ISIT 2009, Seoul, Korea, June 28 - July 3, 2009

Relative Entropy and Score Function: New Information–Estimation Relationships through Arbitrary Additive Perturbation

Dongning Guo
Department of Electrical Engineering & Computer Science
Northwestern University
Evanston, IL 60208, USA

***Abstract*—This paper establishes new information–estimation relationships pertaining to models with additive noise of arbitrary distribution. In particular, we study the change in the relative entropy between two probability measures when both of them are perturbed by a small amount of the same additive noise. It is shown that the rate of the change with respect to the energy of the perturbation can be expressed in terms of the mean squared difference of the score functions of the two distributions, and, rather surprisingly, is unrelated to the distribution of the perturbation otherwise. The result holds true for the classical relative entropy (or Kullback–Leibler distance), as well as two of its generalizations: Rényi's relative entropy and the f-divergence. The result generalizes a recent relationship between the relative entropy and mean squared errors pertaining to Gaussian noise models, which in turn supersedes many previous information–estimation relationships. A generalization of the de Bruijn identity to non-Gaussian models can also be regarded as consequence of this new result.**

I. Introduction

To date, a number of connections between basic information measures and estimation measures have been discovered. By *information measures* we mean notions which describe the amount of information, such as entropy and mutual information, as well as several closely related quantities, such as differential entropy and relative entropy (also known as information divergence or Kullback–Leibler distance). By *estimation measures* we mean key notions in estimation theory, which include in particular the mean squared error (MSE) and Fisher information, among others.

An early such connection is attributed to de Bruijn [1] which relates the differential entropy of an arbitrary random variable corrupted by Gaussian noise and its Fisher information:

$$\frac{\mathrm{d}}{\mathrm{d}\delta} h\left(X+\sqrt{\delta}\,N\right)=\frac{1}{2}J\left(X+\sqrt{\delta}\,N\right) \tag{1}$$

for every $\delta \geq 0$, where X denotes an arbitrary random variable and $N \sim \mathcal{N}(0,1)$ denotes a standard Gaussian random variable independent of X throughout this paper. Here $J(Y)$ denotes the Fisher information of its distribution with respect to (w.r.t.) the location family. The de Bruijn identity is equivalent to a recent connection between the input–output mutual information and the minimum mean-square error (MMSE) of a Gaussian model [2]:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I\left(X;\sqrt{\gamma}\,X+N\right)=\frac{1}{2}\mathsf{mmse}\big(P_X,\gamma\big) \tag{2}$$

where $X \sim P_X$ and $\mathsf{mmse}\big(P_X,\gamma\big)$ denotes the MMSE of estimating X given $\sqrt{\gamma}\,X+N$. The parameter $\gamma \geq 0$ is understood as the signal-to-noise ratio (SNR) of the Gaussian model. By-products of formula (2) include the representation of the entropy, differential entropy and the non-Gaussianness (measured in relative entropy) in terms of the MMSE [2]–[4]. Several generalizations and extensions of the previous results are found in [5]–[7]. Moreover, the derivative of the mutual information and entropy w.r.t. channel gains have also been studied for non-additive-noise channels [7], [8].

Among the aforementioned information measures, relative entropy is the most general and versatile in the sense that it is defined for distributions which are discrete, continuous, or neither, and all the other information measures can be easily expressed in terms of relative entropy. The following relationship between the relative entropy and Fisher information is known [9]: Let $\{p_\theta\}$ be a family of probability density functions (pdfs) parameterized by $\theta \in \mathbb{R}$. Then

$$D(p_{\theta+\delta}\|p_\theta)=\big(\delta^2/2\big)J(p_\theta)+o\big(\delta^2\big) \tag{3}$$

where $J(p_\theta)$ is the Fisher information of p_θ w.r.t. θ.

In a recent work [10], Verdú established an interesting relationship between the relative entropy and mismatched estimation. Let $\mathsf{mse}_Q(P,\gamma)$ represent the MSE for estimating the input X of distribution P to a Gaussian channel of SNR equal to γ based on the channel output, with the estimator assuming the prior distribution of X to be Q. Then

$$\begin{aligned}2\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathsf{D}\left(P*\mathcal{N}\left(0,\gamma^{-1}\right)\|Q*\mathcal{N}\left(0,\gamma^{-1}\right)\right)&\\ =\mathsf{mse}_Q(P,\gamma)-\mathsf{mmse}(P,\gamma)&\end{aligned} \tag{4}$$

where the convolution $P*\mathcal{N}\left(0,\gamma^{-1}\right)$ represents the distribution of $X+N/\sqrt{\gamma}$ with $X \sim P$. Obviously $\mathsf{mse}_Q(P,\gamma)=\mathsf{mmse}(P,\gamma)$ if Q is identical to P. The formula is particularly satisfying because the left-hand side (l.h.s.) is an information-theoretic measure of the mismatch between two distributions, whereas the right-hand side (r.h.s.) measures the mismatch using an estimation-theoretic metric, i.e., the increase in the

This work is supported by the NSF under grant CCF-0644344 and DARPA under grant W911NF-07-1-0028.

978-1-4244-4313-0/09/$25.00 ©2009 IEEE

estimation error due to the mismatch.

In another recent work, Narayanan and Srinivasa [11] consider an additive non-Gaussian noise channel model and provide the following generalization of the de Bruijn identity:

$$\frac{\mathrm{d}}{\mathrm{d}\delta} h\left(X+\sqrt{\delta}\,V\right)\bigg|_{\delta=0^+} = \frac{1}{2}J(X) \tag{5}$$

where the pdf of V is symmetric about 0, twice differentiable, and of unit variance but otherwise arbitrary. The significance of (5) is that the derivative does not depend on the detailed statistics of the noise. Thus, if we view the differential entropy as a manifold of the distribution of the perturbation $\sqrt{\delta}\,V$, then the geometry of the manifold appears to be locally a bowl which is uniform in every direction of the perturbation.

In this work, we consider the change in the relative entropy between two distributions when both of them are perturbed by an infinitesimal amount of the same arbitrary additive noise. We show that the rate of this change can be expressed as the mean-squared difference of the *score functions* of the distributions. Note that the score function is an important notion in estimation theory, whose mean square is the Fisher information. Like formula (5), the new general relationship turns out to be independent of the noise distribution.

The general relationship is found to hold for both the classical relative entropy (or Kullback–Leibler distance) and the more general Rényi's relative entropy, as well as the general f-divergence due to Csiszár and independently Ali and Silvey [12]. In the special case of Gaussian perturbations, it is shown that (1), (2), (4), (5) can all be obtained as consequence of the new result.

II. Main Results

Theorem 1: Let Ψ denote an arbitrary distribution with zero mean and variance δ. Let P and Q be two distributions whose respective pdfs p and q are twice differentiable. If $P \ll Q$ and

$$\lim_{z\to\infty} \frac{\mathrm{d}}{\mathrm{d}z}\left[p(z)\log\frac{p(z)}{q(z)}\right] = 0 \tag{6}$$

then

$$\frac{\mathrm{d}}{\mathrm{d}\delta} D(P*\Psi\|Q*\Psi)\bigg|_{\delta=0^+}$$
$$= -\frac{1}{2}\int_{-\infty}^{\infty} p(z)\left(\nabla\log\frac{p(z)}{q(z)}\right)^2 \mathrm{d}z \tag{7}$$
$$= -\frac{1}{2}\mathsf{E}_P\left\{(\nabla\log p(Z)-\nabla\log q(Z))^2\right\} \tag{8}$$

where the expectation in (8) is taken with respect to $Z\sim P$.

The classical relative entropy (Kullback-Leibler distance) is defined for two probability measures $P\ll Q$ as

$$D(P\|Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P. \tag{9}$$

When the corresponding densities exist, it is also customary to denote the relative entropy by $D(p\|q)$.

The notation $\mathrm{d}/\mathrm{d}\delta$ in Theorem 1 can be understood as taking derivative w.r.t. the variance of the distribution Ψ with its shape fixed, i.e., Ψ is the distribution of $\sqrt{\delta}\,V$ with the random variable V fixed. We note that the r.h.s. of (8) does not depend on the distribution of V, i.e., the change in the relative entropy due to small perturbation is proportional to the variance of the perturbation but independent of its shape. Thus the notation $\mathrm{d}/\mathrm{d}\delta$ is not ambiguous.

For every function f, let ∇f denote its derivative f' for notational convenience. For every differentiable pdf p, the function $\nabla\log p(x) = p'(x)/p(x)$ is known as its score function, hence the r.h.s. of (8) is the mean squared difference of two score functions. As the previous result (4), this is satisfying because both sides of (8) represent some error due to the mismatch between the prior distribution q supplied to the estimator and the actual distribution p. Obviously, if p and q are identical, then both sides of the formula are equal to zero; otherwise, the derivative is negative (i.e., perturbation reduces relative entropy).

Consider now the Rényi relative entropy, which is defined for two probability measures $P\ll Q$ and every $\alpha>0$ as

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^{\alpha-1}\mathrm{d}P \tag{10}$$

where $D_1(P\|Q)$ is defined as the classical relative entropy $D(P\|Q)$ because $\lim_{\alpha\to1}D_\alpha(P\|Q)=D(P\|Q)$.

Theorem 2: Let the distributions P, Q and Ψ be defined the same way as in Theorem 1. Let δ denote the variance of Ψ. If $P\ll Q$ and

$$\lim_{z\to\infty}\frac{\mathrm{d}}{\mathrm{d}z}\left[p^\alpha(z)\,q^{1-\alpha}(z)\right]=0 \tag{11}$$

then

$$\frac{\mathrm{d}}{\mathrm{d}\delta} D_\alpha(P*\Psi\|Q*\Psi)\bigg|_{\delta=0^+}$$
$$= -\frac{\alpha}{2}\int_{-\infty}^{\infty}\left(\nabla\log\frac{p(z)}{q(z)}\right)^2 \frac{p^\alpha(z)q^{1-\alpha}(z)}{\int_{-\infty}^{\infty}p^\alpha(u)q^{1-\alpha}(u)\,\mathrm{d}u}\,\mathrm{d}z. \tag{12}$$

Note that as $\alpha\to1$, the r.h.s. of (12) becomes the r.h.s. of (7). We also point out that, similar to that in (7), the outer integral in (12) can be viewed as the mean square difference of two scores $(\nabla\log p(Z)-\nabla\log q(Z))$ with the pdf of Z being proportional to $p^\alpha(z)q^{1-\alpha}(z)$.

Theorems 1 and 2 are quite general because conditions (6) and (11) are satisfied by most distributions of interest. For example, if both p and q belong to the exponential family, then the derivatives in (6) and (11) also vanish exponentially fast. Not all distributions satisfy those conditions. This is because that although the functions $p(z)$ and $p(z)\log(p(z)/q(z))$ integrate to 1 and $D(P\|Q)$ respectively, they need not vanish. For example, $p(z)$ may consist of narrower and narrower Gaussian pulses of the same height as $z\to\infty$, so that not only $p(z)$ does not vanish, but $p'(z)$ is unbounded.

Another family of generalized relative entropy, called the f-divergence, was introduced by Csiszár and independently by

Ali and Silvey (see e.g., [12]). It is defined for $P \ll Q$ as

$$I_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q. \tag{13}$$

Theorem 3: Let the distributions P, Q and Ψ be defined the same way as in Theorem 1. Let δ denote the variance of Ψ. Suppose the second derivative of $f(\cdot)$ exists and is denoted by $f''(\cdot)$. If $P \ll Q$ and

$$\lim_{z\to\infty} \frac{\mathrm{d}}{\mathrm{d}z}\left[q(z) f\left(\frac{p(z)}{q(z)}\right)\right] = 0 \tag{14}$$

then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta} I_f\left(P*\Psi \big\| Q*\Psi\right)\Big|_{\delta=0^+} \\ = -\frac{1}{2}\int_{-\infty}^{\infty} q(y) f''\left(\frac{p(y)}{q(y)}\right)\left(\nabla\frac{p(y)}{q(y)}\right)^2 \mathrm{d}y\ .\end{aligned} \tag{15}$$

The integral in (15) can still be expressed in terms of the difference of the score functions because

$$\nabla(p(y)/q(y)) = (p(y)/q(y))[\nabla \log p(y) - \nabla \log q(y)]. \tag{16}$$

Note that the special case of $f(t) = t\log t$ corresponds to the Kullback–Leibler distance, whereas the case of $f(t) = (t - t^q)/(q-1)$ corresponds to the Tsallis relative entropy [13]. Indeed, Theorem 1 is a special case of Theorem 3.

III. Proof

The key property that underlies Theorems 1–3 is the following observation of the local geometry of an additive-noise-perturbed distribution made in [11]:

Lemma 1: Let the pdf $p(\cdot)$ of a random variable Z be twice differentiable. Let p_δ denote the pdf of $Z + \sqrt{\delta}V$ where V is of zero mean and unit variance, and is independent of Z. Then for every $y \in \mathbb{R}$, as $\delta \to 0^+$,

$$\frac{\partial}{\partial\delta} p_\delta(y)\Big|_{\delta=0^+} = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}y^2} p(y). \tag{17}$$

Formula (17) allows the derivative w.r.t. the energy of the perturbation δ to be transformed to the second derivative of the original pdf. In Appendix we provide a brief proof for Lemma 1 which is slightly different than that in [11]. Note that Lemma 1 does not require the distribution of the perturbation to be symmetric as is required in [11].

In the following we first prove Theorem 3, which implies Theorem 1 as a special case, and then prove Theorem 2.

A. Proof for Theorem 3

Let V be a random variable with fixed distribution P_V. For convenience, we use a shorthand p_δ to denote the pdf of the random variable $Y = Z + \sqrt{\delta}V$ with $(Z, V) \sim P \times P_V$. Similarly, let q_δ denote the pdf of $Y = Z + \sqrt{\delta}V$ with $(Z, V) \sim Q \times P_V$. Clearly,

$$\frac{\mathrm{d}}{\mathrm{d}\delta} I_f\left(P*\Psi \big\| Q*\Psi\right) = \frac{\mathrm{d}}{\mathrm{d}\delta} I_f(p_\delta\|q_\delta). \tag{18}$$

For any single-variable function g, let g' and g'' denote its first and second derivative respectively. Consider now

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\delta} I_f(p_\delta\|q_\delta) \\ &= \frac{\mathrm{d}}{\mathrm{d}\delta}\int_{-\infty}^{\infty} q_\delta(y) f\left(\frac{p_\delta(y)}{q_\delta(y)}\right)\mathrm{d}y \quad (19)\\ &= \int_{-\infty}^{\infty} \frac{\partial q_\delta(y)}{\partial\delta} f\left(\frac{p_\delta(y)}{q_\delta(y)}\right) + q_\delta(y)\frac{\partial}{\partial\delta} f\left(\frac{p_\delta(y)}{q_\delta(y)}\right)\mathrm{d}y \quad (20)\\ &= \int_{-\infty}^{\infty} \frac{\partial q_\delta(y)}{\partial\delta}\left[f\left(\frac{p_\delta(y)}{q_\delta(y)}\right) - \frac{p_\delta(y)}{q_\delta(y)} f'\left(\frac{p_\delta(y)}{q_\delta(y)}\right)\right] \\ &\qquad + \frac{\partial p_\delta(y)}{\partial\delta} f'\left(\frac{p_\delta(y)}{q_\delta(y)}\right)\mathrm{d}y. \quad (21)\end{aligned}$$

Invoking Lemma 1 on (21) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta} I_f(p_\delta\|q_\delta)\Big|_{\delta=0^+} = \frac{1}{2}\int_{-\infty}^{\infty} p''(y) f'\left(\frac{p(y)}{q(y)}\right) \\ + q''(y)\left[f\left(\frac{p(y)}{q(y)}\right) - \frac{p(y)}{q(y)} f'\left(\frac{p(y)}{q(y)}\right)\right]\mathrm{d}y.\end{aligned} \tag{22}$$

To proceed, we reorganize the integrand in (22) to the desired form. The key technique is integration by parts, which we carry out implicitly with the help of a modest amount of foresight. For convenience, we use p and q as shorthand for $p(y)$ and $q(y)$ respectively. We use the fact $g'h = (gh)' - gh'$ to rewrite the integrand in (22) as

$$\begin{aligned}&p'' f'\left(\frac{p}{q}\right) + q''\left[f\left(\frac{p}{q}\right) - \frac{p}{q} f'\left(\frac{p}{q}\right)\right] \\ &= \left[p' f'\left(\frac{p}{q}\right) + q'\left[f\left(\frac{p}{q}\right) - \frac{p}{q} f'\left(\frac{p}{q}\right)\right]\right]' \\ &\quad - p'\left[f'\left(\frac{p}{q}\right)\right]' - q'\left[f\left(\frac{p}{q}\right) - \frac{p}{q} f'\left(\frac{p}{q}\right)\right]'\ .\end{aligned} \tag{23}$$

Combining the first two terms and simplifying the last term on the r.h.s. of (23) yield

$$\left[q f\left(\frac{p}{q}\right)\right]'' - p'\left[f'\left(\frac{p}{q}\right)\right]' + \frac{q'p}{q}\left[f'\left(\frac{p}{q}\right)\right]'\ . \tag{24}$$

The first term in (24) integrates to zero by assumption (14). The last two terms in (24) can be combined to obtain

$$-q\left(\frac{p}{q}\right)'\left[f'\left(\frac{p}{q}\right)\right]' = -q\left(\nabla\frac{p}{q}\right)^2 f''\left(\frac{p}{q}\right). \tag{25}$$

Collecting the results from (22) to (25), we have

$$\begin{aligned}\frac{\mathrm{d}I_f(P_\delta\|Q_\delta)}{\mathrm{d}\delta}\Big|_{\delta=0^+} \\ = -\frac{1}{2}\int_{-\infty}^{\infty} q(y)\left(\nabla\frac{p(y)}{q(y)}\right)^2 f''\left(\frac{p(y)}{q(y)}\right)\mathrm{d}y\end{aligned} \tag{26}$$

which is equivalent to (15). Hence the proof of Theorem 3.

We note that the preceding calculation is tantamount to two uses of integration by parts. The treatment here, however, requires the minimum regularity conditions on the densities p and q.

B. Proof for Theorem 2

Consider now

$$\frac{\mathrm{d}}{\mathrm{d}\delta} D_\alpha(p_\delta \| q_\delta)$$
$$= \frac{1}{\alpha-1}\frac{\mathrm{d}}{\mathrm{d}\delta}\log\int_{-\infty}^{\infty} p_\delta^\alpha(y) q_\delta^{\alpha-1}(y)\,\mathrm{d}y \tag{27}$$
$$= \frac{1}{\alpha-1}\int_{-\infty}^{\infty}\frac{\partial\left(p_\delta^\alpha(y)q_\delta^{\alpha-1}(y)\right)}{\partial\delta}\,\mathrm{d}y \Bigg/ \int_{-\infty}^{\infty} p_\delta^\alpha(y)q_\delta^{\alpha-1}(y)\,\mathrm{d}y. \tag{28}$$

By Lemma 1, the integral in the numerator in (28) can be written as

$$\int_{-\infty}^{\infty}\alpha\left(\frac{p_\delta(y)}{q_\delta(y)}\right)^{\alpha-1}\frac{\partial p_\delta(y)}{\partial\delta} + (1-\alpha)\left(\frac{p_\delta(y)}{q_\delta(y)}\right)^{\alpha}\frac{\partial q_\delta(y)}{\partial\delta}\,\mathrm{d}y$$
$$= \int_{-\infty}^{\infty}\frac{\alpha}{2}\left(\frac{p(y)}{q(y)}\right)^{\alpha-1}p''(y) + \frac{1-\alpha}{2}\left(\frac{p(y)}{q(y)}\right)^{\alpha}q''(y)\,\mathrm{d}y \tag{29}$$

at $\delta = 0^+$. Note that (11) implies that

$$\alpha\left(\frac{p(y)}{q(y)}\right)^{\alpha-1}p'(y) + (1-\alpha)\left(\frac{p(y)}{q(y)}\right)^{\alpha}q'(y) \tag{30}$$

vanishes as $y \to \infty$. Using integration by parts, we further equate the integral on the r.h.s. of (29) to

$$\int_{-\infty}^{\infty} -\frac{\alpha}{2}p'\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{p}{q}\right)^{\alpha-1} + \frac{\alpha-1}{2}q'\frac{\mathrm{d}}{\mathrm{d}y}\left(\frac{p}{q}\right)^{\alpha}\mathrm{d}y. \tag{31}$$

The integrand in (31) can be written as

$$\frac{\alpha(1-\alpha)}{2}\left(p'\left(\frac{p}{q}\right)^{\alpha-2} - q'\left(\frac{p}{q}\right)^{\alpha-1}\right)\left(\frac{p}{q}\right)'. \tag{32}$$

In the following we omit the coefficient $\alpha(1-\alpha)/2$ to write the remaining terms in (32) as

$$\left(p'\left(\frac{p}{q}\right)^{\alpha-2} - q'\left(\frac{p}{q}\right)^{\alpha-1}\right)\frac{p'q-pq'}{q^2}$$
$$= (p')^2 p^{\alpha-2}q^{1-\alpha} - 2p'q'p^{\alpha-1}q^{-\alpha} + p^\alpha q^{-1-\alpha}(q')^2 \tag{33}$$
$$= (\nabla\log p - \nabla\log q)^2 p^\alpha q^{1-\alpha}. \tag{34}$$

Collecting the preceding results from (28) to (34), we have established (12) in Theorem 2.

IV. Recovering Existing Information–Estimation Relationships Using Theorem 1

A. Mutual information and MMSE

We first use Theorem 1 to recover formula (2) established in [2]. For convenience, consider the following alternative Gaussian model:

$$Z = X + \sigma W \tag{35}$$

where X and $W \sim \mathcal{N}(0,1)$ are independent so that σ^2 represents the noise variance. It suffices to show the following result, which is equivalent to (2),

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)}I(X;X+\sigma W) = -\frac{1}{2\sigma^4}\mathsf{mmse}\left(P_X, \frac{1}{\sigma^2}\right). \tag{36}$$

For any $x \in \mathbb{R}$, let $P_{Z|X=x}$ denote the distribution of Z as the output of the model (35) conditioned on $X = x$. The mutual information can be expressed as

$$I(X;X+\sigma W) = I(X;Z) = D(P_{Z|X}\|P_Z|P_X) \tag{37}$$

which is the average of $D(P_{Z|X=x}\|P_Z)$ over x according to the distribution P_X, which does not depend on σ^2. Let $N \sim \mathcal{N}(0,1)$ be independent of Z. Consider the derivative of $D(P_{Z|X=x}\|P_Z)$ w.r.t. σ^2, or equivalently, by introducing a small perturbation,

$$\frac{\mathrm{d}}{\mathrm{d}\delta}D\Big(P_{Z+\sqrt{\delta}N|X=x}\Big\|P_{Z+\sqrt{\delta}N}\Big)\Big|_{\delta=0^+}$$
$$= -\frac{1}{2}\mathsf{E}\left\{\left(\nabla\log p_{Z|X=x}(Z) - \nabla\log p_Z(Z)\right)^2\right\} \tag{38}$$

due to Theorem 1, where the expectation is over $P_{Z|X=x}$, which is a Gaussian distribution centered at x with variance σ^2. The first score is easy to evaluate: $\nabla\log p_{Z|X=x}(Z) = (x-Z)/\sigma^2$. The second score is determined by the following simple variation of a result due to Esposito [14] (see also Lemma 2 in [2]):

Lemma 2: $\nabla\log p_Z(z) = (\mathsf{E}\{X \mid Z=z\} - z)/\sigma^2$.

Clearly, the r.h.s. of (38) becomes

$$-\mathsf{E}_{P_{Z|X=x}}\left\{(x - \mathsf{E}\{X\mid Z\})^2\right\}/(2\sigma^4), \tag{39}$$

the average of which over x is equal to the r.h.s. of (36). Thus (36) is established, and so is (2).

B. Differential Entropy and MMSE

Consider again the model (35). It is not difficult to see

$$D(P_{Z+\sqrt{\delta}\,N}\|\mathcal{N}(0,\sigma^2+\delta))$$
$$= \frac{1}{2}\log\left(2\pi e\left(\sigma^2+\delta\right)\right) + \frac{\mathsf{E}X^2/2}{\sigma^2+\delta} - h\big(Z+\sqrt{\delta}\,N\big). \tag{40}$$

By Theorem 1 and Lemma 2, we have

$$\frac{\mathrm{d}}{\mathrm{d}\delta}D\left(P_{Z+\sqrt{\delta}\,N}\|\mathcal{N}(0,\sigma^2+\delta)\right)\Big|_{\delta=0^+}$$
$$= -\frac{1}{2}\mathsf{E}\left\{\left(\frac{\mathsf{E}\{X\mid Z\} - z}{\sigma^2} + \frac{z}{\sigma^2}\right)^2\right\} \tag{41}$$
$$= -\mathsf{E}\left\{(\mathsf{E}\{X\mid Z\})^2\right\}\Big/(2\sigma^4). \tag{42}$$

Plugging into (40), we have

$$\frac{\mathrm{d}}{\mathrm{d}(\sigma^2)}\left(h(Z) - \frac{1}{2}\log(2\pi e(1+\sigma^2))\right)$$
$$= \frac{1}{2}\frac{1}{(1+\sigma^2)\sigma^2} - \frac{\mathsf{E}X^2 - (\mathsf{E}\{X\mid Z\})^2}{2\sigma^4}. \tag{43}$$

Note that $\mathsf{E}X^2 - (\mathsf{E}\{X\mid Z\})^2 = \mathsf{mmse}(P_X, 1/\sigma^2)$. Moreover, $h(X) = h(Z)|_{\sigma=0}$, and $h(Z) - \frac{1}{2}\log(2\pi e(1+\sigma^2))$ vanishes as $\sigma^2 \to \infty$. Therefore, by integrating w.r.t. σ^2 from

0 to ∞, we obtain

$$h(X) = \frac{1}{2}\log(2\pi e) + \frac{1}{2}\int_0^\infty \frac{1}{s^2}\mathsf{mmse}\left(P_X, \frac{1}{s}\right) - \frac{1}{s(s+1)}\,\mathrm{d}s \tag{44}$$

which is equivalent to the integral expression in [2], in which we use the SNR as the integral variable.

C. Relative Entropy and MMSE

The connection between relative entropy and MMSE (4) can also be regarded as a special case of Theorem 1. Consider again the model (35) and apply Theorem 1. We have

$$\begin{aligned} &-2\frac{\mathrm{d}}{\mathrm{d}\delta}D\left(P_{X+\sqrt{\delta}\,N}\|Q_{X+\sqrt{\delta}\,N}\right)\\ &= \mathsf{E}_P\left\{\left(\nabla\log p_Z(X+\sqrt{\delta}\,N) - \nabla\log q_Z(X+\sqrt{\delta}\,N)\right)^2\right\}. \end{aligned} \tag{45}$$

By Lemma 2, the r.h.s. of (45) can be rewritten as

$$\begin{aligned} &\mathsf{E}_P\left\{(\mathsf{E}_P\{X \mid Z\} - \mathsf{E}_Q\{X \mid Z\})^2\right\}\\ &= \mathsf{E}_P\left\{[(X - \mathsf{E}_Q\{X \mid Z\}) - (X - \mathsf{E}_P\{X \mid Z\})]^2\right\} &(46)\\ &= \mathsf{E}_P\left\{(X - \mathsf{E}_Q\{X \mid Z\})^2\right\} + \mathsf{E}_P\left\{(X - \mathsf{E}_P\{X \mid Z\})^2\right\}\\ &\quad - 2\mathsf{E}_P\left\{(X - \mathsf{E}_Q\{X \mid Z\})(X - \mathsf{E}_P\{X \mid Z\})\right\}. &(47) \end{aligned}$$

Using the orthogonality of $(X - \mathsf{E}_P\{X \mid Z\})$ and every function of Z under probability measure P, we can replace $\mathsf{E}_Q\{X \mid Z\}$ in the last term by $\mathsf{E}_P\{X \mid Z\}$ (which are both functions of Z), and continue the equality as

$$\begin{aligned} &\mathsf{E}_P\left\{(X - \mathsf{E}_Q\{X \mid Z\})^2\right\} + \mathsf{E}_P\left\{(X - \mathsf{E}_P\{X \mid Z\})^2\right\}\\ &\quad - 2\mathsf{E}_P\left\{(X - \mathsf{E}_P\{X \mid Z\})(X - \mathsf{E}_P\{X \mid Z\})\right\}\\ &= \mathsf{E}_P\left\{(X - \mathsf{E}_Q\{X \mid Z\})^2\right\} - \mathsf{E}_P\left\{(X - \mathsf{E}_P\{X \mid Z\})^2\right\} &(48)\\ &= \underset{Q_X}{\mathsf{mse}}(P_X, \gamma) - \mathsf{mmse}(P_X, \gamma) &(49) \end{aligned}$$

where $\gamma = 1/\delta$. Hence yields the desired formula (4).

D. Differential Entropy and Fisher Information

The generalized de Bruijn identity (5) can be recovered basically by inspection of (8). Consider a distribution Q_Z which is uniform on $[-m, m]$ with m being a large number and vanishes smoothly outside the interval (e.g., a raised-cosine function with roll-off). Then $Q_{Z+\sigma N}$ remains essentially uniform, so that $\nabla\log q_Z(z) \approx 0$ over almost all the probability mass of P_Z. As $m \to \infty$, (8) reduces to (5).

V. Concluding Remarks

The relationships connecting the score function and various forms of relative entropy shown in this paper are the most general for additive-noise models to this date. It is by now clear that such derivative relationships between basic information- and estimation-theoretic measures rely on neither the normality of the additive perturbation, nor the logarithm functional in classical information measures. The results, however, do not directly translate into integral relationships unless the noise is Gaussian, which has the infinite divisibility property.

VI. Acknowledgements

The author would like to thank Sergio Verdú for sharing an earlier conjecture of his on the relationship between the Kullback–Leibler distance and MSE, which inspired this work.

Appendix
Proof of Lemma 1

Proof: Recall that p_δ denotes the distribution of $Y = X + \sqrt{\delta}\,V$. Denote the characteristic function of p_δ as

$$\varphi(u, \delta) = \mathsf{E}\left\{e^{iuY}\right\}. \tag{50}$$

Due to independence of X and V,

$$\begin{aligned} \varphi(s, \delta) &= \mathsf{E}\left\{e^{iuX}\right\}\mathsf{E}\left\{e^{iu\sqrt{\delta}\,V}\right\} &(51)\\ &= \varphi(s, 0)\,\mathsf{E}\left\{\sum_{k=0}^{\infty}\frac{\left(iu\sqrt{\delta}\,V\right)^k}{k!}\right\} &(52)\\ &= \varphi(s, 0)\left[1 + \frac{\delta(iu)^2}{2} + \sum_{k=3}^{\infty}\frac{\left(iu\sqrt{\delta}\right)^k}{k!}\mathsf{E}V^k\right] &(53) \end{aligned}$$

where we have used the assumptions that V has zero mean and unit variance. Note also that the series sum in (53) vanishes as $o(\delta)$. Taking the inverse Fourier transform on both sides of (53) yields

$$p_\delta(y) = p_0(y) + \frac{1}{2}\delta\frac{\partial^2}{\partial y^2}p_0(y) + o(\delta). \tag{54}$$

Hence Lemma 1 is proved as $p_0(y) = p(y)$. ■

References

[1] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, 1959.

[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, "Proof of entropy power inequalities via MMSE," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 1011–1015, Seattle, WA, USA, July 2006.

[4] S. Verdú and D. Guo, "A simple proof of the entropy power inequality," *IEEE Trans. Inform. Theory*, pp. 2165–2166, May 2006.

[5] M. Zakai, "On mutual information, likelihood ratios, and estimation error for the additive Gaussian channel," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3017–3024, Sept. 2005.

[6] D. Guo, S. Shamai, and S. Verdú, "Additive non-Gaussian noise channels: Mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. Inform. Theory*, Adelaide, Australia, Sept. 2005.

[7] D. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inform. Theory*, vol. 53, pp. 453–470, Feb. 2007.

[8] C. Measson, A. Montanari, and R. Urbanke, "Maxwell Construction: The Hidden Bridge Between Iterativew and Maximum a Posteriori Decoding," *IEEE Trans. Inform. Theory*, vol. 54, pp. 5277–5307, 2008.

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

[10] S. Verdú, "Mismatched estimation and relative entropy," in *Proc. IEEE Int. Symp. Inform. Theory*, Seoul, Korea, 2009.

[11] K. R. Narayanan and A. R. Srinivasa, "On the thermodynamic temperature of a general distribution." http://arxiv.org/abs/0711.1460v2, 2007.

[12] I. Csiszr, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, pp. 261–273, 2008.

[13] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[14] R. Esposito, "On a relation between detection and estimation indecision theory," *Information and Control*, vol. 12, pp. 116–120, 1968.