Additive Non-Gaussian Noise Channels: Mutual Information and Conditional Mean Estimation

Dongning Guo
Electrical & Computer Engineering Dept.
Northwestern University
Evanston, IL 60208, USA
Email: dGuo@Northwestern.edu

Shlomo Shamai (Shitz)
Electrical Engineering Dept.
Technion-Israel Inst. of Tech.
32000 Haifa, Israel
Email: sshlomo@ee.technion.ac.il

Sergio Verdú
Electrical Engineering Dept.
Princeton University
Princeton, NJ 08544, USA
Email: Verdu@Princeton.edu

***Abstract*— It has recently been shown that the derivative of the input-output mutual information of Gaussian noise channels with respect to the signal-to-noise ratio is equal to the minimum mean-square error. This paper considers general additive noise channels where the noise may not be Gaussian distributed. It is found that, for every fixed input distribution, the derivative of the mutual information with respect to the signal strength is equal to the correlation of two conditional mean estimates associated with the input and the noise respectively. Special versions of the result are given in the respective cases of additive exponentially distributed noise, Cauchy noise, Laplace noise, and Rayleigh noise. The previous result on Gaussian noise channels is also recovered as a special case.**

I. Introduction

Consider the classical problem of information transmission through a memoryless noisy channel. Two performance measures of fundamental importance in information theory and estimation theory respectively have been well-studied. One is the mutual information between the input and output pair; the other is the minimum mean-square error (MMSE) in estimating the input given the output. It is well-known that the MMSE is achieved by conditional mean estimation. In continuous-time channels observed in white Gaussian noise, the mutual information is equal to the causal MMSE times the signal-to-noise ratio (SNR) regardless of whether the input signal is Gaussian [1]. Recently, more fundamental links which hold in both continuous-time and discrete-time (and some even more general settings) have been shown between mutual information and estimation-theoretic quantities [2]. Most notably, the mutual information is the integral of the MMSE as a function of the SNR in additive Gaussian noise channels. In Poisson channels, the mutual information can also be represented as an integral of a certain measure of the estimation error [3]. The key formula for Gaussian channels in [2] has found applications in nonlinear filtering, multiuser detection and information theory. An operational meaning of the formula leads to the re-definition of the extrinsic information transfer (EXIT) chart that illustrates the dynamics of iterative decoding of error control codes such as the low density parity check (LDPC) codes [4], [5].

This work was supported in part by 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.

A natural question to pose is how general the information-estimation relationship can be. Reference [3] notes the key property of independent increments exploited in the case of Gaussian channels and widens the ground to include Poisson channels, in which the output is a doubly stochastic point process whose rate is affine with the input. This paper takes a different viewpoint and generalizes the information-estimation relationship to additive noise channels of the form

$$Y = h(a; X) + W \tag{1}$$

where the noise W is an arbitrary continuous random variable, and $h(a; X)$ is a deterministic signaling function of the random input X and a parameter a that represents the quality of the channel. It is found that, for an arbitrary but fixed input distribution, the geometry of the probability law of the additive noise channel allows the derivative of the mutual information with respect to the channel quality measure a to be expressed as the correlation of two conditional mean estimates, one associated with the input X and the other with the noise W. The relationship is shown in simple forms for special noise distributions, including the exponential, Cauchy, Laplace, Rayleigh, as well as the Gaussian noise, where in each of these cases a represents the channel gain in the amplitude.

The remainder of this paper is organized as follows. Section II studies the general additive noise channel and presents the main results. Proofs are given in Section III. Special versions of the result in case of well-known noise distributions are shown in Section IV.

II. General Results

Consider the general memoryless additive noise channel described by (1) where X and Y are real-valued random variables that denote the input and output of the channel respectively, and W denotes the additive noise independent of X. Here $h(\cdot;\cdot)$ is a deterministic signaling function and the parameter $a \in \mathbb{R}$ describes the quality of the channel. Suppose that the derivative of $h(\cdot;\cdot)$ with respect to the first argument exists and is denoted by $h'(\cdot;\cdot)$. It is also assumed that $p_W(\cdot)$, the probability density function (pdf) of the noise, exists and is differentiable. Note that the conditional distribution that describes the channel is simply

$$p_{Y|X;a}(y|x;a) = p_W(y - h(a;x)). \tag{2}$$

The unconditional output distribution is thus

$$p_{Y;a}(y;a) = \mathsf{E}\left\{p_W(y - h(a;X))\right\}. \tag{3}$$

Throughout this paper we assume that the input distribution P_X and the noise pdf p_W are arbitrary but fixed and not dependent on a, the channel quality parameter.[1] We also assume that the following conditions are satisfied:[2]

$$\nabla p_W(w) \text{ uniformly continuous on the support of } p_W; \tag{4a}$$

$$\mathsf{E}\left\{\nabla \log p_W(W) \mid Y\right\} \text{ exists}; \tag{4b}$$

$$\mathsf{E}\left\{h'(a;X) \cdot \nabla \log p_W(W) \mid Y\right\} \text{ exists}; \tag{4c}$$

$$\text{For every } a_0, \left|\log p_{Y;a}(y;a) \cdot \frac{\partial}{\partial a} p_{Y;a}(y;a)\right| < f(y)$$
$$\text{in a neighborhood of } a_0 \text{ with some integrable } f(y). \tag{4d}$$

The technical conditions in (4) are satisfied by most input and noise distributions of interest in communication problems. As we shall see in Section III, these conditions are needed to guarantee that the derivatives with respect to y and a penetrate certain expectations. Once the noise distribution p_W is given, the conditions can be regarded as constraints on the input distribution P_X. For example, if the noise distribution is standard Gaussian, conditions (4) simply boil down to the existence of $\mathsf{E}X^2$.

The main result of this paper is the derivative of the input-output mutual information[3] of the channel (1) given by the following theorem, the proof of which is relegated to Section III.

Theorem 1: Consider the general additive noise channel (1). For every fixed input distribution and every fixed noise distribution that satisfy conditions (4),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) \\ = -\mathsf{E}\Big\{\mathsf{E}\left\{h'(a;X)|Y\right\}\mathsf{E}\left\{\nabla \log p_W(W)|Y\right\}\Big\}.\end{aligned} \tag{5}$$

Theorem 1 essentially states that the increase in the mutual information due to improvement in the channel quality is equal to minus the correlation of two conditional mean estimates, one associated with the input, and the other with the noise. It is remarkable that the relationship (5) holds regardless of the input and noise statistics. We observe that reference [5] also considers the derivative of the input-output mutual information of a memoryless channel with respect to some channel parameter, but is short of expressing the derivative in terms of conditional mean estimates in the case of additive noise.

The random variable $\nabla \log p_W(W)$ is known as the *score* of the distribution p_W [6, p. 327]. It is clear that the score has zero mean:[4]

$$\mathsf{E}\left\{\nabla \log p_W(W)\right\} = \int \nabla p_W(w)\, \mathrm{d}w = 0. \tag{6}$$

The variance of the score is the Fisher information of the noise distribution

$$I_0(p_W) = \mathsf{E}\left\{[\nabla \log p_W(W)]^2\right\}. \tag{7}$$

A simple and important scenario of the additive noise channel model (1) is the case where the parameter $a \geq 0$ serves as the signal amplitude described by

$$Y = aX + W. \tag{8}$$

Condition (4c) becomes

$$\mathsf{E}\left\{X \cdot \nabla \log p_W(W) \mid Y\right\} \text{ exists for all } a. \tag{9}$$

In view of Theorem 1, the following result is immediate.

Corollary 1: Consider the additive noise channel (8). For every fixed input distribution and every fixed noise distribution that satisfy conditions (4),

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) = -\mathsf{E}\big\{\mathsf{E}\left\{X \mid Y\right\}\mathsf{E}\left\{\nabla \log p_W(W) \mid Y\right\}\big\}. \tag{10}$$

It is interesting to note that the derivative of the mutual information is directly related to the conditional mean estimate of the input, which is known to achieve the MMSE. In the limit of vanishing channel gain, i.e., $a \to 0$, one finds that

$$\mathsf{E}\left\{X \mid Y = y\right\} = \mathsf{E}X - a\,\sigma_X^2 \nabla \log p_W(y) + o(a) \tag{11}$$

where σ_X^2 stands for the variance of the input X. Using Corollary 1, the rate of increase of the mutual information with respect to the input power gain a^2 is found at $a = 0$:

Corollary 2: Suppose that the Fisher information of the noise distribution is finite. Under the same conditions as in Corollary 1,

$$\lim_{a \to 0} \frac{\mathrm{d}}{\mathrm{d}(a^2)} I(X;Y) = \frac{\sigma_X^2}{2} I_0(P_W) \tag{12}$$

The result was also obtained in [7] and [8, p. 1025]. It is interesting to remark that the rate of increase of the mutual information at zero channel gain depends on the input distribution only through its variance. This result is the generalization to arbitrary additive noise channels of a fact that has been observed in Gaussian channels (e.g., [8], [9], [10], [2], [11]).

One application of Theorem 1 and Corollary 1 is to the computation of mutual information, which can be found as an integral of the correlation of some conditional mean estimates by (10). This suggests a viable technique to obtain or

[1] In general, P_X represents the input probability measure since the pdf may not exist.

[2] Here ∇ stands for the gradient operator. In general, for any differentiable function $f : \mathbb{R}^L \to \mathbb{R}$, its gradient at any $\boldsymbol{y} \in \mathbb{R}^L$ is a column vector $\nabla f(\boldsymbol{y}) = \left[\frac{\partial f}{\partial y_1}(\boldsymbol{y}), \cdots, \frac{\partial f}{\partial y_L}(\boldsymbol{y})\right]^\top$.

[3] The unit of information measures is assumed to be nats throughout.

[4] The integral with respect to w is from $-\infty$ to ∞. For notational simplicity we omit integral limits in this paper whenever they are clear from their context.

bound the mutual information whenever direct calculation is cumbersome. Another application are generalized EXIT charts that describe the dynamics of iterative decoding in memoryless channels (cf. [5]). Furthermore, power allocation with parallel non-Gaussian noise channels can be solved using formulas (5) and (10) (cf. [12]).

It is straightforward to generalize Theorem 1 to vector channels with vector channel parameters. Consider the vector version of the additive noise channel

$$\boldsymbol{Y} = \boldsymbol{h}(\boldsymbol{a}; X) + \boldsymbol{W} \tag{13}$$

where $\boldsymbol{W}$ is a column vector consisting of L independent noise entries with the same pdf p_W, and $\boldsymbol{a}$ is a vector that consists of parameters that describe the channel quality. Denote the pdf of $\boldsymbol{W}$ as $p_{\boldsymbol{W}}$, and the derivative with respect to $\boldsymbol{a}$ using a gradient operator $\nabla_{\boldsymbol{a}}$. We omit the proof of the following result.

Theorem 2: For every fixed input distribution and fixed noise distribution that satisfy the vector version of conditions (4),

$$\begin{aligned}\nabla_{\boldsymbol{a}} I(X; \boldsymbol{Y}) = -\mathsf{E}\big\{ \mathsf{E}\left\{ \nabla_{\boldsymbol{a}} \boldsymbol{h}(\boldsymbol{a}; X) \mid \boldsymbol{Y} \right\} \\ \times \mathsf{E}\left\{ \nabla \log p_{\boldsymbol{W}}(\boldsymbol{W}) \mid \boldsymbol{Y} \right\} \big\}.\end{aligned} \tag{14}$$

III. Proof

Theorem 1 is proved by directly taking the derivative of the mutual information with respect to the channel quality parameter and exploiting the geometry of the conditional density function (2).

Proof: [Theorem 1] We first note that conditions (4b) and (4c) are equivalent to

$$\mathsf{E}\left\{ \nabla p_W(y - h(a; X)) \right\} < \infty \quad \forall a, y \tag{15}$$

and

$$\mathsf{E}\left\{ h'(a; X) \nabla p_W(y - h(a; X)) \right\} < \infty \quad \forall a, y. \tag{16}$$

The input-output mutual information is expressed as

$$I(X; Y) = \mathsf{E}\left\{ \log \frac{p_{Y|X;a}(Y|X; a)}{p_{Y;a}(Y; a)} \right\} \tag{17}$$

$$= \mathsf{E}\left\{ \log \frac{p_W(W)}{p_{Y;a}(Y; a)} \right\}. \tag{18}$$

It is straightforward to calculate the derivative of the mutual information with respect to the parameter a:

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X; Y)$$

$$= -\frac{\mathrm{d}}{\mathrm{d}a} \mathsf{E}\left\{ \log p_{Y;a}(Y; a) \right\} \tag{19}$$

$$= -\int \log p_{Y;a}(y; a) \frac{\partial}{\partial a} p_{Y;a}(y; a) \,\mathrm{d}y \tag{20}$$

$$= -\iint \log p_{Y;a}(y; a) \frac{\partial}{\partial a} p_{Y|X;a}(y|x; a) \,\mathrm{d}y P_X(\,\mathrm{d}x). \tag{21}$$

The exchange of derivative and expectation in (20) is justified by Lebesgue's dominated convergence theorem using condition (4d). Similarly, the derivative with respect to a penetrates the integral over P_X in (21) because of the uniform continuity of $\nabla p_W(\cdot)$ and the integrability of $(\partial/\partial a) p_{Y|X;a}(y|X; a)$, which are guaranteed by conditions (4a) and (16). We note that by (2),

$$\frac{\partial}{\partial a} p_{Y|X;a}(y|x; a) = -h'(a; x) \frac{\partial}{\partial y} p_{Y|X;a}(y|x; a). \tag{22}$$

Thus the derivative (21) reduces to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a} I(X; Y) = \iint h'(a; x) \log p_{Y;a}(y; a) \\ \times \frac{\partial}{\partial y} p_{Y|X;a}(y|x; a) \,\mathrm{d}y P_X(\,\mathrm{d}x).\end{aligned} \tag{23}$$

Integrating by parts (see e.g., [13]) with respect to y, (23) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X; Y)$$

$$\begin{aligned}= -\iint h'(a; x) \frac{p_{Y|X;a}(y|x; a)}{p_{Y;a}(y; a)} \\ \times \frac{\partial}{\partial y} p_{Y;a}(y; a) \,\mathrm{d}y P_X(\,\mathrm{d}x)\end{aligned} \tag{24}$$

$$= -\int \mathsf{E}\left\{ h'(a; X) \mid Y = y \right\} \frac{\partial}{\partial y} p_{Y;a}(y; a) \,\mathrm{d}y \tag{25}$$

where we have used the following

$$\int f(a; x) \frac{p_{Y|X;a}(y|x; a)}{p_{Y;a}(y; a)} P_X(\,\mathrm{d}x)$$

$$= \int f(a; x) P_{X|Y;a}(\,\mathrm{d}x|y; a) \tag{26}$$

$$= \mathsf{E}\left\{ f(a; X) \mid Y = y \right\} \tag{27}$$

which is true for all measurable functions $f(\cdot; \cdot)$. We further note that

$$\frac{\partial}{\partial y} p_{Y;a}(y; a)$$

$$= \int \frac{\partial}{\partial y} p_{Y|X;a}(y|x; a) \, P_X(\,\mathrm{d}x) \tag{28}$$

$$\begin{aligned}= \int \frac{\partial}{\partial y} \log p_{Y|X;a}(y|x; a) \\ \times p_{Y|X;a}(y|x; a) \, P_X(\,\mathrm{d}x)\end{aligned} \tag{29}$$

$$\begin{aligned}= \int \nabla \log p_W(y - h(a; x)) \frac{p_{Y|X;a}(y|x; a)}{p_{Y;a}(y; a)} \\ \times P_X(\,\mathrm{d}x) \, p_{Y;a}(y; a)\end{aligned} \tag{30}$$

$$\begin{aligned}= \mathsf{E}\left\{ \nabla \log p_W(y - h(a; X)) \mid Y = y \right\} \\ \times p_{Y;a}(y; a)\end{aligned} \tag{31}$$

$$= \mathsf{E}\left\{ \nabla \log p_W(W) \mid Y = y \right\} \, p_{Y;a}(y; a) \tag{32}$$

where (31) follows from (27), and the exchange of derivative and expectation in (28) is justified by condition (4b) and dominated convergence. Using (32), (25) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) = -\int p_{Y;a}(y;a) \mathsf{E}\{ h'(a;X) \mid Y=y\} \times \mathsf{E}\{ \nabla \log p_W(W) \mid Y=y\} \, \mathrm{d}y. \tag{33}$$

Theorem 1 is thus proved. ■

Corollary 2 can be proved by examining the Taylor series expansion of $p_W(y-ax)$ at the vicinity of $a=0$.

Proof: [Corollary 2] We first show (11). Using (27),

$$\begin{aligned}
&\mathsf{E}\{ X \mid Y=y\} \\
&= \int x \frac{p_{Y|X;a}(y|x;a)}{p_{Y;a}(y;a)} P_X(\mathrm{d}x) && (34)\\
&= \int x \frac{p_W(y-ax)}{\mathsf{E}\{p_W(y-aX)\}} P_X(\mathrm{d}x) && (35)\\
&= \int x \frac{p_W(y) - axp_W'(y) + o(a)}{p_W(y) - a\,\mathsf{E}Xp_W'(y) + o(a)} P_X(\mathrm{d}x) && (36)\\
&= \mathsf{E}X - a\,\sigma_X^2 p_W'(y)/p_W(y) + o(a) && (37)
\end{aligned}$$

which is equivalent to (11). It is then straightforward to verify (12) using Corollary 1 by noting (6) and the following continuity result:

$$\lim_{a\to 0} \mathsf{E}\{ \nabla \log p_W(W) \mid Y=y\} = \nabla \log p_W(y). \tag{38}$$

■

IV. Special Cases

This section presents the specialization of Corollary 1 in case of several popular noise distributions.

A. Gaussian Noise

The special case of Gaussian noise has been treated in [2] where it is shown that the derivative of the input-output mutual information with respect to the SNR is equal to the MMSE (times a factor of $1/2$).[5]

Let p_W be standard Gaussian. Then

$$\nabla \log p_W(w) = -w. \tag{39}$$

Note that conditions (4b) and (4c) correspond to the existence of $\mathsf{E}X$ and $\mathsf{E}X^2$ respectively. It turns out that as long as $\mathsf{E}X < \infty$, condition (4d) is also satisfied. Plugging into Corollary 1 yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) &= \mathsf{E}\{\mathsf{E}\{ X \mid Y\} \mathsf{E}\{ W \mid Y\}\} && (40)\\
&= \mathsf{E}\{\mathsf{E}\{ X \mid Y\} \mathsf{E}\{ Y - aX \mid Y\}\} && (41)\\
&= \mathsf{E}\{XY\} - a\,\mathsf{E}\left\{(\mathsf{E}\{ X \mid Y\})^2\right\} && (42)\\
&= a\,\mathsf{E}\left\{(X - \mathsf{E}\{ X \mid Y\})^2\right\} && (43)
\end{aligned}$$

where the final expectation is the MMSE. From (43) the key formula in [2] readily follows.

[5]The factor of $\frac{1}{2}$ is absent for complex-valued channels.

B. Laplace Noise

If the noise has Laplace distribution:

$$p_W(w) = \frac{1}{2} e^{-|W|}, \tag{44}$$

then

$$\nabla \log p_W(w) = -\nabla |w| = -\mathrm{sgn}(w). \tag{45}$$

Note that

$$\mathsf{E}\{ \mathrm{sgn}(W) \mid Y\} = 2\,\mathsf{P}\{W > 0 \,|\, Y\} - 1. \tag{46}$$

By Corollary 1, for the additive Laplace noise channel, as long as $\mathsf{E}X < \infty$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) \\
&\quad = \mathsf{E}\Big\{ \mathsf{E}\{ X \mid Y\} \mathsf{E}\{ \mathrm{sgn}(W) \mid Y\} \Big\} && (47)\\
&\quad = 2\,\mathsf{E}\Big\{ \mathsf{E}\{ X \mid Y\} \mathsf{P}\{W > 0 \,|\, Y\} \Big\} - \mathsf{E}X. && (48)
\end{aligned}$$

C. Cauchy Noise

Consider the case of Cauchy distributed noise:

$$p_W(w) = \frac{1}{\pi} \frac{1}{1+w^2}. \tag{49}$$

By Corollary 1, the derivative of the mutual information is found as

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) = \mathsf{E}\left\{ \mathsf{E}\{ X \mid Y\} \mathsf{E}\left\{ \frac{2W}{W^2+1} \,\middle|\, Y \right\} \right\} \tag{50}$$

if $\mathsf{E}X < \infty$, which is immediate by noting that

$$\nabla \log p_W(w) = -\frac{2w}{w^2+1}. \tag{51}$$

D. Rayleigh Noise

Let the input signal X be a positive random variable. Consider the case of Rayleigh distributed noise:

$$p_W(w) = w\, e^{-\frac{1}{2}w^2}. \tag{52}$$

It is clear that $\nabla p_W(\cdot)$ is uniformly continuous in $[0,\infty)$ and

$$\nabla \log p_W(w) = \frac{1}{w} - w. \tag{53}$$

For the additive Rayleigh noise channel, if $\mathsf{E}X < \infty$, then

$$\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) = \mathsf{E}\left\{ \mathsf{E}\{ X \mid Y\} \mathsf{E}\left\{ \frac{1}{W} - W \,\middle|\, Y \right\} \right\}. \tag{54}$$

E. Exponential Noise

The exponential distribution shares some interesting properties with the Gaussian distribution [14]. Let the additive noise W take the exponential distribution with unit mean, i.e.,

$$p_W(w) = e^{-w} u(w) \tag{55}$$

where $u(w)$ is the unit step function. We let $u(0)=1$. The discontinuity of $p_W(w)$ at $w=0$ requires caution. Indeed,

$$\nabla p_W(w) = \delta(w) - e^{-w} u(w) \tag{56}$$

where $\delta(\cdot)$ is the Dirac function.

Condition (4a) is violated at the singular point $w = 0$. Fortunately, for our purposes, it is adequate to write

$$\nabla \log p_W(w) = \delta(w) - 1. \tag{57}$$

Corollary 1 still holds in this case. In fact, for the exponential noise channel with any positive input that satisfies $\mathsf{E}X < \infty$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) \\
&= -\mathsf{E}\{\mathsf{E}\{X \mid Y\}\,\mathsf{E}\{\delta(Y-aX)-1 \mid Y\}\} && (58)\\
&= \mathsf{E}X - \mathsf{E}\{\mathsf{E}\{X \mid Y\}\,\mathsf{E}\{\delta(Y-aX) \mid Y\}\} && (59)\\
&= \mathsf{E}X - \int \mathsf{E}\{X \mid Y=y\}\,\mathsf{E}\{\delta(Y-aX) \mid Y=y\} \\
&\qquad\qquad \times p_{Y;a}(y;a)\,\mathrm{d}y && (60)\\
&= \mathsf{E}X - \int \mathsf{E}\{X \mid Y=y\} \int \delta(y-ax) \\
&\qquad\qquad \times P_{X|Y;a}(\mathrm{d}x|y;a)\,p_{Y;a}(y;a)\,\mathrm{d}y && (61)\\
&= \mathsf{E}X - \iint \mathsf{E}\{X \mid Y=y\}\,\delta(y-ax) \\
&\qquad\qquad \times P_{X,Y;a}(\mathrm{d}x,\,\mathrm{d}y;a) && (62)\\
&= \mathsf{E}X - \iint \mathsf{E}\{X \mid Y=y\}\,\delta(y-ax) \\
&\qquad\qquad \times p_{Y|X;a}(y|x;a)\,\mathrm{d}y\,P_X(\mathrm{d}x) && (63)\\
&= \mathsf{E}X - \int \mathsf{E}\{X \mid Y=ax\}\,P_X(\mathrm{d}x) && (64)
\end{aligned}$$

where (64) is due to $p_W(0) = 1$. Thus

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}a} I(X;Y) &= \mathsf{E}X - \mathsf{E}\{\mathsf{E}\{X \mid Y=aX'\}\} && (65)\\
&= (1/a)\mathsf{E}\{\mathsf{E}\{W \mid Y=aX'\}\} && (66)
\end{aligned}$$

where X' is an independent copy of X. In (65) and (66), $\mathsf{E}\{W \mid Y = aX'\}$ is the conditional mean estimate of the noise with the observation equal to aX', which is the output of the same channel with input X' but free of noise. The right hand side (RHS) of (66) is the expectation of such an estimate with respect to X' taking the input distribution P_X.

In Fig. 1 we verify (66) with a special mixture input:

$$p_X(x) = t\,\delta(x) + (1-t)\,e^{-x}u(x). \tag{67}$$

The mutual information and the RHS of (66) are obtained through numerical integration and plotted as a function of the parameter a with $t = 1/2$. The numerical integral of the latter coincides with the former. The input distribution is capacity achieving if the channel gain is $a = 1/t$, in which case the output distribution is exponential and the conditional mean is

$$\mathsf{E}\{X \mid Y=y\} = \tfrac{1}{a}\,e^{-\frac{1-a}{a}y} \tag{68}$$

and the RHS of (66) is equal to $a/2 - 2/a + 1/a^2$, which is $1/4$ with $a = 2$.

V. Conclusion

This paper unveils an information-estimation relationship that holds in general additive noise channels. It is found that the rate of mutual information increase with respect to some channel quality measure is equal to the correlation between two conditional mean estimates associated with the input and the noise respectively. The relationship is proved for arbitrary but fixed input distribution and noise distribution that satisfy a set of mild technical conditions. The mutual information-MMSE formula for Gaussian channels can be regarded as a special case of the new result.

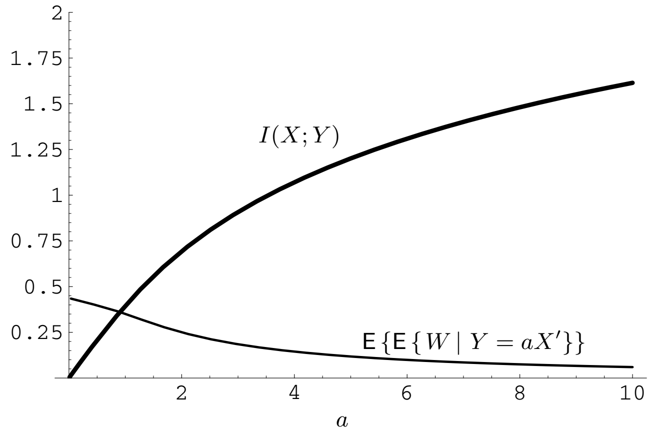

Fig. 1. Verification of Corollary 1 in the special case of exponentially distributed noise and input of a mixture of point mass and exponential distribution.

References

[1] T. E. Duncan, "On the calculation of mutual information," *SIAM Journal of Applied Mathematics*, vol. 19, pp. 215–220, July 1970.

[2] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261–1282, Apr. 2005.

[3] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," in *Proc. IEEE Inform. Theory Workshop*, pp. 265–270, San Antonio, TX, USA, 2004.

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

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

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[7] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[8] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.

[9] A. Lapidoth and S. Shamai, "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1118–1134, May 2002.

[10] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1319–1343, June 2002.

[11] V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1567–1580, Aug. 2004.

[12] A. Lozano, A. Tulino, and S. Verdú, "Mercury/waterfilling: power allocation with arbitrary input constellations," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, 2005.

[13] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, Inc., 3rd ed., 1976.

[14] S. Verdú, "The exponential distribution in information theory," *Problems of Information Transmission*, vol. 32, pp. 86–95, 1996. Translated from Problemy Peredachi Informatsii, Vol. 32, No. 1, pp. 100–111, January–March, 1996.