Representation of Mutual Information Via Input Estimates

Daniel P. Palomar, *Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—A relationship between information theory and estimation theory was recently shown for the Gaussian channel, relating the derivative of mutual information with the minimum mean-square error. This paper generalizes the link between information theory and estimation theory to arbitrary channels, giving representations of the derivative of mutual information as a function of the conditional marginal input distributions given the outputs. We illustrate the use of this representation in the efficient numerical computation of the mutual information achieved by inputs such as specific codes or natural language.**

***Index Terms*—Computation of mutual information, extrinsic information, input estimation, low-density parity-check (LDPC) codes, minimum mean square error (MMSE), mutual information, soft channel decoding.**

I. Introduction and Motivation

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

$$Y = \sqrt{\mathsf{snr}}\, X + N \tag{1}$$

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

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

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) snr. It was also shown in [1] that (2) extends to the linear vector Gaussian channel

$$\boldsymbol{Y} = \sqrt{\mathsf{snr}}\, \mathbf{H}\boldsymbol{X} + \boldsymbol{N} \tag{3}$$

as

$$\frac{d}{d\mathsf{snr}} I\left(\boldsymbol{X}; \sqrt{\mathsf{snr}}\, \mathbf{H}\boldsymbol{X} + \boldsymbol{N}\right) = \mathbb{E}\left[\left\|\mathbf{H}\boldsymbol{X} - \mathbb{E}\left[\mathbf{H}\boldsymbol{X}\,|\sqrt{\mathsf{snr}}\, \mathbf{H}\boldsymbol{X} + \boldsymbol{N}\right]\right\|^2\right] \tag{4}$$

where the right-hand side is the expected squared Euclidean norm of the error in the estimation of $\mathbf{H}\boldsymbol{X}$. 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

$$\nabla_{\mathbf{H}} I\left(\boldsymbol{X}; \mathbf{H}\boldsymbol{X} + \boldsymbol{N}\right) = \mathbf{H}\mathbf{E} \tag{5}$$

where

$$\mathbf{E} \triangleq \mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X} \mid \boldsymbol{Y}\right]\right)\left(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X} \mid \boldsymbol{Y}\right]\right)^{\dagger}\right] \tag{6}$$

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: the mercury/waterfilling optimal power allocation over a set of parallel Gaussian channels [3]; the numerical optimization of linear precoders for multiple-input multiple-output (MIMO) channels [2]; a simple proof for the entropy power inequality [4]; a simple proof of the monotonicity of the non-Gaussianness of independent random variables [5]; and the study of extrinsic information of good codes [6]. Interestingly, as has been recently shown in [7], the derivative of the conditional entropy of the input given the output (or, equivalently, the mutual information) with respect to a channel

Manuscript received August 21, 2005; revised June 27, 2006. 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. The material in this paper was presented in part at the 43rd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, September 2005.

D. P. Palomar was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA. He is now with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: palomar@ust.hk).

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.

Color versions of Figs. 1–3 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2006.889728

0018-9448/$25.00 © 2007 IEEE

parameter can be used as a generalization of the extrinsic information transfer (EXIT) charts (called GEXIT charts) which has very appealing properties as a tool for analyzing the behavior of ensembles of codes using iterative decoding. Along the same lines, [8] analyzed mean-square error (MSE) charts as opposed to the traditional charts based on mutual information [9].

Counterparts of the fundamental relation have been explored for other types of channels; namely, for Poisson channels [10], for additive non-Gaussian channels [11], and for the discrete memoryless channel (DMC) [7, Theorem 1]. As should be expected, the MMSE does not play a role in the representation of mutual information for these channels.

Pursuing the connection between information theory and estimation theory found in [1] in the context of Gaussian channels, the goal of this paper is to generalize that link to arbitrary channels. Generalizing the aforementioned approaches, our main result gives the derivative of mutual information with respect to a channel parameter θ in terms of the input estimate given by the posterior distribution $P^{\theta}_{X|Y}$ as

$$\frac{\partial}{\partial\theta} I(X;Y) = \mathbb{E}\left[\frac{\partial \log_e P^{\theta}_{Y|X}(Y \mid X)}{\partial\theta} \log P^{\theta}_{X|Y}(X \mid Y)\right] \tag{7}$$

where $P^{\theta}_{Y|X}$ is an arbitrary random transformation and the expectation is with respect to the joint distribution $P_X P^{\theta}_{Y|X}$. For the particular case of a memoryless channel, the derivative is expressed in terms of the individual input estimates given by the posterior marginals $P^{\theta}_{X_i|Y^n}$ as

$$\frac{\partial}{\partial\theta} I(X^n;Y^n) = \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial \log_e P^{\theta}_{Y_i|X_i}(Y_i|X_i)}{\partial\theta} \log P^{\theta}_{X_i|Y^n}(X_i \mid Y^n)\right] \tag{8}$$

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

In addition to the theoretical interest of this characterization, it allows the efficient computation of the mutual information $I(X^n;Y^n)$ achieved by a given code over a 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 finds several applications, for example, in studying the concatenation of coding schemes [12], in lower-bounding the size of a code to achieve a desired block error rate or bit error rate, and in predicting the convergence behavior of iterative decoding schemes using EXIT charts [9].

Expressions for the capacity of coded systems over the binary symmetric channel (BSC) and the Gaussian channel were obtained in [12]; however, a numerical evaluation is only possible for very small codes. In [13]–[15], 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 Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm [16]. However, in a more general setting where the source is not a Markov process, the previous 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 $P_{X^n|Y^n}$.

Based on (8), 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 $P_{X^n|Y^n}$ or P_{Y^n}) or, equivalently, the symbol-wise *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 dynamic programming algorithm computes the exact posterior marginals [16]. In other cases, the posterior marginals can only be approximated such as in the turbo decoding for concatenated codes (e.g., [17]), the soft decoding of Reed–Solomon codes [18], and the *sum-product* algorithm for factor graphs (e.g., [19], [20]).

This paper is organized as follows. Section II first obtains the relation between mutual information and the input estimates in a general abstract setting which is then particularized for several channels of interest including the BSC, the binary erasure channel (BEC), the DMC, the scalar/vector Gaussian channel, an arbitrary additive-noise channel, and the Poisson channel. Section III considers practical issues of the computation of mutual information via the derivative based on a Monte Carlo method. Section IV provides numerical results by computing the mutual information achieved by low-density parity-check (LDPC) codes over the BSC and the Gaussian channel, and also by computing the information received by the reader of a novel with typos.

II. Derivative of Mutual Information

We first give a general representation for arbitrary random transformations, memoryless channels, and finite-state Markov channels and then particularize the results for specific types of channels such as the BSC, BEC, DMC, scalar/vector Gaussian channel, arbitrary additive-noise channel, and Poisson channel.

A. General Representation

We start with some notation that will allow us to express our results compactly and in full generality. Let the functions $f^{\theta}_{X|Y}$ and $f^{\theta}_{Y|X}$ denote the Radon–Nikodym derivatives of the probability measures $P^{\theta}_{X|Y}$ and $P^{\theta}_{Y|X}$ with respect to arbitrary measures Q_X and Q_Y such that $P^{\theta}_{X|Y} \ll Q_X$ and $P^{\theta}_{Y|X} \ll Q_Y$.[1]

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. These conditions are explicitly stated in Appendix A and

[1] In the continuous/discrete cases, $f^{\theta}_{X|Y}$ and $f^{\theta}_{Y|X}$ are simply probability density/mass functions.

will be implicitly assumed to be satisfied wherever necessary in the subsequent statement of the results.

The following intermediate result is key in the proof of the main result of the paper and is reminiscent of the regularity condition commonly invoked in estimation theory for the proof of the Cramer–Rao lower bound [21], [22].[2]

Lemma 1: Consider a random transformation $P^{\theta}_{Y|X}$, which is differentiable as a function of the real-valued parameter θ, and a random input with distribution P_X (independent of θ). Then

i)
$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f^{\theta}_{X|Y}(X \mid y) \mid Y=y\right]=0 \tag{9}$$
where the expectation is with respect to $P^{\theta}_{X|Y=y}$, and

ii)
$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f^{\theta}_{Y|X}(Y \mid x) \mid X=x\right]=0 \tag{10}$$
where the expectation is with respect to $P^{\theta}_{Y|X=x}$.

Proof: See Appendix B. □

The following result characterizes the derivative of the mutual information for an arbitrary random transformation with arbitrary input and output alphabets.

Theorem 1: Consider a random transformation $P^{\theta}_{Y|X}$, which is differentiable with respect to θ, and a random input with distribution P_X (independent of θ). Then, the derivative of the mutual information $I(X;Y)$ with respect to θ can be written in terms of the posterior distribution $P^{\theta}_{X|Y}$ as[3]

$$\frac{\partial}{\partial\theta}I(X;Y)=\mathbb{E}\left[\frac{\partial\log_e f^{\theta}_{Y|X}(Y \mid X)}{\partial\theta}\log f^{\theta}_{X|Y}(X \mid Y)\right] \tag{11}$$

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

Proof: Choose an arbitrary $Q_X \gg P_X$. First decompose the mutual information as

$$I(X;Y)=D\left(P^{\theta}_{X|Y} \parallel Q_X \mid P^{\theta}_{Y}\right)-D(P_X \parallel Q_X). \tag{12}$$

Then, since neither P_X nor Q_X depend on θ

$$\begin{aligned}
&\frac{\partial}{\partial\theta}I(X;Y)\\
&\quad=\frac{\partial}{\partial\theta}D\left(P^{\theta}_{X|Y} \parallel Q_X \mid P^{\theta}_{Y}\right)\\
&\quad=\frac{\partial}{\partial\theta}\mathbb{E}\left[\log\frac{dP^{\theta}_{X|Y}}{dQ_X}\right]\\
&\quad=\frac{\partial}{\partial\theta}\int\int\log\left(\frac{dP^{\theta}_{X|Y}}{dQ_X}\right)\frac{dP^{\theta}_{Y|X}}{dQ_Y}dQ_Y dP_X\\
&\quad=\int\int\frac{\partial}{\partial\theta}\left(\log\frac{dP^{\theta}_{X|Y}}{dQ_X}\right)dP^{\theta}_{Y|X}dP_X\\
&\qquad+\int\int\log\left(\frac{dP^{\theta}_{X|Y}}{dQ_X}\right)\frac{\partial}{\partial\theta}\left(\frac{dP^{\theta}_{Y|X}}{dQ_Y}\right)dQ_Y dP_X\\
&\quad=\mathbb{E}\left[\frac{\partial}{\partial\theta}\log\frac{dP^{\theta}_{X|Y}}{dQ_X}\right]\\
&\qquad+\int\int\log\left(\frac{dP^{\theta}_{X|Y}}{dQ_X}\right)\frac{\partial}{\partial\theta}\left(\log_e\frac{dP^{\theta}_{Y|X}}{dQ_Y}\right)dP^{\theta}_{Y|X}dP_X\\
&\quad=0+\mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\log_e\frac{dP^{\theta}_{Y|X}}{dQ_Y}\right)\log\left(\frac{dP^{\theta}_{X|Y}}{dQ_X}\right)\right]
\end{aligned} \tag{13}$$

where the regularity condition RC3 in Appendix A has been used for the interchange of the order of differentiation and integral and Lemma 1 has been invoked. □

The following result gives an alternative expression for the derivative of the mutual information which applies to many parameterizations of interest in applications.

Theorem 2: Consider the setup of Theorem 1 and further assume that the output alphabet is continuous and that the derivative of the random transformation, with probability density function (pdf) $f^{\theta}_{Y|X}$, factorizes as

$$\frac{\partial f^{\theta}_{Y|X}(y \mid x)}{\partial\theta}=-\phi^{\theta}(x)\frac{\partial f^{\theta}_{Y|X}(y \mid x)}{\partial y}\quad\text{for a.e. } x. \tag{14}$$

Then

$$\frac{\partial}{\partial\theta}I(X;Y)=\mathbb{E}\left[\phi^{\theta}(X)\frac{\partial\log f^{\theta}_{X|Y}(X \mid Y)}{\partial y}\right]. \tag{15}$$

Proof: From Theorem 1 and the factorization in (14), we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}I(X;Y)\\
&\quad=-\mathbb{E}\left[\phi^{\theta}(X)\frac{\partial\log_e f^{\theta}_{Y|X}(Y \mid X)}{\partial y}\log f^{\theta}_{X|Y}(X \mid Y)\right]\\
&\quad=-\mathbb{E}\left[\phi^{\theta}(X)\,\mathbb{E}\left[\frac{\partial\log_e f^{\theta}_{Y|X}(Y \mid X)}{\partial y}\log f^{\theta}_{X|Y}(X \mid Y) \mid X\right]\right]\\
&\quad=-\mathbb{E}\left[\phi^{\theta}(X)\int\frac{\partial f^{\theta}_{Y|X}(y \mid X)}{\partial y}\log f^{\theta}_{X|Y}(X \mid y)\,dy\right]\\
&\quad=\mathbb{E}\left[\phi^{\theta}(X)\int f^{\theta}_{Y|X}(y \mid X)\frac{\partial\log f^{\theta}_{X|Y}(X \mid y)}{\partial y}dy\right]\\
&\quad=\mathbb{E}\left[\phi^{\theta}(X)\frac{\partial\log f^{\theta}_{X|Y}(X \mid Y)}{\partial y}\right]
\end{aligned} \tag{16}$$

where we have integrated by parts:

$$\begin{aligned}
&\int\frac{\partial f^{\theta}_{Y|X}(y \mid x)}{\partial y}\log f^{\theta}_{X|Y}(x \mid y)\,dy\\
&\qquad=f^{\theta}_{Y|X}(y \mid x)\log f^{\theta}_{X|Y}(x \mid y)\Big]^{+\infty}_{-\infty}\\
&\qquad\quad-\int f^{\theta}_{Y|X}(y \mid x)\frac{\partial\log f^{\theta}_{X|Y}(x \mid y)}{\partial y}dy\\
&\qquad=0-\int f^{\theta}_{Y|X}(y \mid x)\frac{\partial\log f^{\theta}_{X|Y}(x \mid y)}{\partial y}dy
\end{aligned} \tag{17}$$

[2]In fact, (10) is exactly the regularity condition appearing in the proof of the Cramer–Rao lower bound in classical estimation theory that implies that the "score function" has zero mean [21], [22].

[3]Unless the logarithm basis is indicated, it can be chosen arbitrarily as long as both sides of the equation have the same units.

and the following result, whose proof we omit, has been used:

$$\lim_{|y|\to\infty} f^{\theta}_{Y|X}(y \mid x) \log f^{\theta}_{X|Y}(x \mid y) = 0. \tag{18}$$

□

Observe that the factorization in (14) holds, for example, for the additive-noise channel (cf. Section II-G). In the sequel, we will exhibit the utility of the alternative expressions given by Theorems 1 and 2. For example, in terms of numerical computation, the expression in Theorem 1 seems to be preferable (cf. Sections III and IV), whereas in terms of relating the mutual information with well-known concepts in estimation theory,the expression in Theorem 2 turns out to be more convenient (cf. Section II-E).

Theorem 1 can be readily particularized to the case of an arbitrary channel with transition probability $P^{\theta}_{Y^n|X^n}$, 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 3: Consider a memoryless channel with transition probability $P^{\boldsymbol{\theta}}_{Y^n|X^n} = \prod_{i=1}^{n} P^{\theta_i}_{Y_i|X_i}$, where $P^{\theta_i}_{Y_i|X_i}$ is differentiable as a function of the parameter θ_i (and independent of θ_j for $j \neq i$), and a random input with distribution P_{X^n} (independent of θ_i for all i). Then, the derivative of the mutual information $I(X^n;Y^n)$ with respect to θ_i can be written in terms of the posterior marginal distribution $P^{\boldsymbol{\theta}}_{X_i|Y^n}$ as

$$\frac{\partial}{\partial\theta_i} I(X^n;Y^n) = \mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid X_i)}{\partial\theta_i} \log f^{\boldsymbol{\theta}}_{X_i|Y^n}(X_i \mid Y^n)\right] \tag{19}$$

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

Proof: First, observe that, due to the memoryless assumption

$$\frac{\partial \log f^{\boldsymbol{\theta}}_{Y^n|X^n}(y^n \mid x^n)}{\partial\theta_i} = \sum_j \frac{\partial \log f^{\theta_j}_{Y_j|X_j}(y_j \mid x_j)}{\partial\theta_i} = \frac{\partial \log f^{\theta_i}_{Y_i|X_i}(y_i \mid x_i)}{\partial\theta_i}. \tag{20}$$

Now, from Theorem 1,[4] we have (21) at the bottom of the page, where we have used

$$P^{\boldsymbol{\theta}}_{X^n|Y^n} = P^{\boldsymbol{\theta}}_{X_i|Y^n} P^{\boldsymbol{\theta}}_{X^{n\setminus i}|X_i,Y^n} = P^{\boldsymbol{\theta}}_{X_i|Y^n} P^{\boldsymbol{\theta}}_{X^{n\setminus i}|X_i,Y^{n\setminus i}} \tag{22}$$

from which the same relation follows for the Radon–Nikodym derivatives. The final result follows by noting that the last term is zero (using the memoryless property of the channel) as shown in (23) at the bottom of the page, where we have invoked Lemma 1 to obtain

$$\mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid x_i)}{\partial\theta_i \mid X_i = x_i}\right] = 0. \qquad \square$$

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

$$\frac{\partial}{\partial\theta} I(X^n;Y^n) = \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial \log_e f^{\theta}_{Y|X}(Y_i \mid X_i)}{\partial\theta} \log f^{\theta}_{X_i|Y^n}(X_i \mid Y^n)\right]. \tag{24}$$

An alternative expression of the derivative for the memoryless channel can be given, similarly to Theorem 2, as

$$\frac{\partial}{\partial\theta_i} I(X^n;Y^n) = \mathbb{E}\left[\phi^{\theta_i}_i(X_i) \frac{\partial \log f^{\boldsymbol{\theta}}_{X_i|Y^n}(X_i \mid Y^n)}{\partial y_i}\right]. \tag{25}$$

The key relation (2), derived in [1], can be found as the particularization of (25) to the Gaussian channel (cf. Section II-E).

An interesting application of Theorem 3 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 block length 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 under an equiprobable choice of codewords). Indeed, the practical relevance of Theorem 3 for numerical computation is remarkable since, as already mentioned, the symbolwise APP $P_{X_i|Y^n}$ obtained by an optimum

[4]We denote by $x^{n\setminus i}$ the sequence $x^n = (x_1, \ldots, x_n)$ except the element x_i.

$$\begin{aligned}\frac{\partial}{\partial\theta_i} I(X^n;Y^n) &= \mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid X_i)}{\partial\theta_i} \log f^{\boldsymbol{\theta}}_{X^n|Y^n}(X^n \mid Y^n)\right] \\ &= \mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid X_i)}{\partial\theta_i} \log f^{\boldsymbol{\theta}}_{X_i|Y^n}(X_i \mid Y^n)\right] \\ &\quad + \mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid X_i)}{\partial\theta_i} \log f^{\boldsymbol{\theta}}_{X^{n\setminus i}|X_i,Y^{n\setminus i}}\left(X^{n\setminus i} \mid X_i, Y^{n\setminus i}\right)\right]\end{aligned} \tag{21}$$

$$\mathbb{E}_{X^n}\left[\mathbb{E}_{Y_i|X_i}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_i}(Y_i \mid X_i)}{\partial\theta_i}\right] \mathbb{E}_{Y^{n\setminus i}|X^{n\setminus i}}\left[\log f^{\boldsymbol{\theta}}_{X^{n\setminus i}|X_i,Y^{n\setminus i}}\left(X^{n\setminus i} \mid X_i, Y^{n\setminus i}\right)\right]\right] = 0 \tag{23}$$

soft decoder can be efficiently computed in practice with a message-passing algorithm such as the BCJR, sum-product, or belief-propagation algorithms [19], [20]. 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 inner expectation over X_i conditioned on Y^n through $P_{X_i|Y^n}$ (cf. Section III); then, for a finite input alphabet, (24) becomes (26), shown at the bottom of the page.

Using a similar proof, the result in Theorem 3 for a memoryless channel can be easily extended to a finite-state Markov channel as follows.

Theorem 4: Consider a finite-state Markov channel of memory L with transition probability

$$P^{\boldsymbol{\theta}}_{Y^n|X^n} = \prod_{i=1}^{n} P^{\theta_i}_{Y_i|X_{i-L}^{i}}$$

where $P^{\theta_i}_{Y_i|X_{i-L}^{i}}$ is differentiable as a function of the parameter θ_i (and independent of θ_j for $j \neq i$), and a random input with distribution P_{X^n} (independent of θ_i for all i). Then

$$\frac{\partial}{\partial\theta_i} I\left(X^n;Y^n\right) = \mathbb{E}\left[\frac{\partial \log_e f^{\theta_i}_{Y_i|X_{i-L}^{i}}\left(Y_i \,|\, X_{i-L}^{i}\right)}{\partial\theta_i} \log f^{\boldsymbol{\theta}}_{X_{i-L}^{i}|Y^n}\left(X_{i-L}^{i} \mid Y^n\right)\right] \tag{27}$$

where the expectation is with respect to the joint distribution $P_{X_{i-L}^{i}} P^{\boldsymbol{\theta}}_{Y^n|X_{i-L}^{i}}$.

The following is a convenient result that relates the posterior marginals $P_{X_i|Y^n}$ given all the observations y^n with the posterior marginals $P_{X_i|Y^{n\backslash i}}$ given all observations but the ith one $y^{n\backslash i}$ (sometimes known as *extrinsic information*).

Lemma 2: Consider a memoryless channel

$$P_{Y^n|X^n} = \prod_i P_{Y_i|X_i}$$

and an arbitrary input P_{X^n}. Then, the posterior marginals $P_{X_i|Y^n}$ and $P_{X_i|Y^{n\backslash i}}$ are related as follows:

i)

$$P_{X_i|Y^n}(x_i\,|\,y^n) = \frac{P_{X_i|Y^{n\backslash i}}\left(x_i\,|\,y^{n\backslash i}\right) P_{Y_i|X_i}(y_i\,|\,x_i)}{\mathbb{E}\left[P_{Y_i|X_i}\left(y_i \mid X_i\right) \mid Y^{n\backslash i} = y^{n\backslash i}\right]}, \quad \forall x_i. \tag{28}$$

ii)

$$\log \frac{P_{X_i|Y^n}\left(a \mid y^n\right)}{P_{X_i|Y^n}\left(b \mid y^n\right)} = \log \frac{P_{X_i|Y^{n\backslash i}}\left(a \mid y^{n\backslash i}\right)}{P_{X_i|Y^{n\backslash i}}\left(b \mid y^{n\backslash i}\right)} + \log \frac{P_{Y_i|X_i}\left(y_i \mid a\right)}{P_{Y_i|X_i}\left(y_i \mid b\right)}, \quad \forall a,b : P_{Y_i|X_i}\left(y_i \mid a\right), P_{Y_i|X_i}\left(y_i \mid b\right) > 0. \tag{29}$$

Proof: Both follow from

$$P_{X_i|Y^n} P_{Y^n} = P_{X_i,Y^n} = P_{X_i,Y^{n\backslash i}} P_{Y_i|X_i} = P_{X_i|Y^{n\backslash i}} P_{Y^{n\backslash i}} P_{Y_i|X_i}. \tag{30}$$

□

B. Binary Symmetric Channel (BSC)

The derivative of the mutual information of an arbitrary input over a BSC can be computed in practice by generating realizations of X^n and Y^n to approximate the expectation in Theorem 3 particularized to

$$\frac{d \log_e P^{\delta}_{Y|X}\left(y_i \mid x_i\right)}{d\delta} = \frac{x_i \oplus y_i}{\delta} - \frac{1 - x_i \oplus y_i}{1-\delta} \tag{31}$$

where δ is the channel crossover probability and $\oplus$ denotes the XOR operation or sum in modulo 2. However, to speed up the convergence of such an approximation, the expectation can be partially carried out analytically over X_i as in (26). The following result refines Theorem 3 for the BSC by carrying out analytically the expectation over both X_i and Y_i.

Theorem 5: Consider a BSC with crossover probability $\delta \in (0,1)$ and input distribution P_{X^n}. Then,[5] we get (32) at the bottom of the page, where

$$\lambda_i\left(y^{n\backslash i}\right) \triangleq \log \frac{P^{\delta}_{X_i|Y^{n\backslash i}}\left(0 \mid y^{n\backslash i}\right)}{P^{\delta}_{X_i|Y^{n\backslash i}}\left(1 \mid y^{n\backslash i}\right)} \tag{33}$$

and

$$\gamma \triangleq \log \frac{1-\delta}{\delta}. \tag{34}$$

Proof: See Appendix C. □

[5]The base of the exponential operator $\exp(\cdot)$ in (32) and similar expressions can be chosen arbitrarily as long as both sides of the equation have the same units.

$$\frac{\partial}{\partial\theta} I\left(X^n;Y^n\right) = \sum_{i=1}^{n} \mathbb{E}\left[\sum_{x_i} P^{\theta}_{X_i|Y^n}\left(x_i \mid Y^n\right) \frac{\partial \log_e f^{\theta}_{Y|X}\left(Y_i \mid x_i\right)}{\partial\theta} \log P^{\theta}_{X_i|Y^n}\left(x_i \mid Y^n\right)\right]. \tag{26}$$

$$\frac{d}{d\delta} I\left(X^n;Y^n\right) = \sum_{i=1}^{n} \left(\mathbb{E}\left[\tanh\left(\frac{\lambda_i\left(Y^{n\backslash i}\right)}{2\log e}\right) \log\left(\frac{\exp\left(\lambda_i\left(Y^{n\backslash i}\right)\right) + \exp\left(-\gamma\right)}{\exp\left(\lambda_i\left(Y^{n\backslash i}\right)\right) + \exp\left(\gamma\right)}\right)\right] - 2\gamma P_{X_i}\left(1\right)\right) \tag{32}$$

Using Lemma 2, the log-likelihood ratio in (33) can be rewritten as

$$\lambda_i\left(y^{n\backslash i}\right)=\log\frac{P_{X_i|Y^n}\left(0\mid y^n\right)}{P_{X_i|Y^n}\left(1\mid y^n\right)}-\gamma(-1)^{y_i}. \tag{35}$$

The following result particularizes Theorem 5 to the case of independent and identically distributed (i.i.d.) inputs, which alternatively can be easily obtained from direct computation of the mutual information.

Corollary 1: Consider i.i.d. inputs with distribution P_X over a BSC with crossover probability $\delta\in(0,1)$. Then

$$\frac{d}{d\delta}I(X;Y)=(1-2P_X(1))\log\left(\frac{P_Y(0)}{P_Y(1)}\right)-\gamma. \tag{36}$$

C. Binary Erasure Channel (BEC)

The following result refines Theorem 3 for the BEC by carrying out analytically the expectation over both X_i and Y_i.

Theorem 6: Consider a BEC with erasure probability $\epsilon\in(0,1)$ and input distribution P_{X^n}. Then, we get the equation at the bottom of the page, where

$$\lambda_i\left(y^{n\backslash i}\right)\triangleq\log\frac{P^{\epsilon}_{X_i|Y^{n\backslash i}}\left(0\mid y^{n\backslash i}\right)}{P^{\epsilon}_{X_i|Y^{n\backslash i}}\left(1\mid y^{n\backslash i}\right)}. \tag{37}$$

Proof: Similar to the proof of Theorem 5 (also a straightforward particularization of the more general result for the DMC in Theorem 7 combined with (42)). □

The following result particularizes Theorem 6 to the case of i.i.d. inputs.

Corollary 2: Consider i.i.d. inputs with distribution P_X over a BEC with erasure probability ϵ. Then

$$\frac{d}{d\epsilon}I(X;Y)=-h(P_X(1)) \tag{38}$$

where $h(p)\triangleq-p\log p-(1-p)\log(1-p)$ is the binary entropy function.

D. 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

$$P_{Y^n|X^n}\left(y^n\mid x^n\right)=\prod_{i=1}^{n}P_{Y|X}\left(y_i\mid x_i\right).$$

The channel transition probability can be compactly described by the channel transition matrix $\mathbf{\Pi}$ with (k,l)th element defined as $\pi_{kl}=P_{Y|X}\left(b_k\mid a_l\right)$.

The expectation in Theorem 3 particularizes for the DMC (using Lemma 2) to

$$\begin{aligned}&\frac{\partial}{\partial\theta}I\left(X^n;Y^n\right)\\&=\sum_{i=1}^{n}\sum_{x_i,y_i,y^{n\backslash i}}P_{X_i}(x_i)\,P^{\theta}_{Y^{n\backslash i}|X_i}\left(y^{n\backslash i}\mid x_i\right)\frac{\partial P^{\theta}_{Y|X}\left(y_i\mid x_i\right)}{\partial\theta}\\&\quad\log\left(\frac{P^{\theta}_{X_i|Y^{n\backslash i}}\left(x_i\mid y^{n\backslash i}\right)P^{\theta}_{Y|X}\left(y_i\mid x_i\right)}{\sum_{\tilde{x}_i}P^{\theta}_{X_i|Y^{n\backslash i}}\left(\tilde{x}_i\mid y^{n\backslash i}\right)P^{\theta}_{Y|X}\left(y_i\mid\tilde{x}_i\right)}\right).\end{aligned} \tag{39}$$

An equivalent form of (39) was independently obtained in [7, Theorem 1] where the conditioning is with respect to an extrinsic information random variable Z_i (sufficient statistic of $Y^{n\backslash i}$). The convergence analysis of the decoding of LDPC code ensembles is carried out in [7] 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 3 for the DMC by carrying out analytically the expectation over both X_i and Y_i.

Theorem 7: Consider a DMC with channel transition matrix $\mathbf{\Pi}$ and input distribution P_{X^n}. Then, provided that $\pi_{kl}>0$ [6]

$$\begin{aligned}&\left[\nabla_{\mathbf{\Pi}}I\left(X^n;Y^n\right)\right]_{kl}\\&=-\sum_{i=1}^{n}\mathbb{E}\left[\frac{\log\left(1+\sum_{m\neq l}\frac{\pi_{km}}{\pi_{kl}}\exp\left(\lambda_i^{(m,l)}\left(Y^{n\backslash i}\right)\right)\right)}{1+\sum_{m\neq l}\exp\left(\lambda_i^{(m,l)}\left(Y^{n\backslash i}\right)\right)}\right]\end{aligned} \tag{40}$$

where

$$\lambda_i^{(m,l)}\left(y^{n\backslash i}\right)\triangleq\log\frac{P_{X_i|Y^{n\backslash i}}\left(a_m\mid y^{n\backslash i}\right)}{P_{X_i|Y^{n\backslash i}}\left(a_l\mid y^{n\backslash i}\right)}. \tag{41}$$

Proof: See Appendix D. □

The usefulness of the gradient in Theorem 7 is as an intermediate step in the computation of the derivative with respect to an arbitrary parameter θ via the chain rule for differentiation

$$\frac{\partial}{\partial\theta}I\left(X^n;Y^n\right)=\mathrm{Tr}\left(\nabla^T_{\mathbf{\Pi}}I\left(X^n;Y^n\right)\frac{\partial\mathbf{\Pi}}{\partial\theta}\right) \tag{42}$$

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

[6]The gradient with respect to a matrix $\nabla_{\mathbf{M}}f$ is defined as $[\nabla_{\mathbf{M}}f]_{ij}\triangleq\partial f/\partial[\mathbf{M}]_{ij}$.

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

Using Lemma 2, the log-likelihood ratio in (41) can be rewritten as

$$\lambda_i^{(m,l)}\left(y^{n\backslash i}\right)=\log\frac{P_{X_i|Y^n}\left(a_m\mid y^n\right)}{P_{X_i|Y^n}\left(a_l\mid y^n\right)}-\log\frac{\pi_{km}}{\pi_{kl}} \tag{43}$$

where k is such that $b_k=y_i$.

The following result particularizes Theorem 7 to the case of i.i.d. inputs.

Corollary 3: Consider i.i.d. inputs with distribution P_X over a DMC with channel transition matrix $\mathbf{\Pi}$. Then, provided that $\pi_{kl}>0$,

$$\left[\nabla_{\mathbf{\Pi}}I\left(X;Y\right)\right]_{kl}=P_X\left(a_l\right)\log P_{X|Y}\left(a_l\mid b_k\right). \tag{44}$$

E. Scalar Gaussian Channel

Consider the Gaussian channel signal model in (1) for the real-valued case (with a standard Gaussian noise). This channel has the following transition probability:

$$P_{Y|X}\left(y\mid x\right)=\frac{1}{\sqrt{2\pi}}e^{-\left(y-\sqrt{\mathsf{snr}}x\right)^2/2}. \tag{45}$$

Particularizing Theorem 2 to the memoryless real-valued Gaussian channel with arbitrary input with distribution P_{X^n} (with finite second-order moments) we recover the result in [1]

$$\frac{d}{d\mathsf{snr}}I\left(X^n;Y^n\right)=\frac{\log e}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left|X_i-\mathbb{E}\left[X_i\mid Y^n\right]\right|^2\right]. \tag{46}$$

As an illustration, which is useful in Section IV, the binary $\{\pm1\}$ input distribution P_{X^n} yields (either as a particularization of (46) or of Theorem 3) (47) at the bottom of the page, where

$$\lambda_i\left(y^{n\backslash i}\right)\triangleq\log\frac{P_{X_i|Y^{n\backslash i}}\left(-1\mid y^{n\backslash i}\right)}{P_{X_i|Y^{n\backslash i}}\left(+1\mid y^{n\backslash i}\right)}, \tag{48}$$

and (49), also at the bottom of the page.

Using Lemma 2, the log-likelihood ratio in (48) can be rewritten as

$$\lambda_i\left(y^{n\backslash i}\right)=\log\frac{P_{X_i|Y^n}\left(-1\mid y^n\right)}{P_{X_i|Y^n}\left(+1\mid y^n\right)}+2y_i\sqrt{\mathsf{snr}}\log e. \tag{50}$$

F. Linear Vector Gaussian Channel

Consider now the following signal model corresponding to a linear vector Gaussian channel with n_T transmit dimensions and n_R receive dimensions

$$\boldsymbol{Y}=\mathbf{H}\boldsymbol{X}+\boldsymbol{N} \tag{51}$$

where all quantities are complex-valued, $\boldsymbol{X}$ is the n_T-dimensional transmitted vector, $\mathbf{H}$ is the $n_R\times n_T$ matrix that denotes the linear transformation undergone by the signal, $\boldsymbol{Y}$ is the n_R-dimensional received vector, and $\boldsymbol{N}$ is an n_R-dimensional proper complex Gaussian noise vector independent of $\boldsymbol{X}$. The input and the noise are assumed to have zero mean and covariance matrices denoted by $\mathbf{\Sigma}$ and $\mathbf{\Phi}$, respectively. Observe that the signal model in (51) is a generalization of (1).

We particularize Theorem 2 for the complex-valued memoryless linear vector Gaussian channel with arbitrary input with distribution $P_{\boldsymbol{X}^n}$ (with finite second-order moments) to recover the result in [2] [7]

$$\nabla_{\mathbf{H}}I\left(\boldsymbol{X}^n;\boldsymbol{Y}^n\right)=\left(\log e\right)\mathbf{\Phi}^{-1}\mathbf{H}\sum_{i=1}^{n}\mathbf{E}_i \tag{52}$$

where

$$\mathbf{E}_i\triangleq\mathbb{E}\left[\left(\boldsymbol{X}_i-\mathbb{E}\left[\boldsymbol{X}_i\mid\boldsymbol{Y}^n\right]\right)\left(\boldsymbol{X}_i-\mathbb{E}\left[\boldsymbol{X}_i\mid\boldsymbol{Y}^n\right]\right)^{\dagger}\right]. \tag{53}$$

G. Additive-Noise Channel

Consider the following signal model corresponding to an arbitrary additive-noise (not necessarily Gaussian) channel:

$$Y=s^{\theta}\left(X\right)+N \tag{54}$$

where N is an arbitrarily distributed noise with pdf P_N and $s^{\theta}\left(\cdot\right)$ is a deterministic signaling function of the input dependent on the parameter θ. The corresponding channel transition probability is

$$P_{Y|X}^{\theta}\left(y\mid x\right)=P_N\left(y-s^{\theta}\left(x\right)\right). \tag{55}$$

[7]For complex-valued variables, the gradient with respect to a matrix $\nabla_{\mathbf{M}}f$ is defined as $\left[\nabla_{\mathbf{M}}f\right]_{ij}\triangleq\partial f/\partial\left[\mathbf{M}^*\right]_{ij}$, where $df/dx^*\triangleq\left(\partial f/\partial\mathrm{Re}\left\{x\right\}+j\partial f/\partial\mathrm{Im}\left\{x\right\}\right)/2$.

$$\frac{d}{d\mathsf{snr}}I\left(X^n;Y^n\right)=\frac{-1}{2\sqrt{\mathsf{snr}}}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\Psi\left(\lambda_i\left(Y^{n\backslash i}\right);\mathsf{snr}\right)}{1+\exp\left(\lambda_i\left(Y^{n\backslash i}\right)\right)}+\frac{\Psi\left(-\lambda_i\left(Y^{n\backslash i}\right);\mathsf{snr}\right)}{1+\exp\left(-\lambda_i\left(Y^{n\backslash i}\right)\right)}\right] \tag{47}$$

$$\begin{aligned}\Psi\left(x;\mathsf{snr}\right)&\triangleq\mathbb{E}\left[N\log\left(1+\exp\left(x-2\sqrt{\mathsf{snr}}\left(\sqrt{\mathsf{snr}}+N\right)\log e\right)\right)\right]\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\nu\log\left(1+\exp\left(x-2\sqrt{\mathsf{snr}}\left(\sqrt{\mathsf{snr}}+\nu\right)\log e\right)\right)e^{-\nu^2/2}d\nu.\end{aligned} \tag{49}$$

We can now invoke Theorem 3:

$$\frac{\partial}{\partial\theta}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial \log_e P_N\left(Y_i - s^\theta\left(X_i\right)\right)}{\partial\theta}\log P^\theta_{X_i|Y^n}\left(X_i \mid Y^n\right)\right]. \tag{56}$$

Also, noting that

$$\frac{\partial P^\theta_{Y|X}\left(y \mid x\right)}{\partial\theta} = -\frac{\partial s^\theta\left(x\right)}{\partial\theta}\frac{\partial P^\theta_{Y|X}\left(y \mid x\right)}{\partial y}, \tag{57}$$

we can invoke Theorem 2:

$$\frac{\partial}{\partial\theta}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial s^\theta\left(X_i\right)}{\partial\theta}\frac{\partial \log P^\theta_{X_i|Y^n}\left(X_i \mid Y^n\right)}{\partial y_i}\right]. \tag{58}$$

This result is further refined in the next theorem.

Theorem 8: Consider an additive-noise channel with the transition probability in (55) and a random input with distribution P_{X^n}. Then

$$\frac{\partial}{\partial\theta}I\left(X^n;Y^n\right) = -\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial s^\theta\left(X_i\right)}{\partial\theta} \mid Y^n\right]\mathbb{E}\left[\frac{\partial \log P_N\left(N_i\right)}{\partial n_i} \mid Y^n\right]\right]. \tag{59}$$

Proof: See Appendix E. □

For i.i.d. inputs, Theorem 8 particularizes to the result obtained in [11].

H. Poisson Channel

The canonical Poisson random transformation with mean $\mathbb{E}\left[Y \mid X = x\right] = x$ is the probability mass function

$$P_{Y|X}\left(y \mid x\right) = \frac{1}{y!}x^y e^{-x} \tag{60}$$

where $x \in [0,\infty)$ and $y \in \{0,1,2,\ldots\}$. A general Poisson channel is similarly defined by a transformation whose output is a Poisson random variable conditioned on the input X with its mean equal to $\alpha X + \lambda$ with $\alpha, \lambda \geq 0$, i.e., X scaled by α plus a "dark current" λ

$$P_{Y|X}\left(y \mid x\right) = \frac{1}{y!}\left(\alpha x + \lambda\right)^y e^{-(\alpha x+\lambda)}. \tag{61}$$

The partial derivatives of this random transformation with respect to the parameters λ and α are

$$\frac{\partial}{\partial\lambda}P_{Y|X}\left(y \mid x\right) = P_{Y|X}\left(y \mid x\right)\left(\frac{y}{\alpha x+\lambda} - 1\right) \tag{62}$$

$$\frac{\partial}{\partial\alpha}P_{Y|X}\left(y \mid x\right) = P_{Y|X}\left(y \mid x\right)\left(\frac{y}{\alpha x+\lambda} - 1\right)x. \tag{63}$$

So the direct application of Theorem 3 gives

$$\frac{\partial}{\partial\lambda}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda} - 1\right)\log f_{X_i|Y^n}\left(X_i \mid Y^n\right)\right] \tag{64}$$

$$\frac{\partial}{\partial\alpha}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda} - 1\right)\log f_{X_i|Y^n}\left(X_i|Y^n\right)\right]. \tag{65}$$

The next result further refines these expressions.

Theorem 9: Consider a Poisson channel with the channel transition probability in (61) and a random input with distribution P_{X^n}. Then

$$\frac{\partial}{\partial\lambda}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}\left[\log\left(\alpha X_i+\lambda\right) - \log\mathbb{E}\left[\alpha X_i+\lambda \mid Y^n\right]\right] \tag{66}$$

$$\frac{\partial}{\partial\alpha}I\left(X^n;Y^n\right) = \sum_{i=1}^{n}\mathbb{E}[X_i\log\left(\alpha X_i+\lambda\right) - \mathbb{E}\left[X_i|Y^n\right]\log\mathbb{E}[\alpha X_i+\lambda|Y^n]]. \tag{67}$$

Proof: See Appendix F. □

For i.i.d. inputs, Theorem 9 particularizes to the results obtained in [10].

III. Computation of Mutual Information

A. Direct Computation of Mutual Information

The mutual information for an arbitrary channel under an arbitrary finite input alphabet can be directly approximated as

$$I\left(\theta\right) = \mathbb{E}\left[\log\left(\frac{P^\theta_{Y^n|X^n}\left(Y^n|X^n\right)}{P^\theta_{Y^n}\left(Y^n\right)}\right)\right] \approx \frac{1}{M}\sum_{x^n,y^n}\log\left(\frac{P^\theta_{Y^n|X^n}\left(y^n|x^n\right)}{\sum_{\tilde{x}^n}P_{X^n}\left(\tilde{x}^n\right)P^\theta_{Y^n|X^n}\left(y^n|\tilde{x}^n\right)}\right) \tag{68}$$

where the M realizations (x^n, y^n) are drawn independently from $P_{X^n}P^\theta_{Y^n|X^n}$. However, the computation of

$$P_{Y^n}\left(y^n\right) = \sum_{\tilde{x}^n}P_{X^n}\left(\tilde{x}^n\right)P_{Y^n|X^n}\left(y^n|\tilde{x}^n\right)$$

requires the enumeration of the whole input alphabet (or codebook in case of codes) for each realization (x^n, y^n), which grows exponentially with the size of the input vector n. This is clearly infeasible but for very short block lengths.

If instead we express the mutual information in terms of the posterior distribution $P^\theta_{X^n|Y^n}$, the problem does not become

easier due to the difficulty of computing the posterior distribution. Hence, the direct computation of the mutual information is generally infeasible.

B. Computation of Mutual Information Via its Derivative

The key point of using the derivative of mutual information as an intermediate step is the fact that it only depends on the channel transition probability and the posterior distribution of the input given the output. For memoryless channels, only the posterior *marginals* appear in the expression of the derivative even if the inputs are dependent (see Theorem 3), as in the case of the evaluation of the mutual information achieved by a code. Interestingly, the posterior marginals can be computed very efficiently in practice with a message-passing algorithm. Thus, thanks to the representation of the derivative of mutual information as a function of the posterior marginals it is possible to approximate the mutual information achieved by a code using a soft decoding algorithm.

From Theorem 3, the derivative of mutual information for a memoryless channel is shown in

$$\begin{aligned} I'(\theta) &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial \log_e P^{\theta}_{Y|X}(Y_i \mid X_i)}{\partial \theta} \log P^{\theta}_{X_i|Y^n}(X_i \mid Y^n)\right] \\ &\approx \sum_{i=1}^{n} \frac{1}{M} \sum_{j=1}^{M} \frac{\partial \log_e P^{\theta}_{Y|X}(y_i[j] \mid x_i[j])}{\partial \theta} \\ &\quad \cdot \log P^{\theta}_{X_i|Y^n}(x_i[j] \mid y^n[j]) \end{aligned} \tag{69}$$

where the M realizations $(x^n[j], y^n[j])$ are drawn independently from $P_{X^n} P^{\theta}_{Y^n|X^n}$. To compute the mutual information $I(\theta)$, we just need to know its value at some reference point θ^{ref} and then integrate

$$I(\theta) = I\left(\theta^{\mathrm{ref}}\right) + \int_{\theta^{\mathrm{ref}}}^{\theta} I'(\upsilon)\, d\upsilon. \tag{70}$$

In practice, the integral in (70) will be computed as a sum over a grid of θ's. It is worth mentioning that the samples drawn to compute $I'(\theta)$ at some given θ_0 can be reused to compute $I'(\theta)$ around a neighborhood of θ_0 using a change of measure (this is, in fact, the underlying idea of importance sampling [23])

$$\begin{aligned} \mathbb{E}^{\theta}\left[f^{\theta}(X,Y)\right] &= \mathbb{E}^{\theta_0}\left[w(X,Y;\theta) f^{\theta}(X,Y)\right] \\ &\approx \frac{1}{M} \sum_{i=1}^{M} w(x[i], y[i]; \theta) f^{\theta}(x[i], y[i]) \end{aligned} \tag{71}$$

where the M realizations $(x[i], y[i])$ are drawn independently from $P_X P^{\theta_0}_{Y|X}$ and the weight $w(x, y; \theta)$ is the correction term or change of measure given by

$$w(x, y; \theta) = \frac{P^{\theta}_{Y|X}(y \mid x)}{P^{\theta_0}_{Y|X}(y \mid x)}. \tag{72}$$

Notice that the combination of computation of derivative and integration in (70) can be performed in a multitude of ways. Not only can one choose a different parameter θ over which to integrate, but the range of values in the integration can be conveniently chosen such that the mutual information at the reference point θ^{ref} is easily computed. For example, for a BSC one can only choose θ as the crossover probability δ and a convenient reference point would be either $\delta^{\mathrm{ref}} = 0.5$ or $\delta^{\mathrm{ref}} = 0$. When we use the gradient of the mutual information with respect to a vector (or a matrix) parameter, the same approach can be followed using a line integral of the gradient.[8]

The main disadvantage of computing the mutual information through the derivative as in (70) is the fact that in a practical situation the knowledge of the derivative $I'(\theta)$ is noisy. In that respect, it is important to be able to compute $I(\theta)$ in a robust way from noisy measurements of $I'(\theta)$. This can be easily done as described in Appendix G.

C. Practical Aspects

Theorem 3 gives a closed-form expression for the derivative of mutual information. As mentioned earlier, when the input distribution has no particular structure, the expectation has to be approximated using a Monte Carlo approach as in (69). If the conditional distribution is available, it is advantageous to partially carry out the expectation in a semi-analytical way, for example, the expectation over X_i conditioned on Y^n, as in (26), or even the joint expectation over X_i and Y_i conditioned on $Y^{n\setminus i}$. A partial analytical expectation will make the estimation more accurate or, equivalently, will require fewer samples for the same accuracy.

We now illustrate four different levels in the computation of the derivative of the mutual information for the BSC:

1. Via the joint posterior distribution as in the general result for a random transformation in Theorem 1. For the BSC, the expression becomes as shown in (73) at the bottom of the page, where $d_H(x^n, y^n)$ denotes the Hamming distance between the sequences x^n and y^n.

[8]The line integral of a gradient $\int_{\mathbf{a}}^{\mathbf{b}} \nabla\varphi\, d\boldsymbol{\alpha} = \int_a^b \varphi(\boldsymbol{\alpha}(t))\, \boldsymbol{\alpha}'(t)\, dt$ is independent of the path $\boldsymbol{\alpha}(t)$ in any open connected set in which the gradient is continuous [24] and then $\int_{\mathbf{a}}^{\mathbf{b}} \nabla\varphi\, d\boldsymbol{\alpha} = \varphi(\mathbf{b}) - \varphi(\mathbf{a})$.

$$\frac{d}{d\delta} I(X^n; Y^n) = \mathbb{E}\left[\left(\frac{d_H(X^n, Y^n)}{\delta} - \frac{n - d_H(X^n, Y^n)}{1-\delta}\right) \log P^{\delta}_{X^n|Y^n}(X^n \mid Y^n)\right] \tag{73}$$

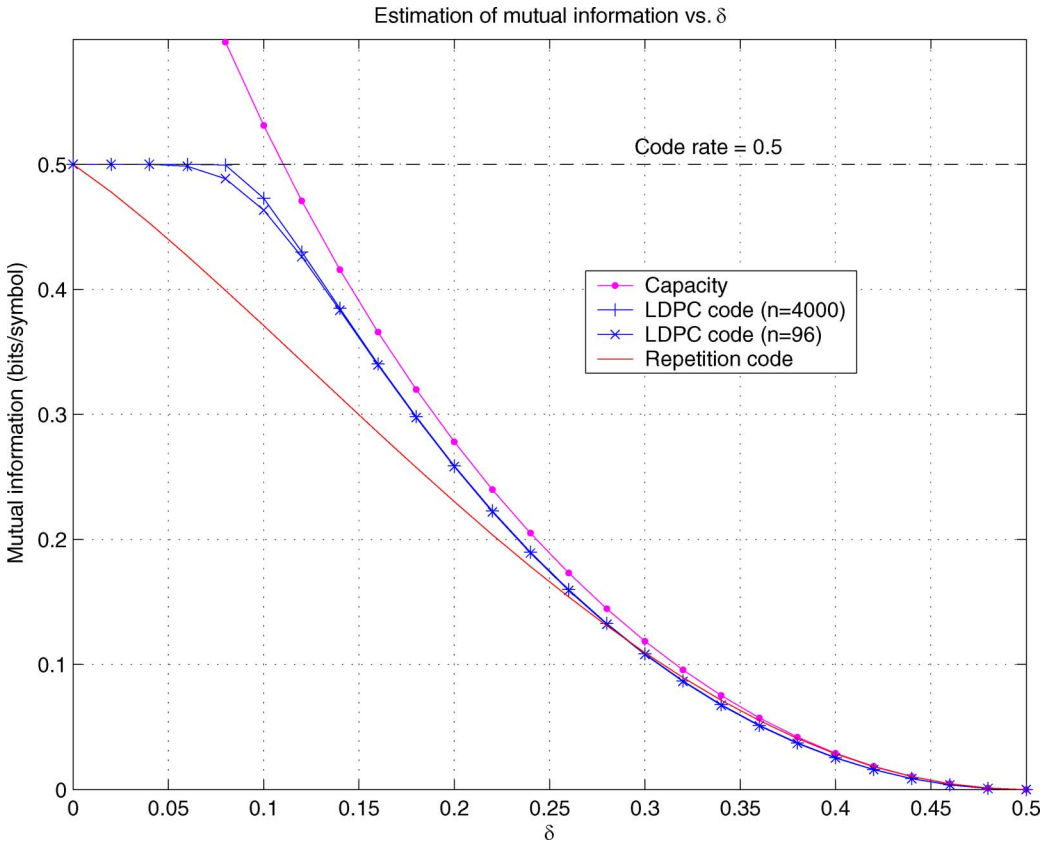

Fig. 1. Estimation of the mutual information of different codes of rate $1/2$ over a BSC.

2. Via the posterior marginal distributions as in Theorem 3. For the BSC, the expression is

$$\frac{d}{d\delta} I(X^n;Y^n) = \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{X_i \oplus Y_i}{\delta} - \frac{1-X_i \oplus Y_i}{1-\delta}\right) \log P^{\delta}_{X_i|Y^n}(X_i|Y^n)\right]. \tag{74}$$

3. Via the posterior marginal distributions but further taking the analytical expectation over X_i conditioned on Y^n (as in (26)).
4. Via the posterior marginal distributions but further taking the analytical expectation over both X_i and Y_i conditioned on $Y^{n\backslash i}$. In other words, via the log likelihood of the posterior marginal distributions as in Theorem 5 for the BSC.

The first method is infeasible but for very small n (as it requires the evaluation of the joint posterior), whereas the last three are feasible because they are based on the posterior marginals. Taking the analytical expectation over X_i, as in method 3, is not just desirable but extremely important[9] (taking also the expectation over Y_i, as in method 4, is simply an option to improve the convergence).

[9]The savings in the number of Monte Carlo samples required for a given accuracy of method 3 over 2 are measured in orders of magnitude.

IV. Applications

A. Computation of Mutual Information of LDPC Codes

We consider the computation of mutual information achieved by LDPC codes over the BSC and the Gaussian channel via the computation of its derivative, which can be efficiently done by estimating the posterior marginals for the LDPC codes with the sum-product algorithm.[10] For the BSC, the computation of the derivative of mutual information with respect to the crossover probability based on the belief propagation algorithm has been shown in [7] to underestimate the true value. Similarly, for the Gaussian channel, the derivative with respect to the SNR can be shown to overestimate the true value. This means that the computation based on the belief propagation algorithm provides upper and lower bounds on the mutual information depending on the reference point used in the integration in (70). We will use the robust formulation in (120) (Appendix G) to obtain a reliable estimation of the mutual information from the noisy knowledge of the derivative.

Fig. 1 shows the mutual information of different codes of rate $1/2$ over a BSC computed via Theorem 5. In particular, two LDPC codes with block lengths $n = 96$ and $n = 4000$ are considered as well as a simple repetition code. Naturally, for $\delta = 0$ all codes achieve a mutual information equal to the code

[10]The sum-product algorithm must not be terminated when a valid codeword has been found (as is done for decoding purposes) but when the estimations of the posterior marginals have converged.

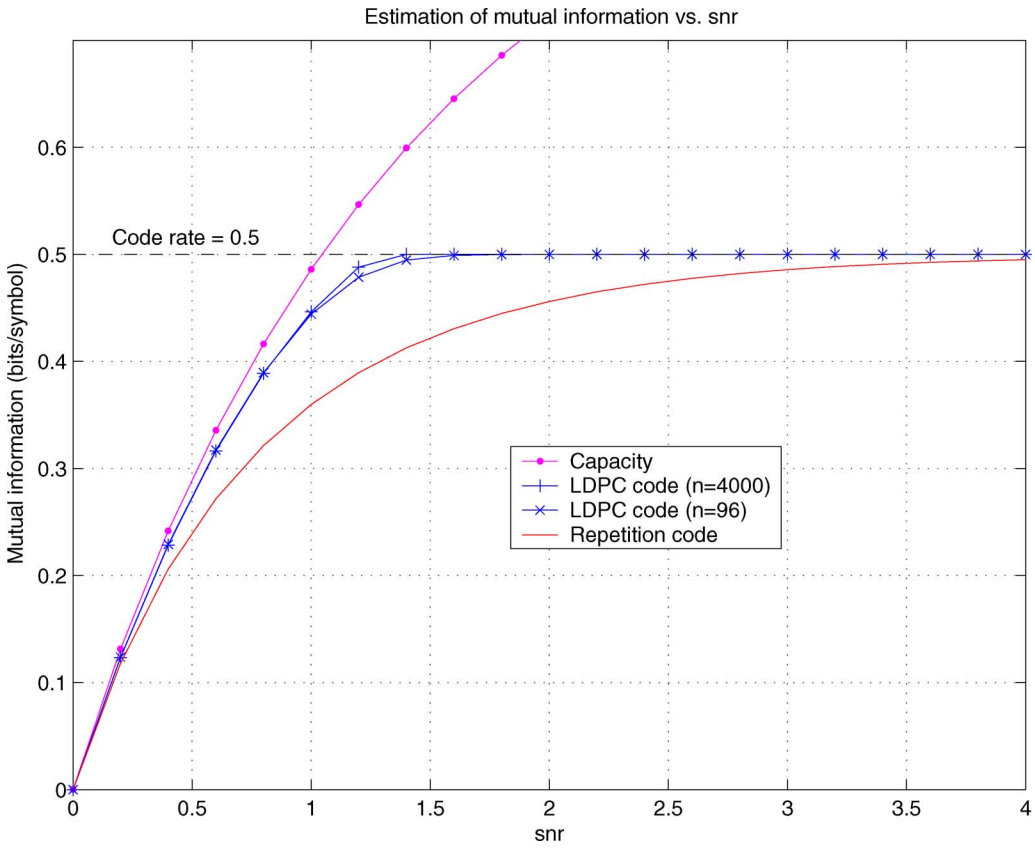

Fig. 2. Estimation of the mutual information of different binary codes of rate $1/2$ over an antipodal Gaussian channel.

rate. What is important is until what value of δ the code can still achieve that value of mutual information. Ideally, we know from the capacity curve that a rate of $1/2$ can be achieved up to a value of $\delta = 0.11$. The repetition code is clearly not a good code for any rate since its mutual information decreases considerably as soon as δ increases. The LDPC codes, on the other hand, are good codes since they do not suffer an appreciable decrease in mutual information up to a certain value δ_0; in particular, for the LDPC code with $n = 96$ we have $\delta_0 = 0.05$, whereas for the LDPC code with $n = 4000$, $\delta_0 = 0.08$. Interestingly, for larger values of δ, both LDPC codes achieve essentially the same mutual information; this means that, in that regime, using a long LDPC code is essentially equivalent to using a short one combined with an outer code (a similar observation was made in [25] for small codes with a sufficiently good decoder).

Fig. 2 shows the mutual information of the same codes over a Gaussian channel computed via (47). Similar observations hold as in the BSC: ideally, a rate $1/2$ can be achieved for SNRs above $\mathsf{snr} = 1.05$, the LDPC code with $n = 96$ does not suffer an appreciable decrease in mutual information down to $\mathsf{snr}_0 = 1.6$, whereas for the LDPC code with $n = 4000$, $\mathsf{snr}_0 = 1.4$.

B. 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 $P_{X_i|Y^{n\backslash i}}$ (e.g., [26], [27]) and then we apply Theorem 7 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.

To be more specific, the universal algorithms in [26], [27] first estimate $P_{Y_i|Y^{n\backslash i}}$. Then $P_{X_i|Y^{n\backslash i}}$ follows straightforwardly as

$$\begin{bmatrix} P_{X_i|Y^{n\backslash i}}\left(a_1 \mid y^{n\backslash i}\right) \\ \vdots \\ P_{X_i|Y^{n\backslash i}}\left(a_{|\mathcal{X}|} \mid y^{n\backslash i}\right) \end{bmatrix} = \left(\mathbf{\Pi}^T\mathbf{\Pi}\right)^{-1}\mathbf{\Pi}^T \begin{bmatrix} P_{Y_i|Y^{n\backslash i}}\left(b_1 \mid y^{n\backslash i}\right) \\ \vdots \\ P_{Y_i|Y^{n\backslash i}}\left(b_{|\mathcal{Y}|} \mid y^{n\backslash i}\right) \end{bmatrix} \quad (75)$$

which requires the channel transition probability matrix $\mathbf{\Pi}$ to be full column-rank (thus, it is assumed that $|\mathcal{X}| \leq |\mathcal{Y}|$).

The input (assumed to be stationary ergodic) is neither accessible nor statistically known and is only observed after passing through the channel. Theorem 7 (combined with (42)) can be

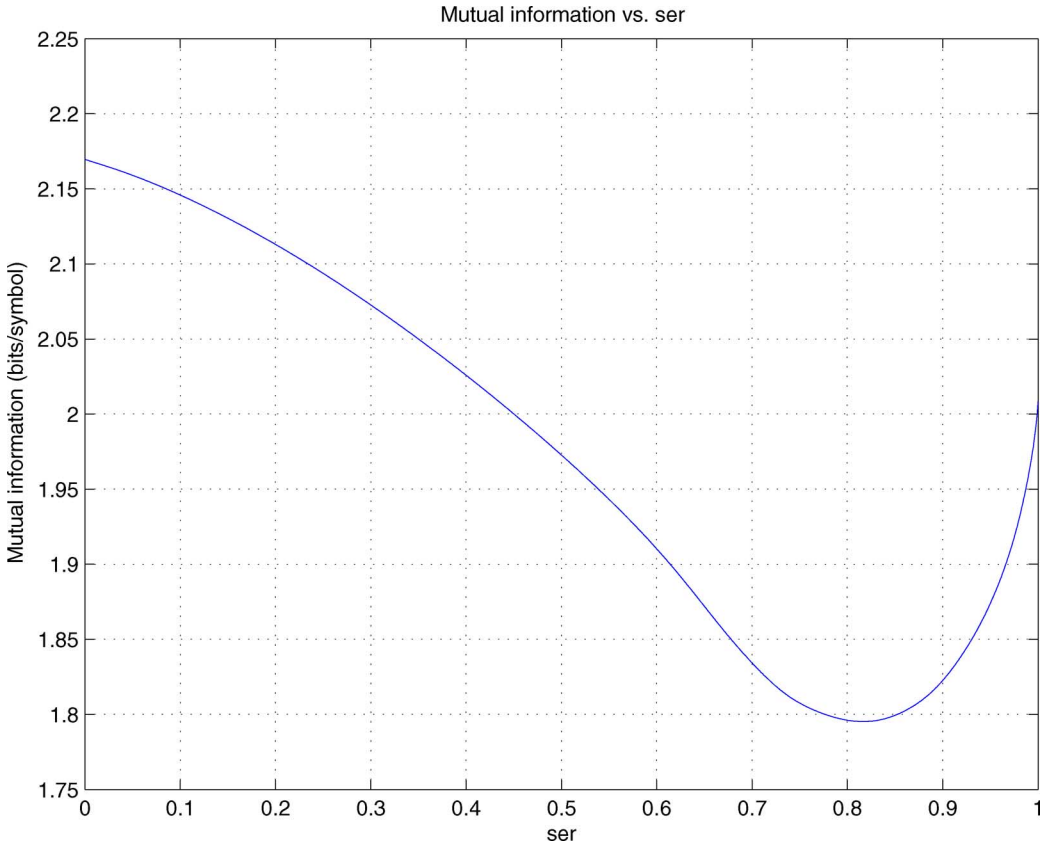

Fig. 3. Input–output mutual information of *Don Quixote de La Mancha* over the typewritter channel as a function of the symbol error probability.

more conveniently rewritten (due to the stationarity and ergodicity) as

$$\frac{\partial}{\partial\theta}\frac{1}{n}I\left(X^n;Y^n\right)\approx\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(\mathbf{R}_i^T\left(y^{n\backslash i}\right)\frac{\partial\mathbf{\Pi}}{\partial\theta}\right) \tag{76}$$

where

$$\left[\mathbf{R}_i\left(y^{n\backslash i}\right)\right]_{kl} = -\frac{\log\left(1+\sum_{m\neq l}\frac{\pi_{km}}{\pi_{kl}}\exp\left(\lambda_i^{(m,l)}\left(y^{n\backslash i}\right)\right)\right)}{1+\sum_{m\neq l}\exp\left(\lambda_i^{(m,l)}\left(y^{n\backslash i}\right)\right)} \tag{77}$$

$$=P_{X_i|Y^{n\backslash i}}\left(a_l\,|\,y^{n\backslash i}\right)\log\left(\frac{P_{X_i|Y^{n\backslash i}}\left(a_l\mid y^{n\backslash i}\right)\pi_{kl}}{\sum_m P_{X_i|Y^{n\backslash i}}\left(a_m\,|\,y^{n\backslash i}\right)\pi_{km}}\right) \tag{78}$$

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* [11] (in English translation) with typos. We model this channel by assuming that each letter is independently flipped, with some symbol error rate (SER) equal to ser, equiprobably into one of its nearest neighbors in the QWERTY keyboard. Fig. 3 shows the mutual information obtained by integrating the derivative from the point of reference $\mathsf{ser}=0$. For $\mathsf{ser}=0$, the mutual information equals the entropy of *Don Quixote de La Mancha* which is 2.17 bits/symbol as computed with the algorithm in [28].

[11]By Miguel de Cervantes Saavedra (1547–1616).

Appendix A
Regularity Conditions

The following are the "regularity conditions" about the interchange of the order of differentiation and integration (expectation) that are required for some results in the paper.

RC1:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{Q_Y}\left[f_{Y|X}^{\theta}\left(Y\mid x\right)\right]=\mathbb{E}_{Q_Y}\left[\frac{\partial}{\partial\theta}f_{Y|X}^{\theta}\left(Y\mid x\right)\right]. \tag{79}$$

RC2:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{P_X}\left[f_{Y|X}^{\theta}\left(y\mid X\right)\right]=\mathbb{E}_{P_X}\left[\frac{\partial}{\partial\theta}f_{Y|X}^{\theta}\left(y\mid X\right)\right]. \tag{80}$$

RC3:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{P_XQ_Y}\left[f_{Y|X}^{\theta}\left(Y\mid X\right)\log f_{X|Y}^{\theta}\left(X\mid Y\right)\right] =\mathbb{E}_{P_XQ_Y}\left[\frac{\partial}{\partial\theta}\left(f_{Y|X}^{\theta}\left(Y\mid X\right)\log f_{X|Y}^{\theta}\left(X\mid Y\right)\right)\right]. \tag{81}$$

Conditions RC1-RC3 are mild properties satisfied in most cases of interest such as for finite alphabets.

$$\mathbb{E}\left[\frac{d\log_e P^{\delta}_{Y|X}(Y_i \mid X_i)}{d\delta}\log P^{\delta}_{X_i|Y^n}(X_i \mid Y^n)\right]$$
$$= \mathbb{E}_{Y^{n\setminus i}}\mathbb{E}_{X_i|Y^{n\setminus i}}\mathbb{E}_{Y_i|X_i}\left[\frac{d\log_e P^{\delta}_{Y|X}(Y_i \mid X_i)}{d\delta}\log\left(\frac{P^{\delta}_{X_i|Y^{n\setminus i}}\left(X_i \mid Y^{n\setminus i}\right)P^{\delta}_{Y|X}(Y_i \mid X_i)}{\sum_{\tilde{x}_i}P^{\delta}_{X_i|Y^{n\setminus i}}\left(\tilde{x}_i \mid Y^{n\setminus i}\right)P^{\delta}_{Y|X}(Y_i \mid \tilde{x}_i)}\right)\right]$$
$$= -\mathbb{E}_{Y^{n\setminus i}}\mathbb{E}_{X_i|Y^{n\setminus i}}\left[\sum_{y_i}\frac{dP^{\delta}_{Y|X}(y_i \mid X_i)}{d\delta}\log\left(\frac{\sum_{\tilde{x}_i}P^{\delta}_{X_i|Y^{n\setminus i}}\left(\tilde{x}_i \mid Y^{n\setminus i}\right)P^{\delta}_{Y|X}(y_i \mid \tilde{x}_i)}{P^{\delta}_{X_i|Y^{n\setminus i}}\left(X_i \mid Y^{n\setminus i}\right)P^{\delta}_{Y|X}(y_i \mid X_i)}\right)\right]. \quad (86)$$

APPENDIX B
PROOF OF LEMMA 1

We first prove the second result

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log_e\frac{dP^{\theta}_{Y|X}}{dQ_Y}(x,Y) \mid X=x\right]$$
$$= \int\frac{dP^{\theta}_{Y|X}}{dQ_Y}(x,y)\frac{\partial}{\partial\theta}\left(\log_e\frac{dP^{\theta}_{Y|X}}{dQ_Y}(x,y)\right)dQ_Y$$
$$= \int\frac{\partial}{\partial\theta}\left(\frac{dP^{\theta}_{Y|X}}{dQ_Y}(x,y)\right)dQ_Y$$
$$= \frac{\partial}{\partial\theta}\int\frac{dP^{\theta}_{Y|X}}{dQ_Y}(x,y)\,dQ_Y$$
$$= \frac{\partial}{\partial\theta}(1)$$
$$= 0 \quad (82)$$

where we have used the regularity condition RC1.

To prove the first result, we could follow exactly the same approach, but it would require the regularity condition

$$\frac{\partial}{\partial\theta}\mathbb{E}_{Q_X}\left[f^{\theta}_{X|Y}(X,y)\right] = \mathbb{E}_{Q_X}\left[\frac{\partial}{\partial\theta}f^{\theta}_{X|Y}(X,y)\right] \quad (83)$$

which is in terms of the posterior $f^{\theta}_{X|Y}$ and hence difficult to verify. Next, we provide an alternative proof that relies on a regularity condition on $f^{\theta}_{Y|X}$ instead.

Using $dP^{\theta}_{X|Y}/dP_X = dP^{\theta}_{Y|X}/dP^{\theta}_Y$ and $dP^{\theta}_{X|Y}/dQ_X = dP^{\theta}_{Y|X}/dP^{\theta}_Y \times dP_X/dQ_X$

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log_e\frac{dP^{\theta}_{X|Y}}{dQ_X}(X,y) \mid Y=y\right]$$
$$= \mathbb{E}\left[\frac{\partial}{\partial\theta}\log_e\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(X,y) \mid Y=y\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\log_e\frac{dP_X}{dQ_X}(X,y) \mid Y=y\right]$$
$$= \int\frac{dP^{\theta}_{X|Y}}{dP_X}(x,y)\frac{\partial}{\partial\theta}\left(\log_e\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\right)dP_X + \mathbb{E}\left[0 \mid Y=y\right]$$
$$= \int\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\frac{\partial}{\partial\theta}\left(\log_e\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\right)dP_X$$
$$= \int\frac{\partial}{\partial\theta}\left(\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\right)dP_X$$
$$= \frac{\partial}{\partial\theta}\int\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\,dP_X$$
$$= \frac{\partial}{\partial\theta}\int\frac{dP^{\theta}_{X|Y}}{dP_X}(x,y)\,dP_X$$
$$= \frac{\partial}{\partial\theta}(1)$$
$$= 0 \quad (84)$$

where we have used the regularity condition

$$\frac{\partial}{\partial\theta}\int\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\,dP_X = \int\frac{\partial}{\partial\theta}\left(\frac{dP^{\theta}_{Y|X}}{dP^{\theta}_Y}(x,y)\right)dP_X \quad (85)$$

which follows from RC2. □

APPENDIX C
PROOF OF THEOREM 5 (BSC)

Invoke Theorem 3 with $\theta = \delta$ and particularize the result as shown in (86) at the top of the page (using $P^{\delta}_{Y^n|X_i} = P^{\delta}_{Y^{n\setminus i}|X_i}P^{\delta}_{Y_i|X_i}$ and (28)).

The term inside the expectation can be rewritten in terms of the log-likelihood ratios (as defined in (33)–(34)) for $x_i = 0$ as

$$\sum_{y_i}\frac{dP^{\delta}_{Y|X}(y_i\,|\,0)}{d\delta}\log\left(1+\frac{P^{\delta}_{X_i|Y^{n\setminus i}}(1\,|\,y^{n\setminus i})P^{\delta}_{Y|X}(y_i\,|\,1)}{P^{\delta}_{X_i|Y^{n\setminus i}}\left(0 \mid y^{n\setminus i}\right)P^{\delta}_{Y|X}(y_i \mid 0)}\right)$$
$$= \log\left(\frac{1+\exp(-\lambda_i)\exp(\gamma)}{1+\exp(-\lambda_i)\exp(-\gamma)}\right) \quad (87)$$

and for $x_i = 1$ as

$$\sum_{y_i}\frac{dP^{\delta}_{Y|X}(y_i\,|\,1)}{d\delta}\log\left(1+\frac{P^{\delta}_{X_i|Y^{n\setminus i}}\left(0 \mid y^{n\setminus i}\right)P^{\delta}_{Y|X}(y_i \mid 0)}{P^{\delta}_{X_i|Y^{n\setminus i}}\left(1 \mid y^{n\setminus i}\right)P^{\delta}_{Y|X}(y_i \mid 1)}\right)$$
$$= \log\left(\frac{1+\exp(\lambda_i)\exp(\gamma)}{1+\exp(\lambda_i)\exp(-\gamma)}\right) \quad (88)$$

where we have used

$$\frac{d}{d\delta}P^{\delta}_{Y|X}(0 \mid 0) = \frac{d}{d\delta}P^{\delta}_{Y|X}(1 \mid 1) = -1 \quad (89)$$
$$\frac{d}{d\delta}P^{\delta}_{Y|X}(1 \mid 0) = \frac{d}{d\delta}P^{\delta}_{Y|X}(0 \mid 1) = 1. \quad (90)$$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{d\log_e P^{\delta}_{Y|X}(Y_i \mid X_i)}{d\delta}\log P^{\delta}_{X_i|Y^n}(X_i \mid Y^n)\right]\\
&\quad=\mathbb{E}_{Y^{n\setminus i}}\left[P^{\delta}_{X_i|Y^{n\setminus i}}\left(0 \mid Y^{n\setminus i}\right)\mathbb{E}_{Y_i|X_i=0}\left[\frac{d\log_e P^{\delta}_{Y|X}(Y_i \mid 0)}{d\delta}\log P^{\delta}_{X_i|Y^n}(0 \mid Y^n)\right]\right.\\
&\qquad\left.+\,P^{\delta}_{X_i|Y^{n\setminus i}}\left(1 \mid Y^{n\setminus i}\right)\mathbb{E}_{Y_i|X_i=1}\left[\frac{d\log_e P^{\delta}_{Y|X}(Y_i \mid 1)}{d\delta}\log P^{\delta}_{X_i|Y^n}(1 \mid Y^n)\right]\right]\\
&\quad=\mathbb{E}_{Y^{n\setminus i}}\left[P^{\delta}_{X_i|Y^{n\setminus i}}\left(0 \mid Y^{n\setminus i}\right)\log\left(\frac{1+\exp(-\lambda_i)\exp(-\gamma)}{1+\exp(-\lambda_i)\exp(\gamma)}\right)\right.\\
&\qquad\left.+\,P^{\delta}_{X_i|Y^{n\setminus i}}\left(1 \mid Y^{n\setminus i}\right)\log\left(\frac{1+\exp(\lambda_i)\exp(-\gamma)}{1+\exp(\lambda_i)\exp(\gamma)}\right)\right]\\
&\quad=\mathbb{E}_{Y^{n\setminus i}}\left[\left(P^{\delta}_{X_i|Y^{n\setminus i}}\left(0 \mid Y^{n\setminus i}\right)-P^{\delta}_{X_i|Y^{n\setminus i}}\left(1 \mid Y^{n\setminus i}\right)\right)\log\left(\frac{\exp(\lambda_i)+\exp(-\gamma)}{\exp(\lambda_i)+\exp(\gamma)}\right)\right]\\
&\qquad-2\gamma\mathbb{E}_{Y^{n\setminus i}}\left[P^{\delta}_{X_i|Y^{n\setminus i}}\left(1 \mid Y^{n\setminus i}\right)\right]\\
&\quad=\mathbb{E}_{Y^{n\setminus i}}\left[\tanh\left(\frac{\lambda_i}{2\log e}\right)\log\left(\frac{\exp(\lambda_i)+\exp(-\gamma)}{\exp(\lambda_i)+\exp(\gamma)}\right)\right]-2\gamma P_{X_i}(1)
\end{aligned}
\tag{91}
$$

Therefore, we get (91) at the top of the following page, where we have used the relation

$$
1-2p=\tanh\left(\frac{1}{2}\log_e\frac{1-p}{p}\right)=\tanh\left(\frac{1}{2\log e}\log\frac{1-p}{p}\right). \tag{92}
$$

□

Appendix D
Proof of Theorem 7 (DMC)

First, observe that

$$
\begin{aligned}
&\left[\nabla_{\mathbf{\Pi}}P_{Y|X}(y_i \mid x_i)\right]_{kl}\\
&\quad=\frac{\partial}{\partial\pi_{kl}}P_{Y|X}(y_i \mid x_i)=1[y_i=b_k]\,1[x_i=a_l].
\end{aligned}
\tag{93}
$$

Then, invoke Theorem 3 with $\theta=\pi_{kl}$ and use $P_{Y^n|X_i}=P_{Y^{n\setminus i}|X_i}P_{Y_i|X_i}$ and (28) to particularize the result as shown in (94) at the bottom of the page. The desired result follows by noting that

$$
P_{X_i|Y^{n\setminus i}}\left(a_l \mid Y^{n\setminus i}\right)=\frac{1}{\left(1+\sum_{m\neq l}\exp\left(\lambda_i^{(m,l)}\right)\right)}. \quad \square
$$

Appendix E
Proof of Theorem 8 (Additive-Noise Channel)

Theorem 2 gives

$$
\frac{\partial}{\partial\theta}I(X^n;Y^n)=\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial s^{\theta}(X_i)}{\partial\theta}\frac{\partial\log f^{\theta}_{X_i|Y^n}(X_i \mid Y^n)}{\partial y_i}\right]. \tag{95}
$$

Now, using $P_{X_i|Y^n}=P_{Y_i|X_i}P_{Y^{n\setminus i}|X_i}P_{X_i}/P_{Y^n}$, we have

$$
\frac{\partial\log f_{X_i|Y^n}}{\partial y_i}=\frac{\partial\log P_{Y_i|X_i}}{\partial y_i}-\frac{\partial\log P_{Y^n}}{\partial y_i}. \tag{96}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\partial\log_e P_{Y|X}(Y_i \mid X_i)}{\partial\pi_{kl}}\log P_{X_i|Y^n}(X_i \mid Y^n)\right]\\
&\quad=\sum_{x_i,y^{n\setminus i}}P_{X_i}(x_i)\,P_{Y^{n\setminus i}|X_i}\left(y^{n\setminus i} \mid x_i\right)\sum_{y_i}\frac{\partial P_{Y|X}(y_i \mid x_i)}{\partial\pi_{kl}}\log\left(\frac{P_{X_i|Y^{n\setminus i}}\left(x_i \mid y^{n\setminus i}\right)P_{Y|X}(y_i \mid x_i)}{\sum_{\tilde{x}_i}P_{X_i|Y^{n\setminus i}}\left(\tilde{x}_i \mid y^{n\setminus i}\right)P_{Y|X}(y_i \mid \tilde{x}_i)}\right)\\
&\quad=-P_{X_i}(a_l)\sum_{y^{n\setminus i}}P_{Y^{n\setminus i}|X_i}\left(y^{n\setminus i} \mid a_l\right)\log\left(\frac{\sum_m P_{X_i|Y^{n\setminus i}}\left(a_m \mid y^{n\setminus i}\right)\pi_{km}}{P_{X_i|Y^{n\setminus i}}\left(a_l \mid y^{n\setminus i}\right)\pi_{kl}}\right)\\
&\quad=-\sum_{y^{n\setminus i}}P_{Y^{n\setminus i}}\left(y^{n\setminus i}\right)P_{X_i|Y^{n\setminus i}}\left(a_l \mid y^{n\setminus i}\right)\log\left(1+\sum_{m\neq l}\exp\left(\lambda_i^{(m,l)}\right)\frac{\pi_{km}}{\pi_{kl}}\right)\\
&\quad=-\mathbb{E}\left[P_{X_i|Y^{n\setminus i}}\left(a_l \mid Y^{n\setminus i}\right)\log\left(1+\sum_{m\neq l}\exp\left(\lambda_i^{(m,l)}\left(Y^{n\setminus i}\right)\right)\frac{\pi_{km}}{\pi_{kl}}\right)\right].
\end{aligned}
\tag{94}
$$

The contribution of the first term inside the expectation is zero

$$\mathbb{E}\left[\frac{\partial}{\partial y_i}\log_e P^{\theta}_{Y_i|X_i}(Y_i|x_i)\,|\,X_i=x_i\right]=\frac{\partial}{\partial y_i}\int P^{\theta}_{Y_i|X_i}(y_i|x_i)\,dy_i$$
$$=\frac{\partial}{\partial y_i}(1)$$
$$=0 \tag{97}$$

and the second term can be rewritten as

$$\begin{aligned}
&\frac{\partial \log_e P_{Y^n}(y^n)}{\partial y_i}\\
&=\frac{1}{P_{Y^n}(y^n)}\frac{\partial}{\partial y_i}\mathbb{E}\left[P^{\theta}_{Y^n|X_i}(y^n\mid X_i)\right]\\
&=\frac{1}{P_{Y^n}(y^n)}\mathbb{E}\left[\frac{\partial P^{\theta}_{Y^n|X_i}(y^n\mid X_i)}{\partial y_i}\right]\\
&=\frac{1}{P_{Y^n}(y^n)}\mathbb{E}\left[P^{\theta}_{Y^{n\backslash i}|X_i}\left(y^{n\backslash i}\mid X_i\right)\frac{\partial P^{\theta}_{Y_i|X_i}(y_i\mid X_i)}{\partial y_i}\right]\\
&=\frac{1}{P_{Y^n}(y^n)}\mathbb{E}\left[P^{\theta}_{Y^n|X_i}(y^n\mid X_i)\frac{\partial \log_e P^{\theta}_{Y_i|X_i}(y_i\mid X_i)}{\partial y_i}\right]\\
&=\mathbb{E}\left[\frac{\partial \log_e P^{\theta}_{Y_i|X_i}(y_i\mid X_i)}{\partial y_i}\mid Y^n=y^n\right]\\
&=\mathbb{E}\left[\frac{\partial \log_e P_N(N_i)}{\partial n_i}\mid Y^n=y^n\right].
\end{aligned}\tag{98}$$

Finally, we can write

$$\begin{aligned}
&\frac{\partial}{\partial\theta}I(X^n;Y^n)\\
&=-(\log e)\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial s^{\theta}(X_i)}{\partial\theta}\mathbb{E}\left[\frac{\partial \log_e P_N(N_i)}{\partial n_i}\mid Y^n\right]\right]\\
&=-(\log e)\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial s^{\theta}(X_i)}{\partial\theta}\mid Y^n\right]\mathbb{E}\left[\frac{\partial \log_e P_N(N_i)}{\partial n_i}\mid Y^n\right]\right].
\end{aligned}\tag{99}$$

□

Appendix F
Proof of Theorem 9 (Poisson Channel)

To start with we obtain an intermediate result that will prove very useful.

Lemma 3: Let X^n be a random variable with distribution P_{X^n} and $P_{Y^n|X^n}$ a time-invariant memoryless Poisson channel with $P_{Y|X}$ given by (61). Then

$$\mathbb{E}[\alpha X_i+\lambda\mid Y^n=y^n]=(y_i+1)\frac{P_{Y^n}\left(y_i+1,y^{n\backslash i}\right)}{P_{Y^n}(y^n)} \tag{100}$$

$$\mathbb{E}\left[\frac{1}{\alpha X_i+\lambda}\mid Y^n=y^n\right]=\frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\backslash i}\right)}{P_{Y^n}(y^n)} \tag{101}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\frac{X_i}{\alpha X_i+\lambda}\mid Y^n=y^n\right]\\
&\quad=\frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\backslash i}\right)}{P_{Y^n}(y^n)}\mathbb{E}\left[X_i\mid Y^n=\left(y_i-1,y^{n\backslash i}\right)\right].
\end{aligned}\tag{102}$$

Also observe that the mean and variance for the Poisson channel are given by

$$\mathbb{E}[Y_i\mid X_i]=\alpha X_i+\lambda \tag{103}$$

$$\mathbb{E}\left[Y_i^2\mid X_i\right]-\mathbb{E}^2[Y_i\mid X_i]=\alpha X_i+\lambda. \tag{104}$$

Now, using $P_{X_i|Y^n}=P_{Y^{n\backslash i}|X_i}P_{Y_i|X_i}P_{X_i}/P_{Y^n}$, we can write the expectation in (64) as

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\frac{dP_{X_i|Y^n}}{dQ_{X_i}}(X_i\mid Y^n)\right]\\
&=\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y^{n\backslash i}|X_i}\left(Y^{n\backslash i}\mid X_i\right)\right]\\
&\quad+\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y|X}(Y_i\mid X_i)\right]\\
&\quad+\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\frac{dP_{X_i}}{dQ_{X_i}}(X_i)\right]\\
&\quad-\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y^n}(Y^n)\right]
\end{aligned}\tag{105}$$

where the first and third terms vanish because of (103). The second term is

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y|X}(Y_i\mid X_i)\right]\\
&=\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\left(\frac{1}{Y_i!}(\alpha X_i+\lambda)^{Y_i}e^{-(\alpha X_i+\lambda)}\right)\right]\\
&=\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\frac{1}{Y_i!}\right]+\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)Y_i\log(\alpha X_i+\lambda)\right]\\
&\quad-\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)(\alpha X_i+\lambda)\right]
\end{aligned}\tag{106}$$

where the second term becomes $\mathbb{E}[\log(\alpha X_i+\lambda)]$ from (103)-(104) and the last term vanishes. Putting everything together, we finally have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\frac{dP_{X_i|Y^n}}{dQ_{X_i}}(X_i\mid Y^n)\right]\\
&=\mathbb{E}[\log(\alpha X_i+\lambda)]-\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]
\end{aligned}\tag{107}$$

where the last term becomes (invoking Lemma 3)

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&=\mathbb{E}\left[\left(\mathbb{E}\left[\frac{1}{\alpha X_i+\lambda}\mid Y^n\right]Y_i-1\right)\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&=\mathbb{E}\left[\left(\frac{P_{Y^n}\left(Y_i-1,Y^{n\backslash i}\right)}{P_{Y^n}(Y^n)}-1\right)\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&=\mathbb{E}\left[\frac{P_{Y^n}\left(Y_i-1,Y^{n\backslash i}\right)}{P_{Y^n}(Y^n)}\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&\quad-\mathbb{E}\left[\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&=\int P_{Y^n}\left(y_i-1,y^{n\backslash i}\right)\log\left(y_i!\,P_{Y^n}(y^n)\right)dy^n\\
&\quad-\mathbb{E}\left[\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]\\
&=\int P_{Y^n}\left(y_i,y^{n\backslash i}\right)\log\left((y_i+1)!\,P_{Y^n}\left(y_i+1,y^{n\backslash i}\right)\right)dy^n\\
&\quad-\mathbb{E}\left[\log\left(Y_i!\,P_{Y^n}(Y^n)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\mathbb{E}\left[\left(\mathbb{E}\left[\frac{X_i}{\alpha X_i+\lambda}\mid Y^n\right]Y_i-\mathbb{E}\left[X_i\mid Y^n\right]\right)\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\mathbb{E}\left[\left(\frac{P_{Y^n}\left(Y_i-1,Y^{n\setminus i}\right)}{P_{Y^n}\left(Y^n\right)}\mathbb{E}\left[X_i\mid Y_i-1,Y^{n\setminus i}\right]-\mathbb{E}\left[X_i\mid Y^n\right]\right)\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\mathbb{E}\left[\frac{P_{Y^n}\left(Y_i-1,Y^{n\setminus i}\right)}{P_{Y^n}\left(Y^n\right)}\mathbb{E}\left[X_i\mid Y_i-1,Y^{n\setminus i}\right]\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]-\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\int P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)\mathbb{E}\left[X_i\mid Y^n=\left(y_i-1,y^{n\setminus i}\right)\right]\log\left(y_i!\,P_{Y^n}\left(y^n\right)\right)dy^n\\
&\qquad-\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\int P_{Y^n}\left(y_i,y^{n\setminus i}\right)\mathbb{E}\left[X_i\mid Y^n=\left(y_i,y^{n\setminus i}\right)\right]\log\left((y_i+1)!\,P_{Y^n}\left(y_i+1,y^{n\setminus i}\right)\right)dy^n\\
&\qquad-\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left((Y_i+1)!\,P_{Y^n}\left(Y_i+1,Y^{n\setminus i}\right)\right)\right]-\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\left((Y_i+1)\frac{P_{Y^n}\left(Y_i+1,Y^{n\setminus i}\right)}{P_{Y^n}\left(Y^n\right)}\right)\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[X_i\mid Y^n\right]\log\mathbb{E}\left[\alpha X_i+\lambda\mid Y^n\right]\right].
\end{aligned}
\tag{110}
$$

□

$$
\begin{aligned}
&=\mathbb{E}\left[\log\left((Y_i+1)!\,P_{Y^n}\left(Y_i+1,Y^{n\setminus i}\right)\right)\right]\\
&\quad-\mathbb{E}\left[\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]\\
&=\mathbb{E}\left[\log\left((Y_i+1)\frac{P_{Y^n}\left(Y_i+1,Y^{n\setminus i}\right)}{P_{Y^n}\left(Y^n\right)}\right)\right]\\
&=\mathbb{E}\left[\log\mathbb{E}\left[\alpha X_i+\lambda\mid Y^n\right]\right].
\end{aligned}
\tag{108}
$$

Proceeding in the same with the derivative with respect to α, we can write the expectation in (65) as

$$
\begin{aligned}
&\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\frac{dP_{X_i|Y^n}}{dQ_{X_i}}\left(X_i\mid Y^n\right)\right]\\
&\quad=\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y|X}\left(Y_i\mid X_i\right)\right]\\
&\qquad-\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log P_{Y^n}\left(Y^n\right)\right]\\
&\quad=\mathbb{E}\left[X_i\log\left(\alpha X_i+\lambda\right)\right]\\
&\qquad-\mathbb{E}\left[X_i\left(\frac{Y_i}{\alpha X_i+\lambda}-1\right)\log\left(Y_i!\,P_{Y^n}\left(Y^n\right)\right)\right]
\end{aligned}
\tag{109}
$$

where the last term is (invoking Lemma 3) as in (110) at the top of the page.

Proof of Lemma 3: First note from (61) that

$$
\frac{1}{\alpha x_i+\lambda}=\frac{1}{y_i}\frac{P_{Y|X}\left(y_i-1\mid x_i\right)}{P_{Y|X}\left(y_i\mid x_i\right)} \tag{111}
$$

$$
\alpha x_i+\lambda=\left(y_i+1\right)\frac{P_{Y|X}\left(y_i+1\mid x_i\right)}{P_{Y|X}\left(y_i\mid x_i\right)}. \tag{112}
$$

Then, using $P_{Y^n|X_i}=P_{Y^{n\setminus i}|X_i}P_{Y|X}$,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{\alpha X_i+\lambda}\mid Y^n=y^n\right]\\
&\quad=\frac{1}{y_i}\mathbb{E}\left[\frac{P_{Y|X}\left(y_i-1\mid X_i\right)}{P_{Y|X}\left(y_i\mid X_i\right)}\mid Y^n=y^n\right]\\
&\quad=\frac{1}{y_i}\int\frac{P_{Y|X}\left(y_i-1\mid x_i\right)}{P_{Y|X}\left(y_i\mid x_i\right)}dP_{X_i|Y^n=y^n}\\
&\quad=\frac{1}{y_iP_{Y^n}(y^n)}\int P_{Y^{n\setminus i}|X_i}\left(y^{n\setminus i}\mid x_i\right)P_{Y|X}\left(y_i-1\mid x_i\right)dP_{X_i}\\
&\quad=\frac{1}{y_iP_{Y^n}\left(y^n\right)}\int P_{Y^n|X_i}\left(y_i-1,y^{n\setminus i}\mid x_i\right)dP_{X_i}\\
&\quad=\frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)}{P_{Y^n}\left(y^n\right)},
\end{aligned}
\tag{113}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\alpha X_i+\lambda\mid Y^n=y^n\right]\\
&\quad=\left(y_i+1\right)\mathbb{E}\left[\frac{P_{Y|X}\left(y_i+1\mid X_i\right)}{P_{Y|X}\left(y_i\mid X_i\right)}\mid Y^n=y^n\right]\\
&\quad=\left(y_i+1\right)\int\frac{P_{Y|X}\left(y_i+1\mid x_i\right)}{P_{Y|X}\left(y_i\mid x_i\right)}dP_{X_i|Y^n=y^n}\\
&\quad=\left(y_i+1\right)\frac{1}{P_{Y^n}\left(y^n\right)}\int P_{Y^n|X_i}\left(y_i+1,y^{n\setminus i}\mid x_i\right)dP_{X_i}\\
&\quad=\left(y_i+1\right)\frac{P_{Y^n}\left(y_i+1,y^{n\setminus i}\right)}{P_{Y^n}\left(y^n\right)},
\end{aligned}
\tag{114}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\frac{X_i}{\alpha X_i+\lambda}\mid Y^n=y^n\right]\\
&\quad=\frac{1}{y_i}\mathbb{E}\left[\frac{P_{Y|X}\left(y_i-1\mid X_i\right)}{P_{Y|X}\left(y_i\mid X_i\right)}X_i\mid Y^n=y^n\right]\\
&\quad=\frac{1}{y_i}\int\frac{P_{Y|X}\left(y_i-1\mid x_i\right)}{P_{Y|X}\left(y_i\mid x_i\right)}x_idP_{X_i|Y^n=y^n}
\end{aligned}
$$

$$= \frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)}{P_{Y^n}\left(y^n\right)}\int \frac{P_{Y^n|X_i}\left(y_i-1,y^{n\setminus i}\,|\,x_i\right)}{P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)}x_i dP_{X_i}$$
$$= \frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)}{P_{Y^n}\left(y^n\right)}\int x_i dP_{X_i|Y^n=\left(y_i-1,y^{n\setminus i}\right)}$$
$$= \frac{1}{y_i}\frac{P_{Y^n}\left(y_i-1,y^{n\setminus i}\right)}{P_{Y^n}\left(y^n\right)}\mathbb{E}\left[X_i\,|\,Y^n=\left(y_i-1,y^{n\setminus i}\right)\right]. \quad (115)$$

□

Appendix G
Robust Computation of a Function From its Noisy Derivative

As previously described, obtaining a function from its derivative and a reference point is straightforward: simply integrating. However, in practice the knowledge of the derivative will be noisy and one may have additional knowledge about the function (e.g., other reference points, monotonicity, convexity, etc.). Hence, it is desirable to have a robust approach to compute the function from the noisy derivative properly using all the available information.

Consider a discrete setting where the values of the function $f_1,\ldots,f_N$ and of its derivative $\dot{f}_1,\ldots,\dot{f}_{N-1}$ are collected in the vectors $\mathbf{f}$ and $\dot{\mathbf{f}}$, respectively. They are related by

$$\dot{\mathbf{f}} = \mathbf{D}\mathbf{f} \quad (116)$$

where $\mathbf{D}$ is the derivative matrix defined as a Toeplitz matrix with first row equal to $[-1,1,0,\ldots,0]$.

In the ideal case of a perfect knowledge of the derivative, the problem formulation would be

$$\begin{aligned} \underset{\mathbf{f}}{\text{find}} \quad & \mathbf{f} \\ \text{subject to} \quad & \dot{\mathbf{f}} = \mathbf{D}\mathbf{f} \\ & f_1 = \alpha_1 \end{aligned} \quad (117)$$

where α_1 is the known reference value at the origin of the function f_1. The solution to this problem is unique and given by

$$\mathbf{f} = \begin{bmatrix} 1 & & \mathbf{0} \\ \vdots & \ddots & \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} \alpha_1 \\ \dot{\mathbf{f}} \end{bmatrix} \quad (118)$$

which coincides with a discretized version of the integral of the derivative starting at the reference point.

In a real situation, the knowledge of the derivative $\dot{\mathbf{f}}$ will be noisy given by

$$\mathbf{z} = \dot{\mathbf{f}} + \boldsymbol{\delta} \quad (119)$$

where $\boldsymbol{\delta}$ is the noise or error. A robust formulation of the estimation of the function would be

$$\begin{aligned} \underset{\boldsymbol{\delta},\mathbf{f},\dot{\mathbf{f}}}{\text{minimize}} \quad & \|\boldsymbol{\delta}\| \\ \text{subject to} \quad & \mathbf{z} = \dot{\mathbf{f}} + \boldsymbol{\delta} \\ & \dot{\mathbf{f}} = \mathbf{D}\mathbf{f} \\ & f_i = \alpha_i \quad \forall i \end{aligned} \quad (120)$$

where the α_i's denote all the available reference points. This problem is convex (all the constraints are linear and the objective is a norm) and, hence, global solutions can be readily obtained [29]. In particular, for the l_2-norm (Euclidean norm) the problem is quadratic, and for the l_1-norm and l_∞-norm (Chebyshev norm) the problem is linear. It is straightforward to add additional constraints to incorporate, for example, the monotonicity of the function, $\mathbf{f}^D \leq (\geq)\,\mathbf{0}$, the concavity of the function, $\mathbf{D}^2\mathbf{f} \leq \mathbf{0}$ (where $\mathbf{D}^2$ denotes the second derivative in matrix form), or knowledge about the error $\boldsymbol{\delta}$ such as the nonnegativeness, $\boldsymbol{\delta} \geq \mathbf{0}$ (e.g., if the noisy knowledge of the derivative $\mathbf{z}$ is known to be an overestimation of the actual value $\dot{\mathbf{f}}$).

References

[1] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[2] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.

[3] A. Lozano, A. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.

[4] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2165–2166, May 2006.

[5] A. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4295–4297, Sep. 2006.

[6] M. Peleg, A. Sanderovich, and S. Shamai (Shitz), "On extrinsic information of good binary codes operating on Gaussian channels," *Europ. Trans. Telecommun.*, vol. 17, no. 6, Aug. 2006.

[7] C. Méasson, R. Urbanke, A. Montanari, and T. Richardson, "Life above threshold: From list decoding to area theorem and MSE," in *Proc. IEEE Information Theory Workshop*, San Antonio, TX, Oct. 2004.

[8] K. Bhattad and K. Narayanan, "An MSE based transfer chart to analyze iterative decoding schemes," in *Proc. 42nd Allerton Conf. Communication, Control, and Computing*, Monticello, IL, 2004.

[9] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[10] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," in *Proc. 2004 IEEE Information Theory Workshop*, San Antonio, TX, Oct. 2004, pp. 265–270.

[11] ——, "Additive non-Gaussian noise channels: Mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. Information Theory* , Adelaide, Australia, Sep. 2005, pp. 719–723.

[12] J. T. Coffey and A. B. Kiely, "The capacity of coded systems," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 113–127, Jan. 1997.

[13] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. 2001 IEEE Int. Conf. Communications (ICC 2001)*, Helsinki, Finland, Jun. 2001, pp. 2692–2695.

[14] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT 2001)*, Washington, DC, Jun. 2001, p. 283.

[15] 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 Conf. (Globecom-2001)*, San Antonio, TX, Nov. 2001, pp. 2992–2996.

[16] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

[17] J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 429–445, Mar. 1996.

[18] R. Koetter and A. Vardy, "Algebraic soft-decision decoding of Reed–Solomon codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 2809–2825, Nov. 2003.

[19] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, 2nd ed. San Francisco, CA: Kaufmann, 1988.
[20] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[21] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[22] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Reading, MA: Addison-Wesley , 1991.
[23] P. J. Smith, M. Shafi, and H. Gao, "Quick simulations: A review of importance sampling techniques in communications systems," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 4, pp. 597–613, May 1997.
[24] T. M. Apostol, *Calculus, Multi-Variable Calculus and Linear Algebra, With Applications to Differential Equations and Probability*, 2nd ed. New York: Wiley, 1969, vol. II.
[25] S. J. MacMullan and O. M. Collins, "The capacity of binary channels that use linear codes and decoders," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 197–214, Jan. 1998.
[26] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.
[27] J. Yu and S. Verdú, "Schemes for bi-directional modeling of discrete stationary sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4789–4807, Nov. 2006.
[28] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal estimation of entropy via block sorting," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1551–1561, Jul. 2004.
[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.