a_1 \cdots a_{|\mathcal{X}|}\}$ , arbitrary finite output alphabet  $\mathcal{Y} = \{b_1 \cdots b_{|\mathcal{Y}|}\}$ , and arbitrary time-invariant memoryless channel transition probability  $P_{Y^n|X^n}(y^n | x^n) = \prod_{i=1}^n P_{Y|X}(y_i | x_i)$ . The channel transition probability can be compactly described with the channel transition matrix  $\mathbf{\Pi}$  with  $(k, l)$ th element defined as  $[\mathbf{\Pi}]_{kl} = \pi_{kl} = P_{Y|X}(b_k | a_l)$ .

Theorem 2 can be readily particularized for the DMC in terms of extrinsic information as

$$\frac{\partial}{\partial \theta} I(X^n; Y^n) = \sum_{i=1}^n \sum_{x_i, y_i, y^{n \setminus i}} P_{X_i}(x_i) P_{Y^{n \setminus i}|X_i}^\theta(y^{n \setminus i} | x_i) \frac{\partial P_{Y|X}^\theta(y_i | x_i)}{\partial \theta} \log \left( \frac{P_{X_i|Y^{n \setminus i}}^\theta(x_i | y^{n \setminus i}) P_{Y|X}^\theta(y_i | x_i)}{\sum_{\tilde{x}_i} P_{X_i|Y^{n \setminus i}}^\theta(\tilde{x}_i | y^{n \setminus i}) P_{Y|X}^\theta(y_i | \tilde{x}_i)} \right). \quad (16)$$

An equivalent form of (16) was independently obtained in [5, Thm. 1] where the conditioning is with respect to an extrinsic information random variable  $Z_i$  instead of  $Y^{n \setminus i}$ . The convergence analysis of the decoding of LDPC code ensembles is carried out in [5] by the GEXIT of the code ensemble (a generalization of the EXIT which is defined as the negative of the derivative of mutual information averaged over the code ensemble).

The following result refines Theorem 2 for the DMC by carrying out the expectation over  $X_i$  and  $Y_i$  analytically and provides an alternative formula in terms of extrinsic information.

**Theorem 5** Consider a DMC with channel transition matrix  $\mathbf{\Pi}$  and input distribution  $P_{X^n}$ . Then, provided that  $\pi_{kl} > 0$ ,<sup>4</sup>

$$[\nabla_{\mathbf{\Pi}} I(X^n; Y^n)]_{kl} = - \sum_{i=1}^n \mathbb{E} \left[ \frac{\log \left( 1 + \sum_{m \neq l} \frac{\pi_{km}}{\pi_{kl}} \exp \left( \lambda_i^{(m,l)}(Y^{n \setminus i}) \right) \right)}{1 + \sum_{m \neq l} \exp \left( \lambda_i^{(m,l)}(Y^{n \setminus i}) \right)} \right] \quad (17)$$

where

$$\lambda_i^{(m,l)}(y^{n \setminus i}) \triangleq \log \frac{P_{X_i|Y^{n \setminus i}}(a_m | y^{n \setminus i})}{P_{X_i|Y^{n \setminus i}}(a_l | y^{n \setminus i})}. \quad (18)$$

<sup>4</sup>The gradient with respect to a matrix  $\nabla_{\mathbf{X}} f$  is defined as  $[\nabla_{\mathbf{X}} f]_{ij} \triangleq \partial f / \partial [\mathbf{X}]_{ij}$ .

The usefulness of the gradient in Theorem 5 is as an intermediate step in the computation of the derivative with respect to an arbitrary parameter  $\theta$  via the chain rule for differentiation:

$$\frac{\partial}{\partial \theta} I(X^n; Y^n) = \text{Tr} \left( \nabla_{\mathbf{\Pi}}^T I(X^n; Y^n) \frac{\partial \mathbf{\Pi}}{\partial \theta} \right) \quad (19)$$

where only the elements of the gradient  $\nabla_{\mathbf{\Pi}} I(X^n; Y^n)$  that are multiplied by nonzero elements of  $\partial \mathbf{\Pi} / \partial \theta$  need to be computed.

## 2.5 Gaussian Channel

For the Gaussian channel, Theorem 2 can be particularized to obtain (2) and (4) (for the case of iid inputs) in agreement with [1] and [2], respectively.

The following result further refines Theorem 2 for the binary input AWGN (BIAWGN) channel by carrying out analytically the expectation over  $X_i$  and  $Y_i$ .

**Theorem 6** *Consider a real-valued Gaussian channel with the channel transition probability  $P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} e^{-(y-\sqrt{\text{snr}}x)^2/2}$  and a binary  $\{\pm 1\}$  input distribution  $P_{X^n}$ . Then,*

$$\frac{d}{d\text{snr}} I(X^n; Y^n) = \frac{-1}{2\sqrt{\text{snr}}} \sum_{i=1}^n \mathbb{E} \left[ \frac{\Psi(\lambda_i(Y^{n \setminus i}); \text{snr})}{1 + \exp(\lambda_i(Y^{n \setminus i}))} + \frac{\Psi(-\lambda_i(Y^{n \setminus i}); \text{snr})}{1 + \exp(-\lambda_i(Y^{n \setminus i}))} \right] \quad (20)$$

where

$$\lambda_i(y^{n \setminus i}) \triangleq \log \frac{P_{X_i|Y^{n \setminus i}}(-1|y^{n \setminus i})}{P_{X_i|Y^{n \setminus i}}(+1|y^{n \setminus i})} \quad (21)$$

and

$$\begin{aligned} \Psi(x; \text{snr}) &\triangleq \mathbb{E} [N \log(1 + \exp(x - 2\sqrt{\text{snr}}(\sqrt{\text{snr}} + N) \log e))] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \nu \log(1 + \exp(x - 2\sqrt{\text{snr}}(\sqrt{\text{snr}} + \nu) \log e)) e^{-\nu^2/2} d\nu. \end{aligned} \quad (22)$$

## 3 Applications

### 3.1 Computation of Mutual Information of LDPC Codes

Consider the computation of the mutual information of an LDPC code over a channel via the derivative. Figure 1(a) shows the mutual information of different codes of rate 1/2 over a BSC versus the channel crossover probability  $\delta$ . The curves were computed via Theorem 3 and the posterior marginals for the LDPC codes were estimated using the sum-product algorithm<sup>5</sup> [10, 11]. In particular, two LDPC codes with blocklengths  $n = 96$  and  $n = 4000$  are considered as well as a simple repetition code. Naturally, for

<sup>5</sup>In general, the belief propagation algorithm will only give an estimate of the posterior marginals; hence, the computation of the mutual information based on these estimations will inevitably be subject to the accuracy of these estimates. A number of algorithms that provide more accurate estimates than the basic belief-propagation algorithm have been recently proposed.

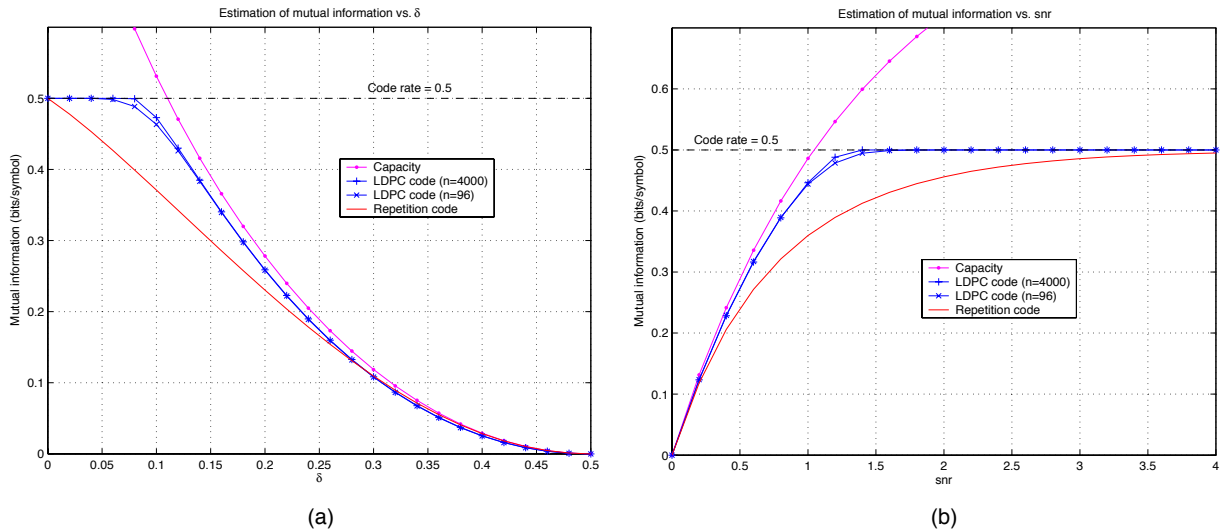


Figure 1: Estimation of the mutual information of different codes of rate 1/2 over: (a) a BSC and (b) an antipodal Gaussian channel.

$\delta = 0$  all codes achieve a mutual information equal to the code rate. The difference among the LDPC codes is not significant (this observation is in terms of mutual information and does not imply that the decoding of these different LDPC codes is expected to give the same error performance whatsoever); the implication is that using a long LDPC code is essentially equivalent to using a short one combined with an outer code.

Figure 1(b) shows the mutual information of the same codes but over a Gaussian channel instead of a BSC. The curves were computed via Theorem 6 and the posterior marginals for the LDPC codes were estimated using the sum-product algorithm. As happened in the BSC, the LDPC codes achieve a mutual information essentially equal to the code rate whenever the rate is below a certain fraction of the channel capacity; the repetition code, however, shows a much worse behavior.

### 3.2 Universal Estimation of the Derivative of Mutual Information

Another application of our results is the estimation of the derivative of the mutual information achieved by inputs which are neither accessible nor statistically known (hence the term universal), as is frequently the case when dealing with text, images, etc. Assuming that the channel is discrete memoryless and known (with full rank transition probability matrix), it is possible to estimate the derivative of the mutual information by simply observing the output. To that end, we use one of the universal algorithms recently developed to estimate the posterior marginals  $P_{X_i|Y^{n \setminus i}}$  [12, 13] and then we apply Theorem 5 for the DMC.

To compute the mutual information by integrating the derivative, we must have access to the outputs corresponding to a grid of channels with a range of qualities starting from a perfect channel. The input (assumed to be stationary ergodic) is neither accessible nor statistically known and is only observed after passing through the channel. Theorem 5 (combined with (19)) can be more conveniently rewritten (due to the stationarity and

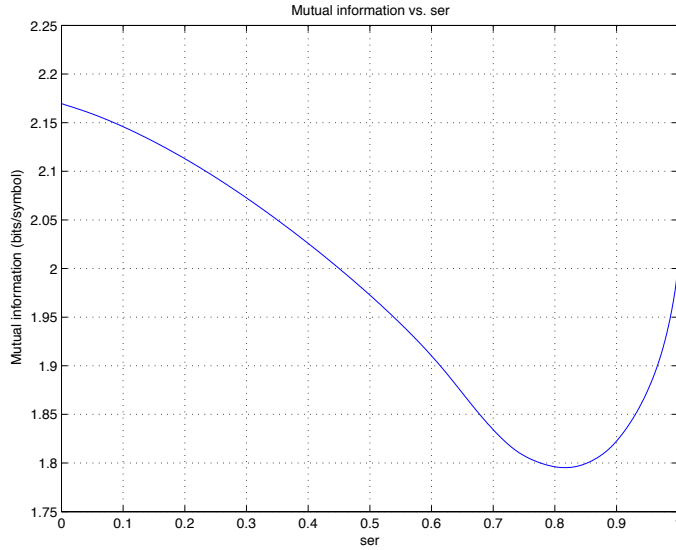


Figure 2: Input-output mutual information of *Don Quixote de La Mancha* over the typewriter channel as a function of the symbol error probability.

ergodicity) as

$$\frac{\partial}{\partial \theta} \frac{1}{n} I(X^n; Y^n) \approx \frac{1}{n} \sum_{i=1}^n \text{Tr} \left( \mathbf{R}_i^T(y^{n \setminus i}) \frac{\partial \Pi}{\partial \theta} \right) \quad (23)$$

where

$$[\mathbf{R}_i(y^{n \setminus i})]_{kl} = - \frac{\log \left( 1 + \sum_{m \neq l} \frac{\pi_{km}}{\pi_{kl}} \exp \left( \lambda_i^{(m,l)}(y^{n \setminus i}) \right) \right)}{1 + \sum_{m \neq l} \exp \left( \lambda_i^{(m,l)}(y^{n \setminus i}) \right)} \quad (24)$$

and the sequence  $y^n$  is obtained by passing an (unknown) sequence  $x^n$  through the channel.

As an illustration of the previous approach, we compute the amount of information about the source received by a reader of the novel *Don Quixote de La Mancha* (in English translation) given that the printed novel contains errors introduced by the typist. We model this channel by assuming that each letter is independently flipped, with some symbol error rate (SER) of `ser`, equiprobably into one of its nearest neighbors in the QWERTY keyboard. Figure 2 shows the mutual information obtained by integrating the derivative from the point of reference `ser` = 0. For `ser` = 0, the mutual information equals the entropy of *Don Quixote de La Mancha* which is 2.17 bits/symbol. Interestingly, the mutual information (as well as the capacity for this channel) is not monotonic; in particular, the mutual information decreases for symbol error rates up to `ser`  $\approx$  0.82 and then increases. The reason for this behavior is that, for a sufficiently high `ser`, each letter and its neighbors happen with roughly the same probability, whereas as `ser`  $\rightarrow$  1, the probability that the intended letter is indeed observed at the output of the channel becomes zero and this reduces the uncertainty about the transmitter letter given the observed one.

## References

- [1] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.
- [2] D. P. Palomar and S. Verdú, “Gradient of mutual information in linear vector Gaussian channels,” in *Proc. 2005 IEEE International Symposium on Information Theory (ISIT 2005)*, Adelaide, Australia, Sept. 4-9, 2005.
- [3] D. Guo, S. Shamai, and S. Verdú, “Mutual information and conditional mean estimation in Poisson channels,” in *Proc. 2004 IEEE Information Theory Workshop*, pp. 265–270, San Antonio, TX, USA, Oct. 24-29, 2004.
- [4] —, “Additive non-Gaussian noise channels: Mutual information and conditional mean estimation,” in *Proc. 2005 IEEE International Symposium on Information Theory (ISIT 2005)*, Adelaide, Australia, Sept. 4-9, 2005.
- [5] C. Méasson, R. Urbanke, A. Montanari, and T. Richardson, “Life above threshold: From list decoding to area theorem and MSE,” in *Proc. IEEE Inform. Theory Workshop*, San Antonio, TX, USA, 2004.
- [6] D. Arnold and H.-A. Loeliger, “On the information rate of binary-input channels with memory,” in *Proc. 2001 IEEE International Conference on Communications (ICC 2001)*, pp. 2692–2695, Helsinki, Finland, June 11-14, 2001.
- [7] V. Sharma and S. K. Singh, “Entropy and channel capacity in the regenerative setup with applications to Markov channels,” in *Proc. 2001 IEEE International Symposium on Information Theory (ISIT 2001)*, p. 283, Washington, DC, USA, June 24-29, 2001.
- [8] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, “On the achievable information rates of finite-state ISI channels,” in *Proc. IEEE 2001 Global Communications Conference (Globecom-2001)*, San Antonio, TX, USA, Nov. 25-29, 2001.
- [9] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, 2nd ed. San Francisco, CA: Kaufmann, 1988.
- [10] N. Wiberg, “Codes and decoding on general graphs,” Ph.D. dissertation, Linköping University, Linköping, Sweden, 1996.
- [11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
- [12] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, “Universal discrete denoising: Known channel,” *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.
- [13] J. Yu and S. Verdú, “Schemes for bi-directional modeling of discrete stationary sources,” in *Proc. 39th IEEE Conference on Information Sciences and Systems (CISS-2005)*, The John Hopkins University, Baltimore, MD, March 16-18, 2005.