Mutual Information and Posterior Estimates in Channels of Exponential Family Type

Maxim Raginsky
ECE Department, Duke University
Durham, NC 27708, USA
Email: m.raginsky@duke.edu

Todd P. Coleman
ECE Department, University of Illinois
Urbana, IL 61801, USA
Email: colemant@illinois.edu

***Abstract*— Recently, there has been a lot of interest in the connections between information-theoretic and estimation-theoretic properties of various noisy channel models. For example, Guo, Shamai, and Verdú have shown that mutual information in Gaussian channels is related in a simple way to minimum mean-square error, regardless of the input distribution. In this paper, we consider the class of E-type channels, i.e., additive noise channels induced by an exponential family of distributions. We derive several differential and integral representations of the mutual information and the posterior information gain that are valid for any E-type channel regardless of input distribution. Next, we establish an extremal property of E-type channels that connects the Bayesian concept of a posterior estimate with a natural rate-distortion problem and makes precise a qualitative observation made by Mitter and Newton concerning information-theoretic properties of optimal nonlinear filters. Finally, we indicate how our results may be used to show monotonicity of the mutual information in E-type channels as a function of a "channel quality" parameter without assuming stochastic degradation.**

I. Introduction

Guo, Shamai, and Verdú [1] have recently shown that mutual information in Gaussian channels is related to the minimum-mean squared error (MMSE), regardless of the input distribution. A natural question to ponder is whether there exists a more general family of parametrized channel laws for which similar connections between information-theoretic and estimation-theoretic quantities may be still be established.

Recent work by Palomar and Verdú [2] has shed light on the general relationship between the mutual information and the posterior distribution of the channel input given the output. Within the Bayesian framework, this posterior distribution is naturally viewed as a (stochastic) input estimator (see, e.g., [3]). The results of [2] can be particularized to a variety of channel laws; for instance, the I-MMSE formula of [1] is a special case when the channel is Gaussian. In this paper, we consider channel laws that are related to the class of exponential families, in a manner we make precise later. We derive a number of results that express the mutual information in these *E-type channels* in terms of posterior inputs estimates and hold regardless of the input statistics. We also establish an extremal property of these channels by connecting the Bayesian framework of posterior estimation to a natural rate-distortion problem. This property makes precise a qualitative observation made by Mitter and Newton [3], [4] in the context of nonlinear filtering, that the posterior input estimate is a lossy information processor operating on the noisy output. Our results are simple consequences of the convex geometry of exponential families. We believe that, even though some of our results can be derived from the general theorems of [2], this close relationship between estimation-theoretic and information-theoretic properties of exponential families (and E-type channels) is interesting in its own right.

We also indicate how the results of this paper may be used to establish monotonicity of the mutual information in E-type channels as a function of a "channel quality" parameter without assuming stochastic degradation. For instance, monotonicity of the mutual information in Gaussian channels as a function of the SNR is a consequence of the infinite divisibility of the Gaussian distribution and the subsequent degradedness condition. Since the I-MMSE formula of [1] is a consequence of this infinite divisibility property, the desired monotonicity result may also be derived from the I-MMSE framework. The benefits of a more general approach to deriving monotonicity are numerous and extend beyond point-to-point systems. For example, consider a (not necessarily stochastically degraded) broadcast channel, where the two individual channels are different parametrizations of the same E-type transition law. If a general statement about monotonicity of mutual information in the appropriate parameter is established *regardless* of the input distribution, then such a broadcast channel is of the "more capable" type, and so its capacity region admits a complete characterization [5]. Similarly, secrecy capacity [6] in such a scenario would subsequently be completely characterized [7].

II. E-type channels

This paper deals with additive noise channels where the law (distribution) of the noise is a member of an exponential family (see, e.g., [8, Sec. 2.7]). Let $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu)$ be a measure space, where μ is a fixed σ-finite *reference measure*. Given a function[1] $\rho : \mathcal{X} \to \mathbb{R}^+$, let us define the set

$$\mathcal{B} \triangleq \left\{ \beta \geq 0 : Z(\beta) \triangleq \int_{\mathcal{X}} e^{-\beta\rho(x)} \mu(dx) < +\infty \right\}. \quad (1)$$

In what follows, we will assume that $\mathcal{B} = (0, \infty)$. We also assume that elements of $\mathcal{X}$ can be added and subtracted, the

[1] From now on, all functions are assumed to be measurable w.r.t. appropriate σ-algebras.

978-1-4244-4983-5/09/$25.00 © 2009 IEEE

two prototypical examples being $\mathcal{X} = \mathbb{R}$ with the usual real-number arithmetic and $\mathcal{X} = \{0, 1, \ldots, M-1\}$ with modulo-M arithmetic. Finally, we require that

$$\int_{\mathcal{X}} e^{-\beta\rho(y-x)}\mu(dx) = Z(\beta), \qquad \forall y \in \mathcal{X}. \tag{2}$$

For example, this will hold when μ is *translation-invariant*, i.e., for any $A \in \mathcal{F}_{\mathcal{X}}$ and $x \in \mathcal{X}$, $\mu(A - x) = \mu(A)$, where $A - x \triangleq \{x' - x : x' \in A\}$ is the translation of A by x.

We say that a channel with input $X \in \mathcal{X}$ and output $Y \in \mathcal{X}$ is of *exponential family type* (or *E-type*, for short) if its transition kernel $P_{Y|X}(A|x)$, $A \in \mathcal{F}_{\mathcal{X}}, x \in \mathcal{X}$, is absolutely continuous w.r.t. μ, and the corresponding Radon–Nikodym (RN) derivative $dP^{\beta}_{Y|X}/d\mu$ has the form

$$p^{\beta}_{Y|X}(y|x) \triangleq \frac{e^{-\beta\rho(y-x)}}{Z(\beta)}$$

for some $\beta > 0$. We will superscript the transition kernel of such a channel by β, as in $P^{\beta}_{Y|X}$, and denote by $\mathcal{E}(\rho) = \{P^{\beta}_{Y|X}\}_{\beta>0}$ the entire family of these channels. We observe that any E-type channel $P^{\beta}_{Y|X}$ is an additive noise channel that effects the random transformation $Y = X + W$, where the additive noise W is independent of the input random variable X and has the law (distribution) $P^{\beta}_W \ll \mu$ with $(dP^{\beta}_W/d\mu)(w) = e^{-\beta\rho(w)}/Z(\beta)$. In other words, the law of W belongs to the *exponential family* induced by ρ [8, Sec. 2.7].

Our interest lies with the mutual information between an input random variable X with a fixed but arbitrary distribution P_X and an output random variable Y related to X via some $P^{\beta}_{Y|X} \in \mathcal{E}(\rho)$. We will denote by P^{β}_{XY} the joint law of X and Y, by P^{β}_Y the marginal law of Y, and by $\mathbb{E}_{\beta}\{\cdot\}$ the expectation w.r.t. P^{β}_{XY}. Then the mutual information is defined through

$$\begin{aligned} I(X;Y) &\triangleq D\Big(P^{\beta}_{XY}\Big\|P_X \times P^{\beta}_Y\Big) \\ &= \begin{cases} \mathbb{E}_{\beta}\left\{\log \frac{dP^{\beta}_{XY}}{d(P_X \times P^{\beta}_Y)}\right\}, & \text{if } P^{\beta}_{XY} \ll P_X \times P^{\beta}_Y \\ +\infty, & \text{otherwise} \end{cases}. \end{aligned}$$

We note for future reference that the *information density* $i_{\beta} = dP^{\beta}_{XY}/d(P_X \times P^{\beta}_Y)$ exists and has the form

$$i_{\beta}(x, y) = \frac{e^{-\beta\rho(y-x)}}{Z(\beta|y)}, \tag{3}$$

where we have defined

$$Z(\beta|y) \triangleq \int_{\mathcal{X}} e^{-\beta\rho(y-x)}P_X(dx). \tag{4}$$

Also, the marginal distribution $P^{\beta}_Y \ll \mu$, and has the density

$$p^{\beta}_Y(y) = \frac{dP^{\beta}_Y}{d\mu}(y) = \frac{Z(\beta|y)}{Z(\beta)}. \tag{5}$$

A. Basic properties of E-type channels

Several properties of E-type channels that will be useful to us later on follow directly from the properties of exponential families. These properties are all consequences of the convex geometry of exponential families, and can be derived from the following (see, e.g., [8, Thm. 2.7.1]):

Lemma 1: The function $Z(\beta)$ is C^{∞} for $\beta > 0$, and its derivatives can be calculated by differentiating under the integral sign in (1).

Let us now define the *log-partition function*

$$\Lambda(\beta) \triangleq -\log Z(\beta) = -\log \int_{\mathcal{X}} e^{-\beta\rho(x)}\mu(dx).$$

Using Lemma 1 and the shift invariance property (2), we can prove the following:

Lemma 2: For every $\beta > 0$ and every input distribution P_X, we have

$$\Lambda'(\beta) = \mathbb{E}_{\beta}\{\rho(Y - X)\}, \ \Lambda''(\beta) = -\operatorname{Var}_{\beta}\{\rho(Y - X)\},$$

where the prime denotes differentiation w.r.t. β.

Given P_X and $P^{\beta}_{Y|X}$, we can also define the corresponding *backward* (or *posterior*) channel $P^{\beta}_{X|Y}$ via the Bayes' rule:

$$P^{\beta}_{X|Y}(A|y) = \frac{\int_A e^{-\beta\rho(x,y)}P_X(dx)}{\int_{\mathcal{X}} e^{-\beta\rho(x,y)}P_X(dx)}, \qquad A \in \mathcal{F}_{\mathcal{X}}, y \in \mathcal{X}.$$

We note that $P^{\beta}_{X|Y} \ll P_X$, and the RN derivative is

$$\frac{dP^{\beta}_{X|Y}}{dP_X}(x|y) = \frac{e^{-\beta\rho(y-x)}}{Z(\beta|y)}. \tag{6}$$

If we use P_X as the reference measure, then the (regular) conditional distribution of X given $Y = y$ also belongs to an exponential family, but one that depends on y. Using this fact and defining the log-partition function

$$\Lambda(\beta|y) \triangleq -\log \int_{\mathcal{X}} e^{-\beta\rho(y-x)}P_x(dx),$$

we can obtain the following analogue of Lemma 2:

Lemma 3: For every $y \in \mathcal{X}$ and every $\beta > 0$, we have

$$\begin{aligned} \Lambda'(\beta|y) &= \mathbb{E}_{\beta}\{\rho(Y - X)|Y = y\}, \\ \Lambda''(\beta|y) &= -\operatorname{Var}_{\beta}\{\rho(Y - X)|Y = y\}. \end{aligned}$$

B. Examples

The two classic examples of E-type channels are the binary symmetric channel (BSC) with $\mathcal{X} = \{0, 1\}$, $\rho(x) = x$, and $Z(\beta) = 1 + e^{-\beta}$, and the Gaussian channel with $\mathcal{X} = \mathbb{R}$, $\rho(x) = x^2$, and $Z(\beta) = \sqrt{\pi/\beta}$.

Another example, perhaps less known, is furnished by the *exponential server timing channel* (ESTC) [9]. The ESTC models the input-output behavior of a $\cdot/M/1$ queue, where the input and the output processes correspond, respectively, to arrivals and departures of packets to and from the queue. For a fixed $T \in (0, \infty)$, define $\mathcal{X}_T$ to be the set of counting

functions on $[0,T]$, i.e., all functions $x : [0,1] \to \mathbb{Z}_+$ that are right-continuous and satisfy the initial condition $x_0 = 0$ (we use the stochastic process notation $x_t \equiv x(t)$ here). Let $\{\mathcal{F}_t : t \in [0,T]\}$ denote the filtration over $\mathcal{X}_T$ defined via $\mathcal{F}_t \stackrel{\triangle}{=} \sigma\{x_s : s \in [0,t]\}$. Following Sundaresan and Verdú [10], we take $\mathcal{F}_{\mathcal{X}} = \mathcal{F}_T$ and let μ be the measure on $(\mathcal{X}, \mathcal{F}_T)$ such that $U = (U_t)_{t\in[0,T]} \sim \mu$ is a point process having constant, unit-rate intensity w.r.t. $\{\mathcal{F}_t : t \in [0,T]\}$. Denote the input process by $X = (X_t)_{t\in[0,T]}$, the output process by $Y = (Y_t)_{t\in[0,T]}$, and define the queue state $Q_t \stackrel{\triangle}{=} X_t - Y_t$. Then the ESTC with the initially empty queue ($Q_{0-} = 0$) and service rate β can be modeled via the transition density $p^{\beta}_{Y|X} = dP^{\beta}_{Y|X}/d\mu$ of the form [10]–[12]

$$p^{\beta}_{Y|X}(y|x) = \exp\left\{\int_0^T \log\left[\beta\rho(q_t)\right] dy_t + \left[1 - \beta\rho(q_t)\right] dt\right\}$$
$$= \beta^{y_T} e^T \exp\left\{-\beta \int_0^T \rho(x_t - y_t) dt\right\}, \qquad (7)$$

where

$$\rho(q) \stackrel{\triangle}{=} \begin{cases} 0, & q = 0 \\ 1, & q > 0 \\ \infty, & q < 0 \end{cases}.$$

Although, strictly speaking, this is not an E-type transition law because of the $\beta^{y_T} e^T$ factor in (7), for a standard input rate constraint of the form $\lim_{T\to\infty} \mathbb{E}[X_T]/T = \lambda < \beta$ it follows from the stability of the queue that

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left\{\log \beta^{Y_T}\right\} = \frac{1}{T} \mathbb{E}\left\{X_T - Q_T\right\} \log \beta = \lambda \log \beta$$

for *any* distribution of the input X satisfying the above rate constraint. As a consequence, asymptotic expressions for information rates and divergence rates, as $T \to \infty$, will be of the form associated with E-type channels.

III. Differential and integral representations

We now present a number of results pertaining to the behavior of the mutual information $I(X;Y)$ and the *posterior information gain* $G(\beta|y) \stackrel{\triangle}{=} D(P^{\beta}_{X|Y=y} \| P_X)$ as functions of the channel parameter $\beta > 0$. In particular, we will derive formulas for the derivatives of $I(X;Y)$ and $G(\beta|y)$ w.r.t. β that involve the quantities $\rho(\beta|y) \stackrel{\triangle}{=} \mathbb{E}_{\beta}\{\rho(Y - X)|Y = y\}$ and $\Lambda(\beta|y)$ and hold *regardless* of the input distribution (provided the latter satisfies mild regularity conditions). We also obtain integral representations of $I(X;Y)$ and $G(\beta|y)$ that, for a fixed $\beta > 0$, involve the behavior of all the channels in $\mathcal{E}_{>\beta}(\rho) \stackrel{\triangle}{=} \{P^{\gamma}_{Y|X}\}_{\gamma>\beta}$. These differential and integral representations are generalizations to arbitrary E-type channels of the results recently obtained by Merhav, Guo and Shamai [13] for the Gaussian channel.

Let us now fix the input distribution P_X, which is assumed to be independent of β. For brevity, we will denote the mutual information between X and Y with $(X,Y) \sim P^{\beta}_{XY}$ by $I(\beta)$; as before, differentiation w.r.t. β will be denoted by the prime, viz., $I'(\beta) = dI(\beta)/d\beta$. The results below are consequences of the identity

$$G(\beta|y) = \Lambda(\beta|y) - \beta\rho(\beta|y), \qquad (8)$$

which follows from (6), and the properties of E-type channels (cf. Sec. II-A). In particular, using (8) and Lemma 2, we get

$$I(\beta) = \mathbb{E}_{\beta}\{\Lambda(\beta|Y)\} - \beta\Lambda'(\beta). \qquad (9)$$

We note parenthetically the following interesting relation between $\mathbb{E}_{\beta}\{\Lambda(\beta|Y)\}$ and the *lautum information* $L(X;Y) \stackrel{\triangle}{=} D(P_X \times P^{\beta}_Y \| P^{\beta}_{XY})$ [14]. An easy calculation shows that $L(\beta) \equiv L(X;Y) = \beta\,\mathbb{E}_{\beta}\{\rho(Y - \bar{X})\} - \mathbb{E}_{\beta}\{\Lambda(\beta|Y)\}$, where $\bar{X} \sim P_X$ is independent of (X,Y). On the other hand, from Jensen's inequality we have

$$\Lambda(\beta|Y) = -\log \mathbb{E}_{\bar{X}}\{e^{-\beta\rho(Y-\bar{X})}\} \le \beta\,\mathbb{E}_{\bar{X}}\{\rho(Y - \bar{X})\}.$$

Hence, the lautum information $L(\beta)$ is equal to the expectation, w.r.t. Y, of the *Jensen divergence* $\Delta(\beta|Y) = \beta\,\mathbb{E}_{\bar{X}}\{\rho(Y - \bar{X})\} - \Lambda(\beta|Y)$ associated with $\Lambda(\beta|Y)$. We now turn to stating and proving the main results of this section.

Theorem 1: Suppose that the input distribution P_X is such that, for all γ in some neighborhood of every $\beta > 0$,

$$\left|\frac{d}{d\gamma}\Big(\Lambda(\gamma|y) p^{\gamma}_Y(y)\Big)\right| \le f(y) \qquad (10)$$

for some nonnegative $f \in L^1(\mu)$. Then

$$I'(\beta) = -\beta\Lambda''(\beta) - \mathrm{Cov}_{\beta}\{\rho(\beta|Y), \Lambda(\beta|Y)\}, \qquad (11)$$

where the expectation in $\mathrm{Cov}_{\beta}\{\cdot,\cdot\}$ is w.r.t. P^{β}_{XY}.

Remark 1: The regularity condition (10) amounts, for a wide variety of cases, to moment conditions of the form $\mathbb{E}\{\rho^m(X)\} < +\infty$.

Proof: Begin by writing

$$\frac{d}{d\beta}\,\mathbb{E}_{\beta}\{\Lambda(\beta|Y)\} = \frac{d}{d\beta}\int \Lambda(\beta|y) p^{\beta}_Y(y)\mu(dy)$$
$$= \int \Lambda'(\beta|y) p^{\beta}_Y(y)\mu(dy) + \int \Lambda(\beta|y)\frac{d}{d\beta} p^{\beta}_Y(y)\mu(dy),$$

where the interchange of derivative and integral is justified by the dominated convergence theorem and the condition (10). From Lemma 3, the first integral above evaluates to

$$\int \rho(\beta|y) p^{\beta}_Y(y)\mu(dy) = \mathbb{E}_{\beta}\{\rho(Y - X)\} = \Lambda'(\beta).$$

Next, we use (5) together with Lemmas 2 and 3 to show that the second integral is equal to

$$\mathbb{E}_{\beta}\{\rho(Y - X)\}\,\mathbb{E}_{\beta}\{\Lambda(\beta|Y)\} - \mathbb{E}_{\beta}\{\rho(Y - X)\Lambda(\beta|Y)\}$$
$$= \mathbb{E}_{\beta}\{\rho(\beta|Y)\}\,\mathbb{E}_{\beta}\{\Lambda(\beta|Y)\} - \mathbb{E}_{\beta}\{\rho(\beta|Y)\Lambda(\beta|Y)\}$$
$$= -\,\mathrm{Cov}_{\beta}\{\rho(\beta|Y), \Lambda(\beta|Y)\}.$$

Putting everything together and using (9), we obtain (11). ∎

For the information gain $G(\beta|y)$, the following is an immediate consequence of (8) and Lemma 3:

Theorem 2:

$$G'(\beta|y) = \beta\,\mathrm{Var}_{\beta}\Big\{\rho(Y - X)\Big|Y = y\Big\}. \qquad (12)$$

We also obtain the following integral representations of $I(\beta)$ and $G(\beta|y)$:

Theorem 3: Suppose that $\lim_{\beta\to\infty} I(P_X,\beta) < +\infty$. Then the mutual information $I(\beta)$ can be expressed as an integral

$$I(\beta) = \int_\beta^\infty \frac{Z(\gamma)}{Z(\beta)}\,\mathbb{E}_\gamma\left\{\rho(Y-X)\log i_\gamma(X,Y)\right\}d\gamma. \quad (13)$$

Proof: Begin by rewriting (11) as

$$\begin{aligned} I'(\beta) &= \beta\,\mathrm{Var}_\beta\{\rho(Y-X)\} - \mathrm{Cov}_\beta\{\rho(Y-X),\Lambda(\beta|Y)\}\\ &= -\,\mathrm{Cov}_\beta\{\rho(Y-X),\log i_\beta(X,Y)\}\\ &= -\,\mathbb{E}_\beta\{\rho(Y-X)\log i_\beta(X,Y)\} + \mathbb{E}_\beta\{\rho(Y-X)\}I(\beta)\\ &= -\,\mathbb{E}_\beta\{\rho(Y-X)\log i_\beta(X,Y)\} + \Lambda'(\beta)I(\beta), \end{aligned}$$

where in the last line Lemma 2 was used. Multiplying both sides by $Z(\beta)$ and rearranging terms, we derive the following differential equation for $I(\beta)$:

$$\Big(Z(\beta)I(\beta)\Big)' = -Z(\beta)\,\mathbb{E}_\beta\{\rho(Y-X)\log i_\beta(X,Y)\}.$$

Its solution is given by $Z(\beta)I(\beta) =$

$$\lim_{\gamma\to\infty} Z(\gamma)I(\gamma) + \int_\beta^\infty Z(\gamma)\,\mathbb{E}_\gamma\{\rho(Y-X)\log i_\gamma(X,Y)\}d\gamma.$$

Since $Z(\gamma)\to 0$ as $\gamma\to\infty$ and $\lim_{\gamma\to\infty} I(\gamma) < +\infty$ by hypothesis, we obtain (13). ∎

Theorem 4:

$$G(\beta|y) = \beta\int_\beta^\infty \frac{G(\gamma|y)}{\gamma^2}d\gamma - \beta\big[\rho(\beta|y) - \rho_0(y)\big], \quad (14)$$

where $\rho_0(y) \triangleq \inf_{x\in\mathcal{X}}\rho(y-x)$.

Proof: Dividing both sides of (8) by β and rearranging, we obtain the differential equation

$$\beta^2\left(\frac{\Lambda(\beta|y)}{\beta}\right)' = -G(\beta|y),$$

which is solved by

$$\Lambda(\beta|y) = \beta\lim_{\gamma\to\infty}\frac{\Lambda(\gamma|y)}{\gamma} + \beta\int_\beta^\infty\frac{G(\gamma|y)}{\gamma^2}d\gamma.$$

The limit in the above expression can be explicitly evaluated using the Laplace principle (see, e.g., [15, Ch. 1]) as

$$\lim_{\gamma\to\infty}\frac{\Lambda(\gamma|y)}{\gamma} = -\lim_{\gamma\to\infty}\frac{1}{\gamma}\log\int e^{-\gamma\rho(y-x)}P_X(dx) = \rho_0(y).$$

Hence,

$$\Lambda(\beta|y) = \beta\rho_0(y) + \beta\int_\beta^\infty\frac{G(\gamma|y)}{\gamma^2}d\gamma. \quad (15)$$

Substituting (15) into (8), we obtain (14). ∎

We find it remarkable that the formulas proved above hold (subject to appropriate regularity conditions) for any E-type channel regardless of the input statistics. Of course, more specific results can be obtained if one makes additional assumptions on the form of ρ. For instance, if $\mathcal{X}=\mathbb{R}$ and ρ is differentiable and homogeneous of degree $k>0$, i.e., $\rho(\beta w) = \beta^k\rho(w), \forall w\in\mathcal{X}, \beta>0$, then the corresponding channel law is of the form assumed in [16] and in Theorem 8 of [2], and their results can be subsequently recovered.

IV. Mutual information and posterior estimates

We now analyze the relationship between the mutual information $I(\beta)$ in an E-type channel and the posterior $P^\beta_{X|Y}$ in the spirit of [2]. Theorems 5 and 6 are simple consequences of the results from the preceding section. Again, we emphasize that they hold for any E-type channel regardless of the input statistics.

Theorem 5: Let $p^\beta_{X|Y}(x|y)$ denote the RN derivative of the posterior $P^\beta_{X|Y}$ w.r.t. the input distribution P_X. Then

$$I'(\beta) = -\,\mathrm{Cov}_\beta\{\rho(Y-X),\log p^\beta_{X|Y}(X|Y)\}. \quad (16)$$

Proof: From the proof of Theorem 3 we have

$$I'(\beta) = -\,\mathrm{Cov}_\beta\{\rho(Y-X),\log i_\beta(X,Y)\}.$$

Then (16) follows from this and from the fact that $i_\beta(x,y) = p^\beta_{X|Y}(x|y)$. ∎

We also have the following representation of the mutual information in terms of the posterior information gain:

Theorem 6:

$$\begin{aligned} I'(\beta) = &-\mathrm{Cov}_\beta\{\rho(\beta|Y),G(\beta|Y)\}\\ &-\mathbb{E}_\beta\left\{\mathrm{Cov}_\beta\{\rho(Y-X),\log i_\beta(X,Y)|Y\}\right\}. \end{aligned} \quad (17)$$

Proof: Use the law of total covariance

$$\mathrm{Cov}\{U,V\} = \mathrm{Cov}\{\mathbb{E}\{U|V'\},\mathbb{E}\{V|V'\}\} + \mathbb{E}\left\{\mathrm{Cov}\{U,V|V'\}\right\}$$

with $U=\rho(Y-X)$, $V=-\log i_\beta(X,Y)$, $V'=Y$, together with the fact that $\mathbb{E}_\beta\{\log i_\beta(X,Y)|Y\} = G(\beta|Y)$. ∎

Let us now return to the Bayesian interpretation of the posterior $P^\beta_{X|Y}(\cdot|y)$ as a (stochastic) estimator of the channel input X given the output $Y=y$. Recently, Mitter and Newton [3] developed a variational approach to nonlinear estimation of a random variable $X\in\mathcal{X}$ from a noisy observation Y. Their approach allows, among other things, to quantify information-theoretically the effect of misspecifications of the prior distribution P_X and the likelihood $P_{Y|X}$ on the quality of estimating the input X from the observed output Y. They showed, in particular, that the posterior estimate $P_{X|Y}$ is optimal in the sense that it minimizes what they termed the "apparent information" about the unknown input contained in the observed output. In a followup work [4], Newton also made a qualitative observation that the posterior can be viewed as a lossy data encoder, where the data are given by the observation Y. One way to make these observations precise is to consider rate-distortion coding of the *output* Y w.r.t. the fidelity criterion given by the difference distortion measure $\rho(y-x)$. In particular, we are interested in stochastic estimators of X from Y that achieve a point on the rate-distortion curve for Y. To this end, we have the following

result, which is in essence a simple consequence of the general variational equations for the rate-distortion function [17]:

Theorem 7: Let X and Y be jointly distributed $\mathcal{X}$-valued random variables, such that $P_Y \ll \mu$ and $I(X;Y) < +\infty$. Then the posterior distribution $P_{X|Y}$ achieves a point on the rate-distortion curve for Y if and only if there is a version of the conditional probability law $P_{Y|X}$ that is a member of the E-type family $\mathcal{E}(\rho)$.

Proof: From the variational equations for the rate-distortion function, we see that the posterior $P_{X|Y}$ (which, owing to the assumed finiteness of $I(X;Y)$, is absolutely continuous w.r.t. P_X) achieves a point (R, D) on the rate-distortion curve for P_Y if and only if

$$\frac{dP_{X|Y}}{dP_X}(x|y) = \frac{e^{-\beta\rho(y-x)}}{\int_{\mathcal{X}} e^{-\beta\rho(y-x)} P_X(dx)} \equiv \frac{e^{-\beta\rho(y-x)}}{Z(\beta|y)},$$

where β is determined from the relation $D = \Lambda'(\beta)$. Thus, $dP_{X|Y}/dP_X$ is of the form (6), which proves the theorem. ■

V. Towards monotonicity

We close with a discussion of how the results proved in this paper can potentially be used to assess whether the mutual information $I(\beta)$ in an E-type channel is a monotonic function of the channel quality parameter β. Such a monotonicity property is an easy consequence of the data processing theorem when the channel family $\mathcal{E}(\rho) = \{P^{\beta}_{Y|X}\}$ can be *ordered by degradation*, i.e., if, for every $\beta_1 < \beta_2$ there exists a transition kernel Q, such that

$$P^{\beta_1}_{Y|X}(A|x) = \int_{\mathcal{X}\times A} P^{\beta_2}_{Y|X}(dy'|x) Q(dy|y')$$

for all $A \in \mathcal{F}_{\mathcal{X}}, x \in \mathcal{X}$. Ordering by degradation is a very strong condition; for additive noise channels, it is equivalent to the following: if $W_j \sim P^{\beta_j}_W$, $j \in \{1,2\}$, and $\beta_1 > \beta_2$, then there exists a random variable U independent of W_1, such that W_2 has the same distribution as $W_1 + U$ [18]. In statistics, the general problem of ordering noisy channels according to their "informativeness" is known under the name of *comparison of experiments*, and dates back to the seminal work of Blackwell [19] (see also [20] and references therein).

When the channel family is not degraded, it may be still possible to prove monotonicity directly by establishing that the right-hand side of (11) is nonnegative. From Theorem 1 we immediately see that $I(\beta) \geq 0$ at a given β if and only if

$$\mathrm{Cov}_\beta\{\rho(\beta|Y), \Lambda(\beta|Y)\} \leq -\beta\Lambda''(\beta).$$

Theorem 5 then tells us that this condition is equivalent to $\mathrm{Cov}_\beta\{\rho(Y-X), \log p^{\beta}_{X|Y}(X|Y)\} \leq 0$. In other words, the mutual information $I(\beta)$ is a monotonically increasing function of β if and only if, for every $\beta > 0$, the "error" $\rho(Y-X)$ and the posterior log-likelihood $\log p^{\beta}_{X|Y}(X|Y)$ are negatively correlated.

Consider channels with real-valued inputs and outputs ($\mathcal{X} = \mathbb{R}$). When the additive noise density $p^{\beta}_W(w)$ is of the form $p^{\beta}_W(w) = \beta f(\beta w)$, where $f(\cdot)$ is a fixed differentiable, unimodal density, this negative correlation [equivalently, monotonicity of $I(\beta)$] follows from the classical results of Stone on the comparison of location experiments [21, Thm. 5]. We conjecture that the same result is valid more generally in the case when the densities $p^{\beta}_W(w)$ are log-concave [i.e., when the function $w \mapsto \rho(w)$ is convex and has a unique minimum].

Acknowledgment

T.P. Coleman would like to acknowledge support provided by the AFOSR Complex Networks Program via award no FA9550-08-1-0079, the DARPA ITMANET Program via US Army RDECOM contract W911NF-07-1-0029, and the NSF CyberTrust Program via award CNS 08-31488.

References

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

[2] D. P. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 453–470, February 2007.

[3] S. K. Mitter and N. J. Newton, "A variational approach to nonlinear estimation," *SIAM J. Control Optim.*, vol. 42, no. 5, pp. 1813–1833, 2003.

[4] N. J. Newton, "Interactive statistical mechanics and nonlinear filtering," *J. Stat. Phys.*, vol. 133, pp. 711–737, 2008.

[5] A. El Gamal, "The capacity of a class of broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 2, pp. 166–169, 1979.

[6] A. Wyner, "The wire-tap channel," *Bell Sys. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, 1975.

[7] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 339–348, 1978.

[8] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

[9] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 4–18, 1996.

[10] R. Sundaresan and S. Verdú, "Capacity of queues via point-process channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2697–2709, June 2006.

[11] P. Brémaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer, 1981.

[12] A. B. Wagner and V. Anantharam, "Zero-rate reliability of the exponential-server timing channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 447–465, February 2005.

[13] N. Merhav, D. Guo, and S. Shamai (Shitz), "Statistical physics of signal estimation in Gaussian noise: theory and examples," *IEEE Trans. Inf. Theory*, 2008, submitted.

[14] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 964–975, March 2008.

[15] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley, 1997.

[16] D. Guo, S. Shamai (Shitz), and S. Verdú, "Additive non-Gaussian noise channels: mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. on Inf. Theory*, Adelaide, Australia, September 2005, pp. 719–723.

[17] I. Csiszár, "On an extremum problem in information theory," *Stud. Sci. Math. Hung.*, vol. 9, pp. 57–70, 1974.

[18] E. L. Lehmann, "Comparing location experiments," *Ann. Statist.*, 1988.

[19] D. Blackwell, "Comparison of experiments," in *Proc. 2nd Berkeley Symp. on Math. Statist. Probab.*, 1951, pp. 93–102.

[20] P. K. Goel and J. Ginebra, "When is one experiment 'always better than' another?" *The Statistician*, vol. 52, no. 4, pp. 515–537, 2003.

[21] M. Stone, "Non-equivalent comparisons of experiments and their use for experiments involving location parameters," *Ann. Statist.*, vol. 32, pp. 326–332, 1961.