Mutual Information and Minimum Mean-Square Error in Gaussian Channels

Dongning Guo, *Member, IEEE*, Shlomo Shamai (Shitz), *Fellow, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—This paper deals with arbitrarily distributed finite-power input signals observed through an additive Gaussian noise channel. It shows a new formula that connects the input–output mutual information and the minimum mean-square error (MMSE) achievable by optimal estimation of the input given the output. That is, the derivative of the mutual information (nats) with respect to the signal-to-noise ratio (SNR) is equal to half the MMSE, regardless of the input statistics. This relationship holds for both scalar and vector signals, as well as for discrete-time and continuous-time noncausal MMSE estimation.**

This fundamental information-theoretic result has an unexpected consequence in continuous-time nonlinear estimation: For any input signal with finite power, the causal filtering MMSE achieved at SNR is equal to the average value of the noncausal smoothing MMSE achieved with a channel whose SNR is chosen uniformly distributed between 0 and SNR.

***Index Terms*—Gaussian channel, minimum mean-square error (MMSE), mutual information, nonlinear filtering, optimal estimation, smoothing, Wiener process.**

I. Introduction

THIS paper is centered around two basic quantities in information theory and estimation theory, namely, the *mutual information* between the input and the output of a channel, and the *minimum mean-square error* (MMSE) in estimating the input given the output. The key discovery is a relationship between the mutual information and MMSE that holds regardless of the input distribution, as long as the input–output pair are related through additive Gaussian noise.

Take for example the simplest scalar real-valued Gaussian channel with an arbitrary fixed input distribution. Let the signal-to-noise ratio (SNR) of the channel be denoted by snr. Both the input–output mutual information and the MMSE are monotone functions of the SNR, denoted by $I(\mathsf{snr})$ and $\mathsf{mmse}(\mathsf{snr})$, respectively. This paper finds that the mutual information in nats and the MMSE satisfy the following relationship regardless of the input statistics:

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{1}$$

Simple as it is, the identity (1) was unknown before this work. It is trivial that one can compute the value of one monotone function given the value of another (e.g., by simply composing the inverse of the latter function with the former); what is quite surprising here is that the overall transformation (1) not only is strikingly simple but is also independent of the input distribution. In fact, this relationship and its variations hold under arbitrary input signaling and the broadest settings of Gaussian channels, including discrete-time and continuous-time channels, either in scalar or vector versions.

In a wider context, the mutual information and mean-square error are at the core of information theory and estimation theory, respectively. The input–output mutual information is an indicator of how much coded information can be pumped through a channel reliably given a certain input signaling, whereas the MMSE measures how accurately each individual input sample can be recovered using the channel output. Interestingly, (1) shows the strong relevance of mutual information to estimation and filtering and provides a noncoding operational characterization for mutual information. Thus, not only is the significance of an identity like (1) self-evident, but the relationship is intriguing and deserves thorough exposition.

At zero SNR, the right-hand side of (1) is equal to one half of the input variance. In that special case, the formula, and in particular, the fact that at low-SNR mutual information is insensitive to the input distribution has been remarked before [1]–[3]. Relationships between the local behavior of mutual information at vanishing SNR and the MMSE of the estimation of the output given the input are given in [4].

Formula (1) can be proved using the new "incremental channel" approach which gauges the decrease in mutual information due to an infinitesimally small additional Gaussian noise. The change in mutual information can be obtained as the input–output mutual information of a derived Gaussian channel whose SNR is infinitesimally small, a channel for which the mutual information is essentially linear in the estimation error, and hence relates the rate of mutual information increase to the MMSE.

Another rationale for the relationship (1) traces to the geometry of Gaussian channels, or, more tangibly, the geometric properties of the likelihood ratio associated with signal detection in Gaussian noise. Basic information-theoretic notions are

Manuscript received March 22, 2004; revised November 9, 2004. This work was supported in part by the National Science Foundation under Grants NCR-0074277 and CCR-0312879.

D. Guo was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA. He is now with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, 60208 USA (e-mail: dGuo@Northwestern.edu).

S. Shamai (Shitz) is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, 32000 Haifa, Israel (e-mail: sshlomo@ee.technion.ac.il).

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.

Digital Object Identifier 10.1109/TIT.2005.844072

firmly associated with the likelihood ratio, and foremost the mutual information is expressed as the expectation of the log-likelihood ratio of conditional and unconditional measures. The likelihood ratio also plays a fundamental role in detection and estimation, e.g., in hypothesis testing it is compared to a threshold to decide which hypothesis to take. Moreover, the likelihood ratio is central in the connection of detection and estimation, in either continuous-time [5]–[7] or discrete-time setting [8]. In fact, Esposito [9] and Hatsell and Nolte [10] noted simple relationships between conditional mean estimation and the gradient and Laplacian of the log-likelihood ratio, respectively, although they did not import mutual information into the picture. Indeed, the likelihood ratio bridges information measures and basic quantities in detection and estimation, and in particular, the estimation errors (e.g., [11]).

In continuous-time signal processing, both the causal (filtering) MMSE and noncausal (smoothing) MMSE are important performance measures. Suppose for now that the input is a stationary process with arbitrary but fixed statistics. Let $\mathsf{cmmse}(\mathsf{snr})$ and $\mathsf{mmse}(\mathsf{snr})$ denote the causal and noncausal MMSEs, respectively, as a function of the SNR. This paper finds that formula (1) holds literally in this continuous-time setting, i.e., the derivative of the mutual information rate is equal to half the noncausal MMSE. Furthermore, by using this new information-theoretic identity, an unexpected fundamental result in nonlinear filtering is unveiled. That is, the filtering MMSE is equal to the mean value of the smoothing MMSE:

$$\mathsf{cmmse}(\mathsf{snr}) = \mathsf{E}\{\mathsf{mmse}(\Gamma)\} \tag{2}$$

where Γ is chosen uniformly distributed between 0 and snr. In fact, stationarity of the input is not required if the MMSEs are defined as time averages.

Relationships between the causal and noncausal estimation errors have been studied for the particular case of linear estimation (or Gaussian inputs) in [12], where a bound on the loss due to the causality constraint is quantified. Capitalizing on earlier research on the "estimator-correlator" principle by Kailath and others (see [13]), Duncan [14], [15], Zakai[1] and Kadota *et al.* [17] pioneered the investigation of relations between the mutual information and causal filtering of continuous-time signals observed in white Gaussian noise. In particular, Duncan showed that the input–output mutual information can be expressed as a time integral of the causal MMSE [15]. Duncan's relationship has proven to be useful in various applications in information theory and statistics [17]–[20]. There are also a number of other works in this area, most notably those of Liptser [21] and Mayer-Wolf and Zakai [22], where the rate of increase in the mutual information between the sample of the input process at the current time and the entire past of the output process is expressed in the causal estimation error and certain Fisher informations. Similar results were also obtained for discrete-time models by Bucy [23]. In [24], Shmelev devised a general, albeit complicated, procedure to obtain the optimal smoother from the optimal filter.

[1]Duncan's theorem was independently obtained by Zakai in the general setting of inputs that may depend causally on the noisy output in a 1969 unpublished Bell Labs Memorandum (see [16, ref. 53]).

The new relationship (1) in continuous-time and Duncan's theorem are proved in this paper using the incremental channel approach with increments in additional noise and additional observation time, respectively. Formula (2) connecting filtering and smoothing MMSE's is then proved by comparing (1) to Duncan's theorem. A non-information-theoretic proof is not yet known for (2).

In the discrete-time setting, identity (1) still holds, while the relationship between the mutual information and the causal MMSEs takes a different form: We show that the mutual information is sandwiched between the filtering error and the prediction error.

The remainder of this paper is organized as follows. Section II gives the central result (1) for both scalar and vector channels along with four different proofs and discussion of applications. Section III gives the continuous-time channel counterpart along with the fundamental nonlinear filtering–smoothing relationship (2), and a fifth proof of (1). Discrete-time channels are briefly dealt with in Section IV. Section V studies general random transformations observed in additive Gaussian noise, and offers a glimpse at feedback channels. Section VI gives new representations for entropy, differential entropy, and mutual information for arbitrary distributions.

II. Scalar and Vector Gaussian Channels

A. The Scalar Channel

Consider a pair or real-valued random variables (X, Y) related by[2]

$$Y = \sqrt{\mathsf{snr}}X + N \tag{3}$$

where $\mathsf{snr} \geq 0$ and the $N \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable independent of X. Then X and Y can be regarded as the input and output, respectively, of a single use of a scalar Gaussian channel with an SNR of snr.[3] The input–output conditional probability density is described by

$$p_{Y|X;\mathsf{snr}}(y|x;\mathsf{snr}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y-\sqrt{\mathsf{snr}}x)^2\right]. \tag{4}$$

Upon the observation of the output Y, one would like to infer the information bearing input X. The *mutual information* between X and Y is

$$I(X;Y) = \mathsf{E}\left\{\log \frac{p_{Y|X;\mathsf{snr}}(Y|X;\mathsf{snr})}{p_{Y;\mathsf{snr}}(Y;\mathsf{snr})}\right\} \tag{5}$$

where $p_{Y;\mathsf{snr}}$ denotes the well-defined marginal probability density function of the output

$$p_{Y;\mathsf{snr}}(y;\mathsf{snr}) = \mathsf{E}\{p_{Y|X;\mathsf{snr}}(y|X;\mathsf{snr})\}. \tag{6}$$

[2]In this paper, random objects are denoted by upper case letters and their values denoted by lower case letters. The expectation $\mathsf{E}\{\cdot\}$ is taken over the joint distribution of the random variables within the brackets.

[3]If $\mathsf{E}X^2 = 1$ then snr complies with the usual notion of signal-to-noise power ratio; otherwise, snr can be regarded as the gain in the output SNR due to the channel. The results in this paper do not require $\mathsf{E}X^2 = 1$.

The mutual information is clearly a function of snr, which we denote by

$$I(\mathsf{snr}) = I(X; \sqrt{\mathsf{snr}}X + N). \tag{7}$$

The error of an estimate, $f(Y)$, of the input X based on the observation Y can be measured in mean-square sense

$$\mathsf{E}\{(X - f(Y))^2\}. \tag{8}$$

It is well known that the minimum value of (8), referred to as the MMSE, is achieved by the conditional mean estimator

$$\widehat{X}(Y; \mathsf{snr}) = \mathsf{E}\{X \mid Y; \mathsf{snr}\}. \tag{9}$$

The MMSE is also a function of snr, which is denoted by

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{mmse}(X \mid \sqrt{\mathsf{snr}}X + N). \tag{10}$$

To start with, consider the special case when the input distribution P_X is standard Gaussian. The input–output mutual information is then the well-known channel capacity under input power constraint [25]

$$I(\mathsf{snr}) = \frac{1}{2}\log(1 + \mathsf{snr}). \tag{11}$$

Meanwhile, the conditional mean estimate of the Gaussian input is merely a scaling of the output

$$\widehat{X}(Y; \mathsf{snr}) = \frac{\sqrt{\mathsf{snr}}}{1 + \mathsf{snr}}Y \tag{12}$$

and, hence, the MMSE is

$$\mathsf{mmse}(\mathsf{snr}) = \frac{1}{1 + \mathsf{snr}}. \tag{13}$$

From (11) and (13), an immediate observation is

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr})\log e \tag{14}$$

where the base of logarithm is consistent with the mutual information unit. To avoid numerous $\log e$ factors, henceforth we adopt natural logarithms and use nats as the unit of all information measures. It turns out that the relationship (14) holds not only for Gaussian inputs, but for any finite-power input.

Theorem 1: Let N be standard Gaussian, independent of X. For every input distribution P_X that satisfies $\mathsf{E}X^2 < \infty$

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X; \sqrt{\mathsf{snr}}X + N) = \frac{1}{2}\mathsf{mmse}(X \mid \sqrt{\mathsf{snr}}X + N). \tag{15}$$

Proof: See Section II-C. □

The identity (15) reveals an intimate and intriguing connection between Shannon's mutual information and optimal estimation in the Gaussian channel (3), namely, the rate of the mutual information increase as the SNR increases is equal to half the MMSE achieved by the optimal (in general nonlinear) estimator.

In addition to the special case of Gaussian inputs, Theorem 1 can also be verified for another simple and important input signaling: ± 1 with equal probability. The conditional mean estimate for such an input is given by

$$\widehat{X}(Y; \mathsf{snr}) = \tanh(\sqrt{\mathsf{snr}}Y). \tag{16}$$

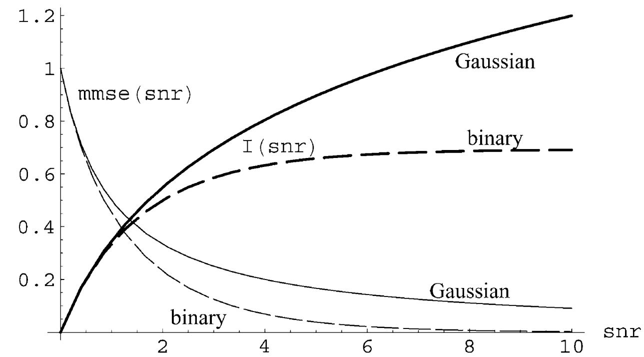

Fig. 1. The mutual information (in nats) and MMSE of scalar Gaussian channel with Gaussian and equiprobable binary inputs, respectively.

The MMSE and the mutual information are obtained as

$$\mathsf{mmse}(\mathsf{snr}) = 1 - \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} \tanh(\mathsf{snr} - \sqrt{\mathsf{snr}}y)\mathrm{d}y \tag{17}$$

and (e.g., [26, p. 274] and [27, Problem 4.22])

$$I(\mathsf{snr}) = \mathsf{snr} - \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} \log\cosh(\mathsf{snr} - \sqrt{\mathsf{snr}}y)\mathrm{d}y \tag{18}$$

respectively. Appendix I verifies that (17) and (18) satisfy (15).

For illustration purposes, the MMSE and the mutual information are plotted against the SNR in Fig. 1 for Gaussian and equiprobable binary inputs.

B. The Vector Channel

Multiple-input multiple-output (MIMO) systems are frequently described by the vector Gaussian channel

$$\boldsymbol{Y} = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{X} + \boldsymbol{N} \tag{19}$$

where $\boldsymbol{H}$ is a deterministic $L \times K$ matrix and the noise $\boldsymbol{N}$ consists of independent standard Gaussian entries. The input $\boldsymbol{X}$ (with distribution $P_{\boldsymbol{X}}$) and the output $\boldsymbol{Y}$ are column vectors of appropriate dimensions.

The input and output are related by a Gaussian conditional probability density

$$p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x}; \mathsf{snr}) = (2\pi)^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{H}\boldsymbol{x}\|^2\right] \tag{20}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. The MMSE in estimating $\boldsymbol{H}\boldsymbol{X}$ is

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{E}\left\{\left\|\boldsymbol{H}\boldsymbol{X} - \boldsymbol{H}\widehat{\boldsymbol{X}}(\boldsymbol{Y}; \mathsf{snr})\right\|^2\right\} \tag{21}$$

where $\widehat{\boldsymbol{X}}(\boldsymbol{Y}; \mathsf{snr})$ is the conditional mean estimate. A generalization of Theorem 1 is the following.

Theorem 2: Let $\boldsymbol{N}$ be a vector with independent standard Gaussian components, independent of $\boldsymbol{X}$. For every $P_{\boldsymbol{X}}$ satisfying $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\boldsymbol{X}; \sqrt{\mathsf{snr}}\boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{22}$$

Proof: See Section II-C. □

A verification of (22) in the special case of Gaussian input with positive-definite covariance matrix $\boldsymbol{\Sigma}$ is straightforward. The covariance of the conditional mean estimation error is

$$\mathsf{E}\{(\boldsymbol{X}-\widehat{\boldsymbol{X}})(\boldsymbol{X}-\widehat{\boldsymbol{X}})^\top\} = (\boldsymbol{\Sigma}^{-1} + \mathsf{snr}\boldsymbol{H}^\top\boldsymbol{H})^{-1} \tag{23}$$

from which one can calculate

$$\mathsf{E}\{\|\boldsymbol{H}(\boldsymbol{X}-\widehat{\boldsymbol{X}})\|^2\} = \mathsf{tr}\{\boldsymbol{H}(\boldsymbol{\Sigma}^{-1} + \mathsf{snr}\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top\}. \tag{24}$$

The mutual information is [28]

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\log\det(\boldsymbol{I} + \mathsf{snr}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}) \tag{25}$$

where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is the unique positive-semidefinite symmetric matrix such that $(\boldsymbol{\Sigma}^{\frac{1}{2}})^2 = \boldsymbol{\Sigma}$. Clearly

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}} I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\mathsf{tr}\{(\boldsymbol{I} + \mathsf{snr}\,\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}})^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}\} \tag{26}$$

$$= \frac{1}{2}\mathsf{E}\{\|\boldsymbol{H}(\boldsymbol{X}-\widehat{\boldsymbol{X}})\|^2\}. \tag{27}$$

C. Incremental Channels

The central relationship given in Sections II-A and -B can be proved in various, rather different, ways. The most enlightening proof is by considering what we call an incremental channel. A proof of Theorem 1 using the SNR-incremental channel is given next, while its generalization to the vector version is omitted but straightforward. Alternative proofs are relegated to later sections.

The key to the incremental-channel approach is to reduce the proof of the relationship for all SNRs to that for the special case of vanishing SNR, a domain in which we can capitalize on the following result.

Lemma 1: As $\delta \to 0$, the input–output mutual information of the canonical Gaussian channel

$$Y = \sqrt{\delta}Z + U \tag{28}$$

where $\mathsf{E}Z^2 < \infty$ and $U \sim \mathcal{N}(0,1)$ is independent of Z, is given by

$$I(Y;Z) = \frac{\delta}{2}\mathsf{E}(Z - \mathsf{E}Z)^2 + o(\delta). \tag{29}$$

Lemma 1 states that the mutual information is, essentially, half the SNR in the vicinity of zero SNR, but insensitive to the shape of the input distribution otherwise. Lemma 1 has been given in [2, Lemma 5.2.1] and [3, Theorem 4] (also implicitly in [1]).[4] Lemma 1 is the special case of Theorem 1 at vanishing SNR, which, by means of the incremental-channel method, can be bootstrapped to a proof of Theorem 1 for all SNRs.

Proof: [Theorem 1] Fix arbitrary $\mathsf{snr} > 0$ and $\delta > 0$. Consider a cascade of two Gaussian channels as depicted in Fig. 2

$$Y_1 = X + \sigma_1 N_1 \tag{30a}$$

$$Y_2 = Y_1 + \sigma_2 N_2 \tag{30b}$$

[4]A proof of Lemma 1 is given in Appendix II for completeness.

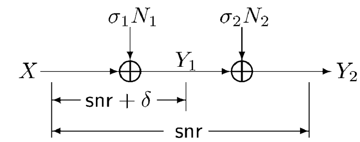

Fig. 2. An SNR-incremental Gaussian channel.

where X is the input, and N_1 and N_2 are independent standard Gaussian random variables. Let $\sigma_1, \sigma_2 > 0$ satisfy

$$\sigma_1^2 = \frac{1}{\mathsf{snr}+\delta} \tag{31a}$$

$$\sigma_1^2 + \sigma_2^2 = \frac{1}{\mathsf{snr}} \tag{31b}$$

so that the SNR of the first channel (30a) is $\mathsf{snr}+\delta$ and that of the composite channel is snr. Such a system is referred to as an SNR-incremental channel since the SNR increases by δ from Y_2 to Y_1.

Theorem 1 is equivalent to that, as $\delta \to 0$

$$I(X;Y_1) - I(X;Y_2) = I(\mathsf{snr}+\delta) - I(\mathsf{snr}) \tag{32}$$

$$= \frac{\delta}{2}\mathsf{mmse}(\mathsf{snr}) + o(\delta). \tag{33}$$

Noting that $X - Y_1 - Y_2$ is a Markov chain

$$I(X;Y_1) - I(X;Y_2) = I(X;Y_1,Y_2) - I(X;Y_2) \tag{34}$$

$$= I(X;Y_1|Y_2), \tag{35}$$

where (35) is the mutual information chain rule [29]. A linear combination of (30a) and (30b) yields

$$(\mathsf{snr}+\delta)Y_1 = \mathsf{snr}(Y_2 - \sigma_2 N_2) + \delta(X + \sigma_1 N_1) \tag{36}$$

$$= \mathsf{snr}Y_2 + \delta X + \sqrt{\delta}N \tag{37}$$

where we have defined

$$N = \frac{1}{\sqrt{\delta}}(\delta\sigma_1 N_1 - \mathsf{snr}\sigma_2 N_2). \tag{38}$$

Clearly, the incremental channel (30) is equivalent to (37) paired with (30b). Due to (31) and mutual independence of (X, N_1, N_2), N is a standard Gaussian random variable independent of X. Moreover, $(X, N, \sigma_1 N_1 + \sigma_2 N_2)$ are mutually independent since

$$\mathsf{E}\{N(\sigma_1 N_1 + \sigma_2 N_2)\} = \frac{1}{\sqrt{\delta}}\left(\delta\sigma_1^2 - \mathsf{snr}\sigma_2^2\right) = 0 \tag{39}$$

also due to (31). Therefore, N is independent of (X, Y_2) by (30). From (37), it is clear that

$$I(X;Y_1|Y_2=y_2) = I(X;\mathsf{snr}Y_2 + \delta X + \sqrt{\delta}N \mid Y_2 = y_2) \tag{40}$$

$$= I(X;\sqrt{\delta}X + N \mid Y_2 = y_2). \tag{41}$$

Hence, given $Y_2 = y_2$, (37) is equivalent to a Gaussian channel with SNR equal to δ where the input distribution is $P_{X|Y_2=y_2}$. Applying Lemma 1 to such a channel conditioned on $Y_2 = y_2$, one obtains

$$I(X;Y_1|Y_2=y_2) = \frac{\delta}{2}\mathsf{E}\left\{(X - \mathsf{E}\{X|Y_2=y_2\})^2 | Y_2 = y_2\right\} + o(\delta). \tag{42}$$

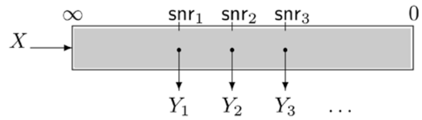

Fig. 3. A Gaussian pipe where noise is added gradually.

Taking the expectation over Y_2 on both sides of (42) yields

$$I(X;Y_1|Y_2) = \frac{\delta}{2}\mathsf{E}\left\{(X - \mathsf{E}\{X \mid Y_2\})^2\right\} + o(\delta) \qquad (43)$$

which establishes (33) by (35) together with the fact that

$$\mathsf{E}\left\{(X - \mathsf{E}\{X \mid Y_2\})^2\right\} = \mathsf{mmse}(\mathsf{snr}). \qquad (44)$$

Hence the proof of Theorem 1. □

Underlying the incremental-channel proof of Theorem 1 is the chain rule for information

$$I(X;Y_1,\ldots,Y_n) = \sum_{i=1}^{n} I(X;Y_i \mid Y_{i+1},\ldots,Y_n). \qquad (45)$$

When $X - Y_1 - \cdots - Y_n$ is a Markov chain, (45) becomes

$$I(X;Y_1) = \sum_{i=1}^{n} I(X;Y_i \mid Y_{i+1}) \qquad (46)$$

where we let $Y_{n+1} \equiv 0$. This applies to a train of outputs tapped from a Gaussian pipe where noise is added gradually until the SNR vanishes, as depicted in Fig. 3. The sum in (46) converges to an integral as $\{Y_i\}$ becomes a finer and finer sequence of Gaussian channel outputs. To see this, note from (43) that each conditional mutual information in (46) corresponds to a low-SNR channel and is essentially proportional to the MMSE times the SNR increment. This viewpoint leads us to an equivalent form of Theorem 1

$$I(\mathsf{snr}) = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma)\mathrm{d}\gamma. \qquad (47)$$

Therefore, as is illustrated by the curves in Fig. 1, the mutual information is equal to an accumulation of the MMSE as a function of the SNR due to the fact that an infinitesimal increase in the SNR adds to the total mutual information an increase proportional to the MMSE.

The infinite divisibility of Gaussian distributions, namely, the fact that a Gaussian random variable can always be decomposed as the sum of independent Gaussian random variables of smaller variances, is crucial in establishing the incremental channel (or the Markov chain). This property enables us to study the mutual information increase due to an infinitesimal increase in the SNR, and thus obtain the differential (15) and (22) in Theorems 1 and 2.

The following corollaries are immediate from Theorem 1 together with the fact that $\mathsf{mmse}(\mathsf{snr})$ is monotone decreasing.

Corollary 1: The mutual information $I(\mathsf{snr})$ is a concave function in snr.

Corollary 2: The mutual information can be bounded as

$$\begin{aligned}\mathsf{E}\{\mathsf{var}\{X|Y;\mathsf{snr}\}\} &= \mathsf{mmse}(\mathsf{snr}) &(48)\\ &\le \frac{2}{\mathsf{snr}} I(\mathsf{snr}) &(49)\\ &\le \mathsf{mmse}(0) = \mathsf{var}\{X^2\}. &(50)\end{aligned}$$

D. Applications and Discussions

1) Some Applications of Theorems 1 and 2: The newly discovered relationship between the mutual information and MMSE finds one of its first uses in relating code-division multiple-access (CDMA) channel spectral efficiencies (mutual information per dimension) under joint and separate decoding in the large-system limit [30], [31]. Under an arbitrary finite-power input distribution, Theorem 1 is invoked in [30] to show that the spectral efficiency under joint decoding is equal to the integral of the spectral efficiency under separate decoding as a function of the system load. The practical lesson therein is the optimality in the large-system limit of successive single-user decoding with cancellation of interference from already decoded users, and an individually optimal detection front end against yet undecoded users. This is a generalization to arbitrary input signaling of previous results that successive cancellation with a linear MMSE front end achieves the CDMA channel capacity under Gaussian inputs [32]–[35].

Relationships between information theory and estimation theory have been identified occasionally, yielding results in one area taking advantage of known results from the other. This is exemplified by the classical capacity–rate distortion relations, that have been used to develop lower bounds on estimation errors [36]. The fact that the mutual information and the MMSE determine each other by a simple formula also provides a new means to calculate or bound one quantity using the other. An upper (resp., lower) bound for the mutual information is immediate by bounding the MMSE for all SNRs using a suboptimal (resp., genie-aided) estimator. Lower bounds on the MMSE, e.g., [37], lead to new lower bounds on the mutual information.

An important example of such relationships is the case of Gaussian inputs. Under power constraints, Gaussian inputs are most favorable for Gaussian channels in information-theoretic sense (they maximize the mutual information); on the other hand, they are least favorable in estimation-theoretic sense (they maximize the MMSE). These well-known results are seen to be immediately equivalent through Theorem 1 (or Theorem 2 for the vector case). This also points to a simple proof of the result that Gaussian inputs achieve capacity by observing that the linear estimation upper bound for MMSE is achieved for Gaussian inputs.[5]

Another application of the new results is in the analysis of sparse-graph codes, where [38] has recently shown that the so-called generalized extrinsic information transfer (GEXIT) function plays a fundamental role. This function is defined for arbitrary codes and channels as minus the derivative of the input–output mutual information per symbol with respect to a channel quality parameter when the input is equiprobable on the codebook. According to Theorem 2, in the special case of the Gaussian channel the GEXIT function is equal to minus one half of the average MMSE of individual input symbols given the channel outputs. Moreover, [38] shows that (1) leads to a simple interpretation of the "area property" for Gaussian channels (cf. [39]). Inspired by Theorem 1, [40] also advocated

[5]The observations here are also relevant to continuous-time Gaussian channels in view of results in Section III.

using the mean-square error as the EXIT function for Gaussian channels.

As another application, the central theorems also provide an intuitive proof of de Bruijn's identity as is shown next.

2) De Bruijn's Identity: An interesting insight is that Theorem 2 is equivalent to the (multivariate) de Bruijn identity [41], [42]

$$\frac{\mathrm{d}}{\mathrm{d}t} h(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\boldsymbol{N}) = \frac{1}{2}\mathsf{tr}\{\boldsymbol{J}(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\boldsymbol{N})\} \tag{51}$$

where $\boldsymbol{N}$ is a vector with independent standard Gaussian entries, independent of $\boldsymbol{X}$. Here, $h(\cdot)$ stands for differential entropy and $\boldsymbol{J}(\cdot)$ for Fisher's information matrix [43], which is defined as[6]

$$\boldsymbol{J}(\boldsymbol{y}) = \mathsf{E}\left\{[\nabla \log p_{\boldsymbol{Y}}(\boldsymbol{y})][\nabla \log p_{\boldsymbol{Y}}(\boldsymbol{y})]^\top\right\}. \tag{52}$$

Let $\mathsf{snr} = 1/t$ and $\boldsymbol{Y} = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{X} + \boldsymbol{N}$. Then

$$h(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\boldsymbol{N}) = I(\boldsymbol{X};\boldsymbol{Y}) - \frac{L}{2}\log\frac{\mathsf{snr}}{2\pi e}. \tag{53}$$

Meanwhile

$$\boldsymbol{J}(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\boldsymbol{N}) = \mathsf{snr}\boldsymbol{J}(\boldsymbol{Y}). \tag{54}$$

Note that

$$p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) = \mathsf{E}\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\} \tag{55}$$

where $p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr})$ is the Gaussian density in (20). It can be shown that

$$\nabla \log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\widehat{\boldsymbol{X}}(\boldsymbol{y};\mathsf{snr}) - \boldsymbol{y}. \tag{56}$$

Plugging (56) into (52) and (54) gives

$$\boldsymbol{J}(\boldsymbol{Y}) = \boldsymbol{I} - \mathsf{snr}\,\boldsymbol{H}\mathsf{E}\{(\boldsymbol{X} - \widehat{\boldsymbol{X}})(\boldsymbol{X} - \widehat{\boldsymbol{X}})^\top\}\boldsymbol{H}^\top. \tag{57}$$

Now de Bruijn's identity (51) and Theorem 2 prove each other by (53) and (57). Noting this equivalence, the incremental-channel approach offers an intuitive alternative to the conventional technical proof of de Bruijn's identity obtained by integrating by parts (e.g., [29]). Although equivalent to de Bruijn's identity, Theorem 2 is important since mutual information and MMSE are more canonical operational measures than differential entropy and Fisher's information.

The Cramér-Rao bound states that the inverse of Fisher's information is a lower bound on estimation accuracy. The bound is tight for Gaussian channels, where Fisher's information matrix and the covariance of conditional mean estimation error determine each other by (57). In particular, for a scalar channel

$$J(\sqrt{\mathsf{snr}}\,X + N) = 1 - \mathsf{snr}\cdot\mathsf{mmse}(\mathsf{snr}). \tag{58}$$

[6]The gradient operator can be written as

$$\nabla = \left[\frac{\partial}{\partial y_1}, \dots, \frac{\partial}{\partial y_L}\right]^\top$$

symbolically. For any differentiable function $f : \mathbb{R}^L \to \mathbb{R}$, its gradient at any $\boldsymbol{y}$ is a column vector

$$\nabla f(\boldsymbol{y}) = \left[\frac{\partial f}{\partial y_1}(\boldsymbol{y}), \dots, \frac{\partial f}{\partial y_L}(\boldsymbol{y})\right]^\top.$$

3) Derivative of the Divergence: Consider an input–output pair (X,Y) connected through (3). The mutual information $I(X;Y)$ is the average value over the input X of the divergence $\mathsf{D}(P_{Y|X=x}\|P_Y)$. Refining Theorem 1, it is possible to directly obtain the derivative of the divergence given any value of the input.

Theorem 3: For every input distribution P_X that satisfies $\mathsf{E}X^2 < \infty$

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\mathsf{D}(P_{Y|X=x}\|P_Y) = \frac{1}{2}\mathsf{E}\bigg\{|X - X'|^2 \mid X = x\} - \frac{1}{2\sqrt{\mathsf{snr}}}\mathsf{E}\{X'N \mid X = x\bigg\} \tag{59}$$

where X' is an auxiliary random variable which is independent and identically distributed (i.i.d.) with X conditioned on $Y = \sqrt{\mathsf{snr}}\,X + N$.

The auxiliary random variable X' has an interesting physical meaning. It can be regarded as the output of the "retrochannel" [30], [31], which takes Y as the input and generates a random variable according to the posterior probability distribution $p_{X|Y;\mathsf{snr}}$. The joint distribution of (X,Y,X') is unique.

4) Multiple-Access Channel: A multiuser system in which users may transmit at different SNRs can be modeled by

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\Gamma}\boldsymbol{X} + \boldsymbol{N} \tag{60}$$

where $\boldsymbol{H}$ is deterministic $L \times K$ matrix known to the receiver

$$\boldsymbol{\Gamma} = \mathrm{diag}\{\sqrt{\mathsf{snr}_1}, \dots, \sqrt{\mathsf{snr}_K}\}$$

consists of the square-root of the SNRs of the K users, and $\boldsymbol{N}$ consists of independent standard Gaussian entries. The following theorem addresses the derivative of the total mutual information with respect to an individual user's SNR.

Theorem 4: For every input distribution $P_{\boldsymbol{X}}$ that satisfies $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$

$$\frac{\partial}{\partial \mathsf{snr}_k} I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\sum_{i=1}^{K}\sqrt{\frac{\mathsf{snr}_i}{\mathsf{snr}_k}}[\boldsymbol{H}^\top\boldsymbol{H}]_{ki}\mathsf{E}\{\mathsf{Cov}\{X_k, X_i|\boldsymbol{Y};\boldsymbol{\Gamma}\}\} \tag{61}$$

where $\mathsf{Cov}\{\cdot,\cdot|\cdot\}$ denotes conditional covariance.

The proof of Theorem 4 follows that of Theorem 2 in Appendix IV and is omitted. Theorems 1 and 2 can be recovered from Theorem 4 by setting $\mathsf{snr}_k = \mathsf{snr}$ for all k.

E. Alternative Proofs of Theorems 1 and 2

In this subsection, we give an alternative proof of Theorem 2, which is based on the geometric properties of the likelihood ratio between the output distribution and the noise distribution. This proof is a distilled version of the more general result of Zakai [44] (follow-up to this work) that uses the Malliavin calculus and shows that the central relationship between the mutual information and estimation error holds also in the abstract Wiener space. This alternative approach of Zakai makes use of relationships between conditional mean estimation and likelihood ratios due to Esposito [9] and Hatsell and Nolte [10].

As mentioned earlier, the central theorems also admit several other proofs. In fact, a third proof using the de Bruijn identity is already evident in Section II-D. A fourth proof of Theorems 1 and 2 by taking the derivative of the mutual information is given in Appendices III and IV. A fifth proof, taking advantage of results in the continuous-time domain, is relegated to Section III.

It suffices to prove Theorem 2 assuming $\boldsymbol{H}$ to be the identity matrix since one can always regard $\boldsymbol{HX}$ as the input. Let $\boldsymbol{Z} = \sqrt{\mathsf{snr}}\,\boldsymbol{X}$. Then the channel (19) is represented by the canonical L-dimensional Gaussian channel

$$\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{N}. \tag{62}$$

The mutual information, which is a conditional divergence, admits the following decomposition [1]:

$$I(\boldsymbol{Y};\boldsymbol{Z}) = \mathsf{D}(P_{\boldsymbol{Y}|\boldsymbol{Z}} \| P_{\boldsymbol{Y}} | P_{\boldsymbol{Z}}) \tag{63}$$

$$= \mathsf{D}(P_{\boldsymbol{Y}|\boldsymbol{Z}} \| P_{\boldsymbol{Y}'} | P_{\boldsymbol{Z}}) - \mathsf{D}(P_{\boldsymbol{Y}} \| P_{\boldsymbol{Y}'}) \tag{64}$$

where $P_{\boldsymbol{Y}'}$ is an arbitrary distribution as long as the two divergences on the right-hand side of (64) are well defined. Choose $\boldsymbol{Y}' = \boldsymbol{N}$. Then the mutual information can be expressed in terms of the divergence between the unconditional output distribution and the noise distribution

$$I(\boldsymbol{Y};\boldsymbol{Z}) = \frac{1}{2}\mathsf{E}\|\boldsymbol{Z}\|^2 - \mathsf{D}(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}). \tag{65}$$

Hence Theorem 2 is equivalent to the following.

Theorem 5: For every $P_{\boldsymbol{X}}$ satisfying $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\mathsf{D}(P_{\sqrt{\mathsf{snr}}\boldsymbol{X}+\boldsymbol{N}} \| P_{\boldsymbol{N}}) = \frac{1}{2}\mathsf{E}\{\|\mathsf{E}\{\boldsymbol{X} \mid \sqrt{\mathsf{snr}}\boldsymbol{X} + \boldsymbol{N}\}\|^2\}. \tag{66}$$

Theorem 5 can be proved using geometric properties of the likelihood ratio

$$l(\boldsymbol{y}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{N}}(\boldsymbol{y})}. \tag{67}$$

The following lemmas are important steps.

Lemma 2 (Esposito [9]): The gradient of the log-likelihood ratio gives the conditional mean estimate

$$\nabla \log l(\boldsymbol{y}) = \mathsf{E}\{\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\}. \tag{68}$$

Lemma 3 (Hatsell and Nolte [10]): The log-likelihood ratio satisfies Poisson's equation[7]

$$\nabla^2 \log l(\boldsymbol{y}) = \mathsf{E}\{\|\boldsymbol{Z}\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\} - \|\mathsf{E}\{\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\}\|^2. \tag{69}$$

From Lemmas 2 and 3

$$\mathsf{E}\{\|\boldsymbol{Z}\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\} = \nabla^2 \log l(\boldsymbol{y}) + \|\nabla \log l(\boldsymbol{y})\|^2. \tag{70}$$

[7]For any differentiable $\boldsymbol{f} : \mathbb{R}^L \to \mathbb{R}^L$

$$\nabla \cdot \boldsymbol{f} = \sum_{l=1}^{L} \frac{\partial f_l}{\partial y_l}.$$

If $\boldsymbol{f}$ is doubly differentiable, its Laplacian is defined as

$$\nabla^2 f = \nabla \cdot (\nabla f) = \sum_{l=1}^{L} \frac{\partial^2 f}{\partial y_l^2}.$$

The following result is immediate.

Lemma 4:

$$\mathsf{E}\{\|\boldsymbol{Z}\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\} = l^{-1}(\boldsymbol{y})\nabla^2 l(\boldsymbol{y}). \tag{71}$$

A Proof of Theorem 5 is obtained by taking the derivative directly.

Proof: [Theorem 5] Note that the likelihood ratio can be expressed as

$$l(\boldsymbol{y}) = \frac{\mathsf{E}\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\}}{p_{\boldsymbol{N}}(\boldsymbol{y})} \tag{72}$$

$$= \mathsf{E}\left\{\exp\left[\sqrt{\mathsf{snr}}\,\boldsymbol{y}^\top\boldsymbol{X} - \frac{\mathsf{snr}}{2}\|\boldsymbol{X}\|^2\right]\right\}. \tag{73}$$

Also, for any function $f(\cdot)$

$$\mathsf{E}\left\{f(\boldsymbol{X})\exp\left[\sqrt{\mathsf{snr}}\,\boldsymbol{y}^\top\boldsymbol{X} - \frac{\mathsf{snr}}{2}\|\boldsymbol{X}\|^2\right]\right\} = l(\boldsymbol{y})\mathsf{E}\{f(\boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{y}\}. \tag{74}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}}\,l(\boldsymbol{y}) = \frac{1}{2}l(\boldsymbol{y})\left[\frac{1}{\sqrt{\mathsf{snr}}}\,\boldsymbol{y}^\top\mathsf{E}\{\boldsymbol{X} \mid \boldsymbol{Y}=\boldsymbol{y}\} - \mathsf{E}\{\|\boldsymbol{X}\|^2 \mid \boldsymbol{Y}=\boldsymbol{y}\}\right] \tag{75}$$

$$= \frac{1}{2\mathsf{snr}}[l(\boldsymbol{y})\boldsymbol{y}^\top\nabla\log l(\boldsymbol{y}) - \nabla^2\log l(\boldsymbol{y})] \tag{76}$$

where (76) is due to Lemmas 2 and 4. Note that the order of expectation with respect to $P_{\boldsymbol{X}}$ and the derivative with respect to the SNR can be exchanged as long as the input has finite power by the Lebesgue (dominated) convergence theorem, [45], [46] (see also Lemma 8 in Appendix IV).

The divergence can be written as

$$\mathsf{D}(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}) = \int p_{\boldsymbol{Y}}(\boldsymbol{y}) \log \frac{p_{\boldsymbol{Y}}(\boldsymbol{Y})}{P_{\boldsymbol{N}}(\boldsymbol{Y})}\mathrm{d}\boldsymbol{y} \tag{77}$$

$$= \mathsf{E}\left\{l(\boldsymbol{N})\log l(\boldsymbol{N})\right\} \tag{78}$$

and its derivative

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}}\mathsf{D}(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}) = \mathsf{E}\left\{\log l(\boldsymbol{N})\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}}l(\boldsymbol{N})\right\}. \tag{79}$$

Again, the order of derivative and expectation can be exchanged by the Lebesgue convergence theorem. By (76), the derivative (79) can be evaluated as

$$\begin{aligned}
&\frac{1}{2\mathsf{snr}}\mathsf{E}\left\{l(\boldsymbol{N})\log l(\boldsymbol{N})\boldsymbol{N}\cdot\nabla\log l(\boldsymbol{N})\right\} \\
&\qquad - \frac{1}{2\mathsf{snr}}\mathsf{E}\left\{\log l(\boldsymbol{N})\nabla^2 l(\boldsymbol{N})\right\} \\
&= \frac{1}{2\mathsf{snr}}\mathsf{E}\{\nabla\cdot[l(\boldsymbol{N})\log l(\boldsymbol{N})\nabla\log l(\boldsymbol{N})] \\
&\qquad - \log l(\boldsymbol{N})\nabla^2 l(\boldsymbol{N})\} &(80)\\
&= \frac{1}{2\mathsf{snr}}\mathsf{E}\left\{l(\boldsymbol{N})\|\nabla\log l(\boldsymbol{N})\|^2\right\} &(81)\\
&= \frac{1}{2\mathsf{snr}}\mathsf{E}\|\nabla\log l(\boldsymbol{Y})\|^2 &(82)\\
&= \frac{1}{2}\mathsf{E}\|\mathsf{E}\left\{\boldsymbol{X} \mid \boldsymbol{Y}\right\}\|^2 &(83)
\end{aligned}$$

where to write (80) we used the following relationship (which can be checked by integration by parts) satisfied by a standard Gaussian vector $\boldsymbol{N}$:

$$\mathsf{E}\left\{\boldsymbol{N}^\top\boldsymbol{f}(\boldsymbol{N})\right\} = \mathsf{E}\left\{\nabla\cdot\boldsymbol{f}(\boldsymbol{N})\right\} \tag{84}$$

for every vector-valued differentiable function $\boldsymbol{f} : \mathbb{R}^l \to \mathbb{R}^l$ that satisfies $f_i(\boldsymbol{n})e^{-\frac{1}{2}n_i^2} \to 0$ as $n_i \to \infty$, $i = 1, \dots, L$. $\square$

F. Asymptotics of Mutual Information and MMSE

In the following, the asymptotics of the mutual information and MMSE at low and high SNRs are studied mainly for the scalar Gaussian channel.

The Lebesgue convergence theorem guarantees continuity of the MMSE estimate

$$\lim_{\mathsf{snr}\to 0} \mathsf{E}\{X \mid Y; \mathsf{snr}\} = \mathsf{E}X \tag{85}$$

and hence,

$$\lim_{\mathsf{snr}\to 0} \mathsf{mmse}(\mathsf{snr}) = \mathsf{mmse}(0) = \sigma_X^2 \tag{86}$$

where $\sigma_{(\cdot)}^2$ denotes the variance of a random variable. It has been shown in [3] that symmetric (proper-complex in the complex case) signaling is second-order optimal in terms of mutual information for in the low SNR regime.

A more refined study of the asymptotics is possible by examining the Taylor series expansion of a family of well-defined functions

$$q_i(y; \mathsf{snr}) = \mathsf{E}\left\{X^i p_{Y|X;\mathsf{snr}}(y \mid X; \mathsf{snr})\right\}, \qquad i = 0, 1, \dots. \tag{87}$$

Clearly

$$p_{Y;\mathsf{snr}}(y; \mathsf{snr}) = q_0(y; \mathsf{snr})$$

and the conditional mean estimate is expressed as

$$\mathsf{E}\{X \mid Y = y; \mathsf{snr}\} = \frac{q_1(y; \mathsf{snr})}{q_0(y; \mathsf{snr})}. \tag{88}$$

Meanwhile, by definition (5) and noting that $p_{Y|X;\mathsf{snr}}$ is Gaussian, one has

$$I(\mathsf{snr}) = -\frac{1}{2}\log(2\pi e) - \int q_0(y; \mathsf{snr}) \log q_0(y; \mathsf{snr}) \mathrm{d}y. \tag{89}$$

As $\mathsf{snr} \to 0$

$$\begin{aligned} q_i(y; \mathsf{snr}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \times \mathsf{E}\Bigg\{ X^i \Bigg[& 1 + Xy\mathsf{snr}^{\frac{1}{2}} + \frac{X^2}{2}(y^2 - 1)\mathsf{snr} \\ & + \frac{X^3}{6}(y^2 - 3)y\mathsf{snr}^{\frac{3}{2}} \\ & + \frac{X^4}{24}(y^4 - 6y^2 + 3)\mathsf{snr}^2 \\ & + \frac{X^5}{120}(15y - 10y^3 + y^5)\mathsf{snr}^{\frac{5}{2}} \\ & + \frac{X^6}{720}(y^6 - 15y^4 + 45y^2 - 15)\mathsf{snr}^3 \\ & + \mathcal{O}\left(\mathsf{snr}^{\frac{7}{2}}\right) \Bigg] \Bigg\}. \end{aligned} \tag{90}$$

Without loss of generality, it is assumed that the input X has zero mean and unit variance. Using (88)–(90), a finer characterization of the MMSE and mutual information is obtained as

$$\begin{aligned} \mathsf{mmse}(\mathsf{snr}) = 1 - \mathsf{snr} + \mathsf{snr}^2 - \frac{1}{6}\Big[& (\mathsf{E}X^4)^2 - 6\mathsf{E}X^4 \\ & - 2(\mathsf{E}X^3)^2 + 15 \Big] \mathsf{snr}^3 + \mathcal{O}(\mathsf{snr}^4) \end{aligned} \tag{91}$$

and

$$\begin{aligned} I(\mathsf{snr}) = \frac{1}{2}\mathsf{snr} - \frac{1}{4}\mathsf{snr}^2 + \frac{1}{6}\mathsf{snr}^3 - \frac{1}{48}\Big[& (\mathsf{E}X^4)^2 - 6\mathsf{E}X^4 \\ & - 2(\mathsf{E}X^3)^2 + 15 \Big] \mathsf{snr}^4 + \mathcal{O}(\mathsf{snr}^5) \end{aligned} \tag{92}$$

respectively. It is interesting to note that higher order moments than the mean and variance have no impact on the mutual information to the third order of the SNR.

The asymptotic properties carry over to the vector channel model (19) for finite-power inputs. The MMSE of a real-valued vector channel is obtained to the second order as

$$\begin{aligned} \mathsf{mmse}(\mathsf{snr}) = & \,\mathsf{tr}\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\} \\ & - \mathsf{snr} \cdot \mathsf{tr}\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\} + \mathcal{O}(\mathsf{snr}^2) \end{aligned} \tag{93}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the input vector. The input–output mutual information is straightforward by Theorem 2 (see also [4]). The asymptotics can be refined to any order of the SNR using the Taylor series expansion.

At high SNRs, the mutual information is upper-bounded for finite-alphabet inputs such as the binary one (18), whereas it can increase at the rate of $\frac{1}{2}\log \mathsf{snr}$ for Gaussian inputs. By Shannon's entropy power inequality [25], [29], given any symmetric input distribution with a density, there exists an $\alpha \in (0, 1]$ such that the mutual information of the scalar channel is bounded

$$\frac{1}{2}\log(1 + \alpha\mathsf{snr}) \le I(\mathsf{snr}) \le \frac{1}{2}\log(1 + \mathsf{snr}). \tag{94}$$

The MMSE behavior at high SNR depends on the input distribution. The decay can be as slow as $\mathcal{O}(1/\mathsf{snr})$ for Gaussian input, whereas for binary input, the MMSE decays as $e^{-2\mathsf{snr}}$. In fact, the MMSE can be made to decay faster than any given exponential for sufficiently skewed binary inputs [31].

III. Continuous-Time Gaussian Channels

The success in the discrete-time Gaussian channel setting in Section II can be extended to more technically challenging continuous-time models. Consider the following continuous-time Gaussian channel:

$$R_t = \sqrt{\mathsf{snr}} X_t + N_t, \qquad t \in [0, T] \tag{95}$$

where $\{X_t\}$ is the input process and $\{N_t\}$ a white Gaussian noise with a flat double-sided power spectrum density of unit height. Since $\{N_t\}$ is not second order, it is mathematically more convenient to study an equivalent model obtained by integrating the observations in (95). In a concise form, the input and output processes are related by a standard Wiener process $\{W_t\}$ independent of the input [47], [48]

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}} X_t \mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0, T]. \tag{96}$$

Also known as Brownian motion, $\{W_t\}$ is a continuous Gaussian process that satisfies

$$\mathsf{E}\{W_t W_s\} = \min(t, s), \qquad \forall\, t, s. \tag{97}$$

Instead of scaling the Brownian motion (as is customary in the literature), we choose to scale the input process so as to minimize notation in the analysis and results.

A. Mutual Information and MMSEs

We are concerned with three quantities associated with the model (96), namely, the causal MMSE achieved by optimal filtering, the noncausal MMSE achieved by optimal smoothing, and the mutual information between the input and output processes. As a convention, let X_a^b denote the process $\{X_t\}$ in the interval $[a,b]$. Also, let μ_X denote the probability measure induced by $\{X_t\}$ in the interval of interest, which, for concreteness we assume to be $[0,T]$. The input–output mutual information is defined by [49], [50]

$$I\left(X_0^T;Y_0^T\right)=\int\log\Phi\mathrm{d}\mu_{XY} \tag{98}$$

if the Radon-Nikodym derivative

$$\Phi=\frac{\mathrm{d}\mu_{XY}}{\mathrm{d}\mu_X\mathrm{d}\mu_Y} \tag{99}$$

exists. The causal and noncausal MMSEs at any time $t\in[0,T]$ are defined in the usual way

$$\mathsf{cmmse}(t,\mathsf{snr})=\mathsf{E}\left\{\left(X_t-\mathsf{E}\left\{X_t\mid Y_0^t;\mathsf{snr}\right\}\right)^2\right\} \tag{100}$$

and

$$\mathsf{mmse}(t,\mathsf{snr})=\mathsf{E}\left\{\left(X_t-\mathsf{E}\left\{X_t\mid Y_0^T;\mathsf{snr}\right\}\right)^2\right\}. \tag{101}$$

Recall the mutual information rate (mutual information per unit time) defined as

$$I(\mathsf{snr})=\frac{1}{T}I\left(X_0^T;Y_0^T\right). \tag{102}$$

Similarly, the average causal and noncausal MMSEs (per unit time) are defined as

$$\mathsf{cmmse}(\mathsf{snr})=\frac{1}{T}\int_0^T\mathsf{cmmse}(t,\mathsf{snr})\mathrm{d}t \tag{103}$$

and

$$\mathsf{mmse}(\mathsf{snr})=\frac{1}{T}\int_0^T\mathsf{mmse}(t,\mathsf{snr})\mathrm{d}t \tag{104}$$

respectively.

To start with, let $T\to\infty$ and assume that the input to the continuous-time model (96) is a stationary[8] Gaussian process with power spectrum $S_X(\omega)$. The mutual information rate was obtained by Shannon [51]

$$I(\mathsf{snr})=\frac{1}{2}\int_{-\infty}^{\infty}\log(1+\mathsf{snr}\,S_X(\omega))\frac{\mathrm{d}\omega}{2\pi}. \tag{105}$$

With Gaussian input, both optimal filtering and smoothing are linear. The noncausal MMSE is due to Wiener [52]

$$\mathsf{mmse}(\mathsf{snr})=\int_{-\infty}^{\infty}\frac{S_X(\omega)}{1+\mathsf{snr}\,S_X(\omega)}\frac{\mathrm{d}\omega}{2\pi} \tag{106}$$

and the causal MMSE is due to Yovits and Jackson [53]

$$\mathsf{cmmse}(\mathsf{snr})=\frac{1}{\mathsf{snr}}\int_{-\infty}^{\infty}\log(1+\mathsf{snr}\,S_X(\omega))\frac{\mathrm{d}\omega}{2\pi}. \tag{107}$$

From (105) and (106), it is easy to see that the derivative of the mutual information rate is equal to half the noncausal MMSE, i.e., the central formula (1) holds literally in this case. Moreover, (105) and (107) show that the mutual information rate is equal to the causal MMSE scaled by half the SNR, although, interestingly, this connection escaped Yovits and Jackson [53].

In fact, these relationships are true not only for Gaussian inputs.

Theorem 6: For every input process $\{X_t\}$ to the Gaussian channel (96) with finite average power, i.e.,

$$\int_0^T\mathsf{E}X_t^2\mathrm{d}t<\infty \tag{108}$$

the input–output mutual information rate and the average noncausal MMSE are related by

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}I(\mathsf{snr})=\frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{109}$$

Proof: See Section III-C. □

Theorem 7 (Duncan [15]): For any input process with finite average power

$$I(\mathsf{snr})=\frac{\mathsf{snr}}{2}\mathsf{cmmse}(\mathsf{snr}). \tag{110}$$

Together, Theorems 6 and 7 show that the mutual information, the causal MMSE and the noncausal MMSE satisfy a triangle relationship. In particular, using the information rate as a bridge, the causal MMSE is found to be equal to the noncausal MMSE averaged over SNR.

Theorem 8: For any input process with finite average power

$$\mathsf{cmmse}(\mathsf{snr})=\frac{1}{\mathsf{snr}}\int_0^{\mathsf{snr}}\mathsf{mmse}(\gamma)\mathrm{d}\gamma. \tag{111}$$

Equality (111) is a surprising fundamental relationship between causal and noncausal MMSEs. It is quite remarkable considering the fact that nonlinear filtering is usually a hard problem and few analytical expressions are known for the optimal estimation errors.

Although in general the optimal anticausal filter is different from the optimal causal filter, an interesting observation that follows from either Theorem 7 or Theorem 8 is that for stationary inputs, the average anticausal MMSE per unit time is equal to the average causal one. To see this, note that both mutual information and the average noncausal MMSE remains the same in reversed time and that white Gaussian noise is reversible.

It is worth pointing out that Theorems 6–8 are still valid if the time averages in (102)–(104) are replaced by their limits as $T\to\infty$. This is particularly relevant to the case of stationary inputs.

Besides Gaussian inputs, another example of the relation in Theorem 8 is an input process called the random telegraph waveform, where $\{X_t\}$ is a stationary Markov process with two equally probable states ($X_t=\pm 1$). See Fig. 4 for an illustration. Assume that the transition rate of the input Markov process is ν, i.e., for sufficiently small h

$$\mathsf{P}\{X_{t+h}=X_t\}=1-\nu h+o(h), \tag{112}$$

the expressions for the MMSEs, achieved by optimal filtering and smoothing, are obtained as [54], [55]

$$\mathsf{cmmse}(\mathsf{snr})=\frac{\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}e^{-\frac{2\nu u}{\mathsf{snr}}}\mathrm{d}u}{\int_1^\infty u^{\frac{1}{2}}(u-1)^{-\frac{1}{2}}e^{-\frac{2\nu u}{\mathsf{snr}}}\mathrm{d}u} \tag{113}$$

[8]For stationary input it would be more convenient to shift $[0,T]$ to $[-T/2,T/2]$ and then let $T\to\infty$ so that the causal and noncausal MMSEs at any time $t\in(-\infty,\infty)$ is independent of t. We stick to $[0,T]$ in this paper for notational simplicity in case of general inputs.

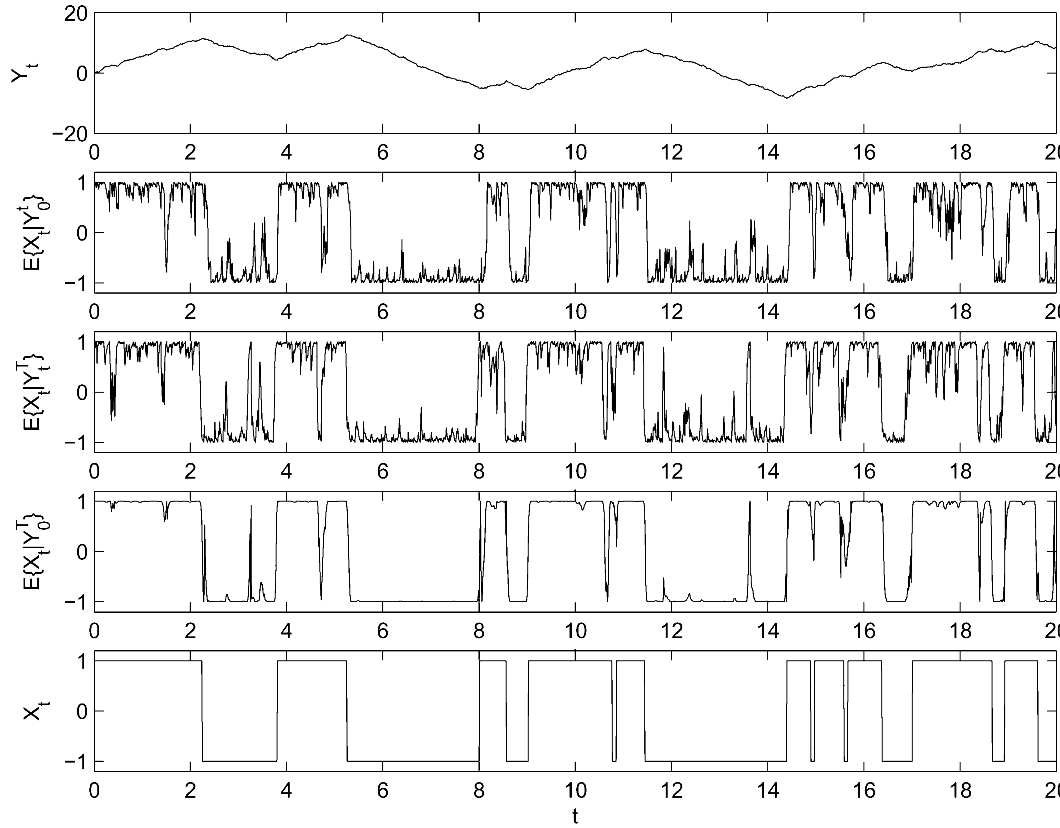

Fig. 4. Sample path of the input and output processes of an additive white Gaussian noise channel, the output of the optimal causal and anticausal filters, as well as the output of the optimal smoother. The input $\{X_t\}$ is a random telegraph waveform with unit transition rate. The SNR is 15 dB.

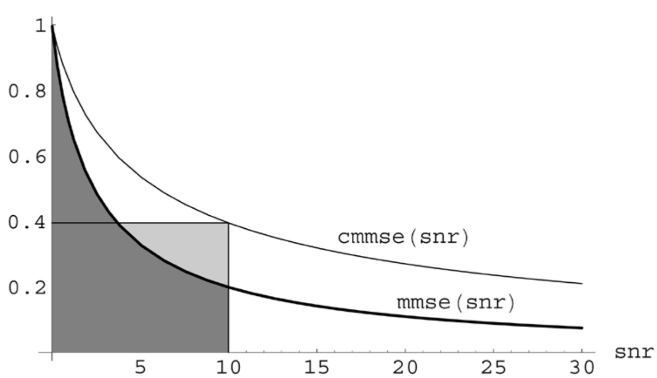

Fig. 5. The causal and noncausal MMSEs of the continuous-time Gaussian channel and a random telegraph waveform input with unit transition rate. The two shaded regions have the same area due to Theorem 8.

and

$$\mathsf{mmse}(\mathsf{snr}) = \frac{\int_{-1}^{1}\int_{-1}^{1} \frac{(1+xy)\exp\left[-\frac{2\nu}{\mathsf{snr}}\left(\frac{1}{1-x^2}+\frac{1}{1-y^2}\right)\right]}{-(1-x)^3(1-y)^3(1+x)(1+y)}\mathrm{d}x\mathrm{d}y}{\left[\int_1^\infty u^{\frac{1}{2}}(u-1)^{-\frac{1}{2}}e^{-\frac{2\nu u}{\mathsf{snr}}}\mathrm{d}u\right]^2} \tag{114}$$

respectively. The relationship (111) is verified in Appendix V. The MMSEs are plotted in Fig. 5 as functions of the SNR for unit transition rate.

Fig. 4 shows experimental results of the filtering and smoothing of the random telegraph signal corrupted by additive white Gaussian noise. The optimal causal filter follows Wonham [54]

$$\mathrm{d}\widehat{X}_t = -\left[2\nu\widehat{X}_t + \mathsf{snr}\,\widehat{X}_t\left(1-\widehat{X}_t^2\right)\right]\mathrm{d}t + \sqrt{\mathsf{snr}}\left(1-\widehat{X}_t^2\right)\mathrm{d}Y_t \tag{115}$$

where

$$\widehat{X}_t = \mathsf{E}\left\{X_t \mid Y_0^t\right\}. \tag{116}$$

The anticausal filter is merely a time reversal of the filter of the same type. The smoother is due to Yao [55]

$$\mathsf{E}\left\{X_t \mid Y_0^T\right\} = \frac{\mathsf{E}\left\{X_t \mid Y_0^t\right\} + \mathsf{E}\left\{X_t \mid Y_t^T\right\}}{1 + \mathsf{E}\left\{X_t \mid Y_0^t\right\}\mathsf{E}\left\{X_t \mid Y_t^T\right\}}. \tag{117}$$

B. Low- and High-SNR Asymptotics

Based on Theorem 8, one can study the asymptotics of the mutual information and MMSE under low SNRs. The causal and noncausal MMSE relationship implies that

$$\lim_{\mathsf{snr}\to 0} \frac{\mathsf{mmse}(0) - \mathsf{mmse}(\mathsf{snr})}{\mathsf{cmmse}(0) - \mathsf{cmmse}(\mathsf{snr})} = 2 \tag{118}$$

where

$$\mathsf{cmmse}(0) = \mathsf{mmse}(0) = \frac{1}{T}\int_0^T \mathsf{E}X_t^2 \mathrm{d}t. \tag{119}$$

Hence, the initial rate of decrease (with snr) of the noncausal MMSE is twice that of the causal MMSE.

In the high-SNR regime, there exist inputs that make the MMSE vanish exponentially fast with SNR. However, in the case of Gauss–Markov input processes, Steinberg *et al.* [56] observed that the causal MMSE is asymptotically twice the noncausal MMSE, as long as the input–output relationship is described by

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}h(X_t)\mathrm{d}t + \mathrm{d}W_t \tag{120}$$

where $h(\cdot)$ is a differentiable and increasing function. In the special case where $h(X_t) = X_t$, Steinberg *et al.* 's observation can be justified by noting that in the Gauss–Markov case, the smoothing MMSE satisfies [57]

$$\mathsf{mmse}(\mathsf{snr}) = \frac{c}{\sqrt{\mathsf{snr}}} + o\left(\frac{1}{\mathsf{snr}}\right) \tag{121}$$

which implies according to (111) that

$$\lim_{\mathsf{snr}\to\infty} \frac{\mathsf{cmmse}(\mathsf{snr})}{\mathsf{mmse}(\mathsf{snr})} = 2. \tag{122}$$

Unlike the universal factor of 2 result in (118) for the low-SNR regime, the 3-dB loss incurred by the causality constraint fails to hold in general in the high-SNR asymptote. For example, for the random telegraph waveform input, the causality penalty increases in the order of $\log \mathsf{snr}$ [55].

C. The SNR-Incremental Channel

Theorem 6 can be proved using the SNR-incremental channel approach developed in Section II. Consider a cascade of two Gaussian channels with independent noise processes

$$\mathrm{d}Y_{1t} = X_t\mathrm{d}t + \sigma_1\mathrm{d}W_{1t} \tag{123a}$$

$$\mathrm{d}Y_{2t} = \mathrm{d}Y_{1t} + \sigma_2\mathrm{d}W_{2t} \tag{123b}$$

where $\{W_{1t}\}$ and $\{W_{2t}\}$ are independent standard Wiener processes also independent of $\{X_t\}$, and σ_1 and σ_2 satisfy (31) so that the SNRs of the first channel and the composite channel are $\mathsf{snr} + \delta$ and snr, respectively. Following steps similar to those that lead to (37), it can be shown that

$$(\mathsf{snr} + \delta)\mathrm{d}Y_{1t} = \mathsf{snr}\,\mathrm{d}Y_{2t} + \delta X_t\mathrm{d}t + \sqrt{\delta}\mathrm{d}W_t \tag{124}$$

where $\{W_t\}$ is a standard Wiener process independent of $\{X_t\}$ and $\{Y_{2t}\}$. Hence, conditioned on the process $\{Y_{2t}\}$ in $[0, T]$, (124) can be regarded as a Gaussian channel with an SNR of δ. Similar to Lemma 1, the following result holds.

Lemma 5: As $\delta \to 0$, the input–output mutual information of the following Gaussian channel:

$$\mathrm{d}Y_t = \sqrt{\delta}Z_t\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0, T] \tag{125}$$

where $\{W_t\}$ is a standard Wiener process independent of the input $\{Z_t\}$, which satisfies

$$\int_0^T \mathsf{E}Z_t^2\mathrm{d}t < \infty \tag{126}$$

is given by the following:

$$\lim_{\delta\to 0} \frac{1}{\delta} I\left(Z_0^T; Y_0^T\right) = \frac{1}{2}\int_0^T \mathsf{E}(Z_t - \mathsf{E}Z_t)^2\mathrm{d}t. \tag{127}$$

Proof: See Appendix VI. □

Applying Lemma 5 to the Gaussian channel (124) conditioned on $\{Y_{2t}\}$ in $[0, T]$, one has

$$I\left(X_0^T; Y_{1,0}^T | Y_{2,0}^T\right) = \frac{\delta}{2}\int_0^T \mathsf{E}\left\{\left(X_t - \mathsf{E}\left\{X_t \mid Y_{2,0}^T\right\}\right)^2\right\}\mathrm{d}t + o(\delta). \tag{128}$$

Since $\{X_t\} - \{Y_{1t}\} - \{Y_{2t}\}$ is a Markov chain, the left-hand side of (128) is recognized as the mutual information increase

$$I\left(X_0^T; Y_{1,0}^T \mid Y_{2,0}^T\right) = I\left(X_0^T; Y_{1,0}^T\right) - I\left(X_0^T; Y_{2,0}^T\right) \tag{129}$$

$$= T[I(\mathsf{snr} + \delta) - I(\mathsf{snr})]. \tag{130}$$

By (130) and definition of the noncausal MMSE (101), (128) can be rewritten as

$$I(\mathsf{snr} + \delta) - I(\mathsf{snr}) = \frac{\delta}{2T}\int_0^T \mathsf{mmse}(t, \mathsf{snr})\mathrm{d}t + o(\delta). \tag{131}$$

Hence the proof of Theorem 6.

The property that independent Wiener processes sum up to a Wiener process is essential in the above proof. The incremental channel device is very useful in proving integral equations such as in Theorem 6.

D. The Time-Incremental Channel

Note that Duncan's theorem (Theorem 7) that links the mutual information and the causal MMSE is also an integral equation, although implicit, where the integral is with respect to time on the right-hand side of (110). Analogous to the SNR-incremental channel, one can investigate the mutual information increase due to an infinitesimal additional observation time of the channel output using a "time-incremental channel." This approach leads to a more intuitive proof of Duncan's theorem than the original one in [15], which relies on intricate properties of likelihood ratios and stochastic calculus.

Duncan's theorem is equivalent to

$$I\left(X_0^{t+\delta}; Y_0^{t+\delta}\right) - I\left(X_0^t; Y_0^t\right) = \delta\frac{\mathsf{snr}}{2}\mathsf{E}\left\{\left(X_t - \mathsf{E}\left\{X_t \mid Y_0^t\right\}\right)^2\right\} + o(\delta) \tag{132}$$

which is to say the mutual information increase due to the extra observation time is proportional to the causal MMSE. The left-hand side of (132) can be written as

$$\begin{aligned} &I\left(X_0^{t+\delta}; Y_0^{t+\delta}\right) - I\left(X_0^t; Y_0^t\right) \\ &= I\left(X_0^t, X_t^{t+\delta}; Y_0^t, Y_t^{t+\delta}\right) - I\left(X_0^t; Y_0^t\right) &(133)\\ &= I\left(X_t^{t+\delta}; Y_t^{t+\delta} \mid Y_0^t\right) + I\left(X_0^t; Y_t^{t+\delta} \mid X_t^{t+\delta}, Y_0^t\right) \\ &\quad + I\left(X_0^t, X_t^{t+\delta}; Y_0^t\right) - I\left(X_0^t; Y_0^t\right) &(134)\\ &= I\left(X_t^{t+\delta}; Y_t^{t+\delta} \mid Y_0^t\right) + I\left(X_0^t; Y_t^{t+\delta} \mid X_t^{t+\delta}, Y_0^t\right) \\ &\quad + I\left(X_t^{t+\delta}; Y_0^t \mid X_0^t\right). &(135)\end{aligned}$$

Since $Y_0^t - X_0^t - X_t^{t+\delta} - Y_t^{t+\delta}$ is a Markov chain, the last two mutual informations in (135) vanish due to conditional independence. Therefore,

$$I\left(X_0^{t+\delta}; Y_0^{t+\delta}\right) - I\left(X_0^t; Y_0^t\right) = I\left(X_t^{t+\delta}; Y_t^{t+\delta} | Y_0^t\right) \tag{136}$$

i.e., the increase in the mutual information is the conditional mutual information between the input and output during the extra time interval given the past observation. Note that conditioned on Y_0^t, the probability law of the channel in $(t, t+\delta)$ remains the same but with different input statistics due to conditioning on Y_0^t. Let us denote this new channel by

$$\mathrm{d}\tilde{Y}_t = \sqrt{\mathsf{snr}}\tilde{X}_t \mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0, \delta] \tag{137}$$

where the time duration is shifted to $[0, \delta]$, and the input process $\tilde{X}_0^\delta$ has the same law as $X_t^{t+\delta}$ conditioned on Y_0^t. Instead of looking at this new problem of an infinitesimal time interval $[0, \delta]$, we can convert the problem to a familiar one by an expansion in the time axis. Since $\sqrt{\delta} W_{t/\delta}$ is also a standard Wiener process, the channel (137) in $[0, \delta]$ is equivalent to a new channel described by

$$\mathrm{d}\tilde{\tilde{Y}}_\tau = \sqrt{\delta \mathsf{snr}}\tilde{\tilde{X}}_\tau \mathrm{d}\tau + \mathrm{d}W'_\tau, \qquad \tau \in [0, 1] \tag{138}$$

where $\tilde{\tilde{X}}_\tau = \tilde{X}_{\tau\delta}$, and $\{W'_t\}$ is a standard Wiener process. The channel (138) is of (fixed) unit duration but a diminishing SNR of $\delta\mathsf{snr}$. It is interesting to note that the trick here performs a "time–SNR" transform. By Lemma 5, the mutual information is

$$\begin{aligned} &I\left(X_t^{t+\delta}; Y_t^{t+\delta} | Y_0^t\right) \\ &= I\left(\tilde{\tilde{X}}_0^1; \tilde{\tilde{Y}}_0^1\right) && (139) \\ &= \frac{\delta \mathsf{snr}}{2} \int_0^1 \mathsf{E}\left(\tilde{\tilde{X}}_\tau - \mathsf{E}\tilde{\tilde{X}}_\tau\right)^2 \mathrm{d}\tau + o(\delta) && (140) \\ &= \frac{\delta \mathsf{snr}}{2} \int_0^1 \mathsf{E}\left\{\left(X_{t+\tau\delta} - \mathsf{E}\left\{X_{t+\tau\delta} \mid Y_0^t; \mathsf{snr}\right\}\right)^2\right\} \mathrm{d}\tau \\ &\quad + o(\delta) && (141) \\ &= \frac{\delta \mathsf{snr}}{2} \mathsf{E}\left\{\left(X_t - \mathsf{E}\left\{X_t \mid Y_0^t; \mathsf{snr}\right\}\right)^2\right\} + o(\delta) && (142) \end{aligned}$$

where (142) is justified by the continuity of the MMSE. The relation (132) is then established by (136) and (142), and hence the proof of Duncan's theorem.

Similar to the discussion in Section II-C, the integral equations in Theorems 6 and 7 proved by using the SNR- and time-incremental channels are also consequences of the mutual information chain rule applied to a Markov chain of the channel input and degraded versions of channel outputs. The independent-increment properties of Gaussian processes both SNR-wise and time-wise are quintessential in establishing the results.

E. A Fifth Proof of Theorem 1

A fifth proof of the mutual information and MMSE relation in the random variable/vector model can be obtained using continuous-time results. For simplicity, Theorem 1 is proved using Theorem 7. The proof can be easily modified to show Theorem 2, using the vector version of Duncan's theorem [15].

A continuous-time counterpart of the model (3) can be constructed by letting $X_t \equiv X$ for $t \in [0, 1]$ where X is a random variable independent of t

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, X \mathrm{d}t + \mathrm{d}W_t. \tag{143}$$

For every $u \in [0, 1]$, Y_u is a sufficient statistic of the observation Y_0^u for X (and X_0^u). This is because the process $\{Y_t - (t/u)Y_u\}$, $t \in [0, u]$, is independent of X (and X_0^u). Therefore, the input–output mutual information of the scalar channel (3) is equal to the mutual information of the continuous-time channel (143)

$$I(\mathsf{snr}) = I(X; Y_1) = I\left(X_0^1; Y_0^1\right). \tag{144}$$

Integrating both sides of (143), one has

$$Y_u = \sqrt{\mathsf{snr}} u X + W_u, \qquad u \in [0, 1] \tag{145}$$

where $W_u \sim \mathcal{N}(0, u)$. Note that (145) is a scalar Gaussian channel with a time-varying SNR which grows linearly from 0 to snr. Due to the sufficiency of Y_u, the MMSE of the continuous-time model given the observation Y_0^u, i.e., the causal MMSE at time u, is equal to the MMSE of a scalar Gaussian channel with an SNR of $u\mathsf{snr}$

$$\mathsf{cmmse}(u, \mathsf{snr}) = \mathsf{mmse}(u\mathsf{snr}). \tag{146}$$

By Duncan's theorem, the mutual information can be written as

$$\begin{aligned} I\left(X_0^1; Y_0^1\right) &= \frac{\mathsf{snr}}{2} \int_0^1 \mathsf{cmmse}(u, \mathsf{snr}) \mathrm{d}u && (147) \\ &= \frac{\mathsf{snr}}{2} \int_0^1 \mathsf{mmse}(u\mathsf{snr}) \mathrm{d}u && (148) \\ &= \frac{1}{2} \int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma) \mathrm{d}\gamma. && (149) \end{aligned}$$

Thus, Theorem 1 follows by also noticing (144).

Note that for constant input applied to a continuous-time Gaussian channel, the noncausal MMSE (101) at any time t is equal to the MMSE of a scalar channel with the same SNR

$$\mathsf{mmse}(t, u\mathsf{snr}) = \mathsf{mmse}(u\mathsf{snr}), \qquad \forall\, t \in [0, T]. \tag{150}$$

Together, (146) and (150) yield (111) for constant input by averaging over time u. Indeed, during any observation time interval of the continuous-time channel output, the SNR of the desired signal against noise is accumulated over time. The integral over time and the integral over SNR are interchangeable in this case. This is another example of the "time–SNR" transform which appeared in Section III-D.

Regarding the preceding proof, note that the constant input can be replaced by a general form of $Xh(t)$, where $h(t)$ is a deterministic signal.

IV. Discrete-Time Gaussian Channels

A. Mutual Information and MMSE

Consider a real-valued discrete-time Gaussian-noise channel of the form

$$Y_i = \sqrt{\mathsf{snr}} X_i + N_i, \qquad i = 1, 2, \ldots \tag{151}$$

where the noise $\{N_i\}$ is a sequence of independent standard Gaussian random variables, independent of the input process $\{X_i\}$. Let the input statistics be fixed and not dependent on snr.

The finite-horizon version of (151) corresponds to the vector channel (19) with $\boldsymbol{H}$ equal to the identity matrix. Let $\boldsymbol{X}^n = [X_1, \ldots, X_n]^\top$, $\boldsymbol{Y}^n = [Y_1, \ldots, Y_n]^\top$, and $\boldsymbol{N}^n = [N_1, \ldots, N_n]^\top$. The relation (22) between the mutual information and the MMSE holds due to Theorem 2.

Corollary 3: If $\sum_{i=1}^{n} \mathsf{E} X_i^2 < \infty$, then

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\boldsymbol{X}^n; \sqrt{\mathsf{snr}}\boldsymbol{X}^n + \boldsymbol{N}^n) = \frac{1}{2}\sum_{i=1}^{n} \mathsf{mmse}(i, \mathsf{snr}) \quad (152)$$

where

$$\mathsf{mmse}(i, \mathsf{snr}) = \mathsf{E}\left\{(X_i - \mathsf{E}\{X_i \mid \boldsymbol{Y}^n; \mathsf{snr}\})^2\right\} \quad (153)$$

is the noncausal MMSE at time i given the entire observation $\boldsymbol{Y}^n$.

It is also interesting to consider optimal filtering and prediction in this setting. Denote the filtering MMSE as

$$\mathsf{cmmse}(i, \mathsf{snr}) = \mathsf{E}\left\{\left(X_i - \mathsf{E}\left\{X_i \mid \boldsymbol{Y}^i; \mathsf{snr}\right\}\right)^2\right\} \quad (154)$$

and the one-step prediction MMSE as

$$\mathsf{pmmse}(i, \mathsf{snr}) = \mathsf{E}\left\{\left(X_i - \mathsf{E}\left\{X_i \mid \boldsymbol{Y}^{i-1}; \mathsf{snr}\right\}\right)^2\right\}. \quad (155)$$

Theorem 9: The input–output mutual information satisfies

$$\frac{\mathsf{snr}}{2}\sum_{i=1}^{n} \mathsf{cmmse}(i, \mathsf{snr}) \le I(\boldsymbol{X}^n; \boldsymbol{Y}^n) \quad (156a)$$

$$\le \frac{\mathsf{snr}}{2}\sum_{i=1}^{n} \mathsf{pmmse}(i, \mathsf{snr}). \quad (156b)$$

Proof: We study the increase in the mutual information due to an extra sample of observation by considering a conceptual time-incremental channel. Since $\boldsymbol{Y}^i - \boldsymbol{X}^i - X_{i+1} - Y_{i+1}$ is a Markov chain, the mutual information increase is equal to

$$I\left(\boldsymbol{X}^{i+1}; \boldsymbol{Y}^{i+1}\right) - I\left(\boldsymbol{X}^i; \boldsymbol{Y}^i\right) = I\left(X_{i+1}; Y_{i+1} \mid \boldsymbol{Y}^i\right) \quad (157)$$

using an argument similar to the one that leads to (136). This conditional mutual information can be regarded as the input–output mutual information of the simple scalar channel (3) where the input distribution is replaced by the conditional distribution $P_{X_{i+1}|\boldsymbol{Y}^i}$. By Corollary 2

$$\mathsf{E}\left\{\mathsf{var}\left\{X_{i+1}|\boldsymbol{Y}^{i+1}; \mathsf{snr}\right\}\right\} \le \frac{2}{\mathsf{snr}} I\left(X_{i+1}; Y_{i+1} \mid \boldsymbol{Y}^i\right) \quad (158)$$

$$\le \mathsf{E}\left\{\mathsf{var}\left\{X_{i+1}|\boldsymbol{Y}^i; \mathsf{snr}\right\}\right\} \quad (159)$$

or equivalently

$$\frac{\mathsf{snr}}{2}\mathsf{cmmse}(i, \mathsf{snr}) \le I\left(X_{i+1}; Y_{i+1}|\boldsymbol{Y}^i\right) \le \frac{\mathsf{snr}}{2}\mathsf{pmmse}(i, \mathsf{snr}). \quad (160)$$

Finally, we obtain the desired bounds in Theorem 9 summing (160) over n and using (157). □

Corollary 3 and Theorem 9 are still valid if all sides are normalized by n and we then take the limit as $n \to \infty$. As a result, the derivative of the mutual information rate (average mutual information per sample) is equal to half the average noncausal MMSE per symbol. Also, the mutual information rate is sandwiched between half the SNR times the average causal and prediction MMSEs per symbol.

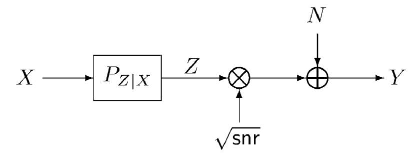

Fig. 6. General additive-noise channel.

B. Discrete Time Versus Continuous Time

In previous sections, the mutual information and the estimation errors have been shown to satisfy similar relations in both discrete- and continuous-time random process models. Indeed, discrete-time processes and continuous-time processes are related fundamentally. For example, discrete-time process can be regarded as the result of integrate-and-dump sampling of the continuous-time one.

It is straightforward to recover the discrete-time results using the continuous-time ones by considering an equivalent of the discrete-time model (151) as a continuous-time one with piecewise-constant input

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, X_{\lceil t \rceil}\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0, \infty). \quad (161)$$

During the time interval $(i-1, i]$, the input to the continuous-time model is equal to the random variable X_i. The samples of $\{Y_t\}$ at natural numbers are sufficient statistics for the input process $\{X_n\}$. Thus, Corollary 3 follows directly from Theorem 6. Analogously, Duncan's theorem can be used to prove Theorem 9 [31].

Conversely, for sufficiently smooth input processes, the continuous-time results (Theorem 6 and Duncan's theorem) can be derived from the discrete-time ones (Corollary 3 and Theorem 9). This can be accomplished by sampling the continuous-time channel outputs and taking the limit of all sides of (156) with vanishing sampling interval. However, in their full generality, the continuous-time results are not a simple extension of the discrete-time ones. A complete analysis of the continuous-time model involves stochastic calculus as developed in Section III.

V. GENERALIZATIONS

A. General Additive-Noise Channel

Consider a general setting where the input is preprocessed arbitrarily before contamination by additive Gaussian noise. The scalar channel setting as depicted in Fig. 6 is first considered for simplicity.

Let X be a random object jointly distributed with a real-valued random variable Z. The channel output is expressed as

$$Y = \sqrt{\mathsf{snr}}\, Z + N \quad (162)$$

where the noise $N \sim \mathcal{N}(0, 1)$ is independent of X and Z. The preprocessor can be regarded as a channel with arbitrary conditional probability distribution $P_{Z|X}$. Since $X - Z - Y$ is a Markov chain

$$I(X; Y) = I(Z; Y) - I(Z; Y \mid X). \quad (163)$$

Note that given (X, Z), the channel output Y is Gaussian. Two applications of Theorem 1 to the right-hand side of (163) give the following.

Theorem 10: Let $X - Z - Y$ be a Markov chain and $Y = \sqrt{\mathsf{snr}}\, Z + N$. If $\mathsf{E}Z^2 < \infty$, then

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X;Y) = \frac{1}{2}\mathsf{E}\left\{(Z - \mathsf{E}\{Z \mid Y;\mathsf{snr}\})^2\right\} - \frac{1}{2}\mathsf{E}\{(Z - \mathsf{E}\{Z \mid Y, X;\mathsf{snr}\})^2\}. \quad (164)$$

The special case of this result for vanishing SNR is given in [4, Theorem 1]. As a simple illustration of Theorem 10, consider a scalar channel where $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ and $P_{Z|X}$ is a Gaussian channel with noise variance σ^2. Then straightforward calculations yield

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{\mathsf{snr}\,\sigma_X^2}{1 + \mathsf{snr}\,\sigma^2}\right) \quad (165)$$

the derivative of which is equal to half the difference of the two MMSEs

$$\frac{1}{2}\left[\frac{\sigma_X^2 + \sigma^2}{1 + \mathsf{snr}\,(\sigma_X^2 + \sigma^2)} - \frac{\sigma^2}{1 + \mathsf{snr}\,\sigma^2}\right]. \quad (166)$$

In the special case where the preprocessor is a deterministic function of the input, e.g., $Z = g(X)$ where $g(\cdot)$ is an arbitrary deterministic mapping, the second term on the right-hand side of (164) vanishes. If, furthermore, $g(\cdot)$ is a one-to-one transformation, then $I(X;Y) = I(g(X);Y)$, and

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X; \sqrt{\mathsf{snr}}\, g(X) + N) = \frac{1}{2}\mathsf{E}\left\{(g(X) - \mathsf{E}\{g(X) \mid Y;\mathsf{snr}\})^2\right\}. \quad (167)$$

Hence, (15) holds verbatim, where the MMSE in this case is defined as the minimum error in estimating $g(X)$. Indeed, the vector channel in Theorem 2 is merely a special case of the vector version of this general result.

One of the many scenarios in which the general result can be useful is the intersymbol interference channel. The input Z_i to the Gaussian channel is the desired symbol X_i corrupted by a function of the previous symbols $(X_{i-1}, X_{i-2}, \ldots)$. Theorem 10 can possibly be used to calculate (or bound) the mutual information given a certain input distribution. Another domain of applications of Theorem 10 is the case of fading channels known or unknown at the receiver, e.g., the channel input $Z = AX$ where A is the multiplicative fading coefficient.

Using similar arguments as in the above, nothing prevents us from generalizing Theorem 6 to a much broader family of models

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, Z_t \mathrm{d}t + \mathrm{d}W_t \quad (168)$$

where $\{Z_t\}$ is a random process jointly distributed with X, and $\{W_t\}$ is a Wiener process independent of X and $\{Z_t\}$.

Theorem 11: As long as the input $\{Z_t\}$ to the channel (168) has finite average power

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}} I\left(X; Y_0^T\right) = \frac{1}{2}\int_0^T \mathsf{E}\left\{\left(Z_t - \mathsf{E}\left\{Z_t \mid Y_0^T;\mathsf{snr}\right\}\right)^2\right\} - \mathsf{E}\left\{\left(Z_t - \mathsf{E}\left\{Z_t \mid Y_0^T, X;\mathsf{snr}\right\}\right)^2\right\}\mathrm{d}t. \quad (169)$$

In case $Z_t = g_t(X)$, where $g_t(\cdot)$ is an arbitrary deterministic one-to-one time-varying mapping, Theorems 6–8 hold verbatim except that the finite-power requirement now applies to $g_t(X)$, and the MMSEs in this case refer to the minimum errors in estimating $g_t(X)$.

B. Gaussian Channels With Feedback

Duncan's theorem (Theorem 7) can be generalized to the continuous-time additive white Gaussian noise channel with feedback [17]

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, Z\left(t, Y_0^t, X\right)\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0,T] \quad (170)$$

where X is any random message (including a random process indexed by t) and the channel input $\{Z_t\}$ is dependent on the message and past output only. The input–output mutual information of this channel with feedback can be expressed as the time average of the optimal filtering mean-square error.

Theorem 12 (Kadota, Zakai, and Ziv [17]): If the power of the input $\{Z_t\}$ to the channel (170) is finite, then

$$I\left(X; Y_0^T\right) = \frac{\mathsf{snr}}{2}\int_0^T \mathsf{E}\Big\{\Big(Z\left(t, Y_0^t, X\right) - \mathsf{E}\left\{Z\left(t, Y_0^t, X\right) | Y_0^t;\mathsf{snr}\right\}\Big)^2\Big\}\mathrm{d}t. \quad (171)$$

Theorem 12 is proved by showing that Duncan's proof of Theorem 7 remains essentially intact as long as the channel input at any time is independent of the future noise process [17]. A new proof can be conceived by considering the time-incremental channel, for which (136) holds literally. Naturally, the counterpart of the discrete-time result (Theorem 9) in the presence of feedback is also feasible.

One is tempted to also generalize the relationship between the mutual information and smoothing error (Theorem 6) to channels with feedback. Unfortunately, it is not possible to construct a meaningful SNR-incremental channel like (123) in this case, since changing the SNR affects not only the amount of Gaussian noise, but also the statistics of the feedback, and consequently the transmitted signal itself. We give two examples to show that in general the derivative of the mutual information with respect to the SNR has no direct connection to the noncausal MMSE, and, in particular

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}} I\left(X; Y_0^T\right) \neq \frac{1}{2}\int_0^T \mathsf{E}\Big\{\Big(Z\left(t, Y_0^t, X\right) - \mathsf{E}\left\{Z\left(t, Y_0^t, X\right) \mid Y_0^T;\mathsf{snr}\right\}\Big)^2\Big\}\mathrm{d}t. \quad (172)$$

Having access to feedback allows the transmitter to determine the SNR as accurately as desired by transmitting known signals and observing the realization of the output for long enough.[9] Once the SNR is known, one can choose a pathological signaling

$$Z\left(t, Y_0^t, X\right) = X/\sqrt{\mathsf{snr}}. \quad (173)$$

[9]The same technique applies to discrete-time channels with feedback. If instead the received signal is in the form of

$$\mathrm{d}Y_t = Z_t\left(t, Y_0^t, X\right)\mathrm{d}t + (1/\sqrt{\mathsf{snr}})\mathrm{d}W_t$$

then the SNR can also be determined by computing the quadratic variation of Y_t during an arbitrarily small interval.

Clearly, the output of channel (170) remains the same regardless of the SNR. Hence, the mutual information has zero derivative, while the MMSE is nonzero. In fact, one can choose to encode the SNR in the channel input in such a way that the derivative of the mutual information is arbitrary (e.g., negative).

The same conclusion can be drawn from an alternative viewpoint by noting that feedback can help to achieve essentially symbol error-free communication at channel capacity by using a signaling specially tailored for the SNR, e.g., capacity-achieving error-control codes. More interesting is the variable-duration modulation scheme of Turin [58] for the infinite-bandwidth continuous-time Gaussian channel, where the capacity-achieving input is an explicit deterministic function of the message and the feedback. From this scheme, we can derive a suboptimal noncausal estimator of the channel input by appending the encoder at the output of the decoder. Since arbitrarily low block error rate can be achieved by the coding scheme of [58] and the channel input has bounded power, the smoothing MMSE achieved by the suboptimal noncausal estimator can be made as small as desired. On the other hand, achieving channel capacity requires that the mutual information be nonnegligible.

Note that a fundamental proviso for our mutual information–MMSE relationship is that the input distribution not be allowed to depend on SNR. However, in general, feedback removes such restrictions.

C. Generalization to Vector Models

Just as Theorem 1 obtained under a scalar model has its counterpart (Theorem 2) under a vector model, all the results in Sections III and IV can be generalized to vector models, under either discrete- or the continuous-time setting. For example, the vector continuous-time model takes the form of

$$\mathrm{d}\boldsymbol{Y}_t = \sqrt{\mathsf{snr}}\,\boldsymbol{X}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t \tag{174}$$

where $\{\boldsymbol{W}_t\}$ is an m-dimensional Wiener process, and $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$ are m-dimensional random processes. Theorem 6 holds literally, while the mutual information rate, estimation errors, and power are now defined with respect to the vector signals and their Euclidean norms. In fact, Duncan's theorem was originally given in vector form [15]. It should be noted that the incremental-channel devices are directly applicable to the vector models.

In view of the above generalizations, the discrete- and continuous-time results in Sections V-A and -B also extend straightforwardly to vector models.

Furthermore, colored additive Gaussian noise can be treated by first filtering the observation to whiten the noise and recover the canonical model of the form (162).

D. Complex-Valued Channels

The results in the discrete-time regime (Theorems 1–5 and Corollaries 1–3) hold verbatim for complex-valued channel and signaling if the noise samples are i.i.d. circularly symmetric complex Gaussian, whose real and imaginary components have unit variance. In particular, the factor of $1/2$ in (15), (22), (152), and (156) remains intact. However, with the more common definition of snr in complex-valued channels where the complex noise has real and imaginary components with variance $1/2$ each, the factor of $1/2$ in the formulas disappears.

The above principle holds also under continuous-time models as long as the complex-valued Wiener process is appropriately defined. This is straightforward by noting that, in general, complex-valued models can be regarded as two independent uses of the real-valued ones (with possibly correlated inputs in the two uses).

VI. New Representation of Information Measures

The relationship between mutual information and MMSE enables other information measures such as entropy and divergence to be expressed as a function of MMSE as well.

Consider a discrete random variable X. Assume $N \sim \mathcal{N}(0,1)$ independent of the input throughout this section. The mutual information between X and its observation through a Gaussian channel converges to the entropy of X as the SNR of the channel goes to infinity.

Lemma 6: For every discrete real-valued random variable X

$$H(X) = \lim_{\mathsf{snr}\to\infty} I(X; \sqrt{\mathsf{snr}}\,X + N). \tag{175}$$

Proof: See Appendix VII. □

Note that if $H(X)$ is infinity then the mutual information in (175) also increases without bound as $\mathsf{snr} \to \infty$. Moreover, the result holds if X is subject to an arbitrary one-to-one mapping $g(\cdot)$ before going through the channel. In view of (167) and (175), the following theorem is immediate.

Theorem 13: For any discrete random variable X taking values in $\mathcal{A}$, the entropy of X is given by (in nats)

$$H(X) = \frac{1}{2}\int_0^\infty \mathsf{E}\left\{\left(g(X) - \mathsf{E}\{g(X) \mid \sqrt{\mathsf{snr}}\,g(X) + N\}\right)^2\right\}\mathrm{d}\mathsf{snr} \tag{176}$$

for any one-to-one mapping $g : \mathcal{A} \to \mathbb{R}$.

It is interesting to note that the integral on the right-hand side of (176) is not dependent on the choice of $g(\cdot)$, which is not evident from estimation-theoretic properties alone.

The "non-Gaussianness" of a random variable (divergence between its distribution and a Gaussian distribution with the same mean and variance) and, thus, the differential entropy can also be written in terms of MMSE. To that end, we need the following auxiliary result.

Lemma 7: Let X be any real-valued random variable and X' be Gaussian with the same mean and variance as X, i.e., $X' \sim \mathcal{N}\left(\mathsf{E}X, \sigma_X^2\right)$. Let Y and Y' be the output of the channel (3) with X and X' as the input, respectively. Then

$$\mathsf{D}(P_X \| P_{X'}) = \lim_{\mathsf{snr}\to\infty} \mathsf{D}(P_Y \| P_{Y'}). \tag{177}$$

Proof: By monotone convergence and the fact that data processing reduces divergence. □

Note that in case the divergence between P_X and $P_{X'}$ is infinity, the divergence between P_Y and $P_{Y'}$ also increases without bound. Since

$$\mathsf{D}(P_Y \| P_{Y'}) = I(X';Y') - I(X;Y) \tag{178}$$

the following result is straightforward by Theorem 1.

Theorem 14: For every random variable X with $\sigma_X^2 < \infty$, its non-Gaussianness is given by

$$D_X = \mathsf{D}\left(P_X \| \mathcal{N}\left(\mathsf{E}X, \sigma_X^2\right)\right) \tag{179}$$

$$= \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\mathsf{snr}\,\sigma_X^2} - \mathsf{mmse}(X|\sqrt{\mathsf{snr}}\,X+N)\,\mathrm{d}\,\mathsf{snr}. \tag{180}$$

Note that the integrand in (179) is always positive since for the same variance, Gaussian inputs maximize the MMSE. Also, Theorem 14 holds even if the divergence is infinity, for example, in the case that X is a discrete random variable. In light of Theorem 14, the differential entropy of X can be expressed as

$$h(X) = \frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - D_X \tag{181}$$

$$= \frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\mathsf{snr}\,\sigma_X^2} - \mathsf{mmse}(X|\sqrt{\mathsf{snr}}\,X+N)\,\mathrm{d}\,\mathsf{snr}. \tag{182}$$

According to (179), $\gamma_X = e^{-D_X}$ is a parameter that measures the difficulty of estimating X when observed in Gaussian noise across the full range of SNRs. Note that $0 \le \gamma_X \le 1$ with the upper bound attained when X is Gaussian, and the lower bound attained when X is discrete. Adding independent random variables results in a random variable that is harder to estimate in the sense of the following inequality:

$$\alpha\gamma_{X_1}^2 + (1-\alpha)\gamma_{X_2}^2 \le \gamma_{X_1+X_2}^2 \tag{183}$$

where X_1 and X_2 are independent random variables and α is the ratio of the variance of X_1 to the sum of the variances of X_1 and X_2. Of course, (183) is nothing but Shannon's entropy power inequality [25]. It would be interesting to see if (183) can be proven from estimation-theoretic principles.

Another observation is that Theorem 10 provides a new means of representing the mutual information between an arbitrary random variable X and a real-valued random variable Z

$$I(X;Z) = \frac{1}{2}\int_0^\infty \mathsf{E}\Big\{(\mathsf{E}\{Z \mid \sqrt{\mathsf{snr}}\,Z+N, X\})^2 - (\mathsf{E}\{Z \mid \sqrt{\mathsf{snr}}\,Z+N\})^2\Big\}\mathrm{d}\,\mathsf{snr}. \tag{184}$$

An arbitrary discrete-valued Z can be handled as in Theorem 13 by means of an adequate one-to-one mapping.

The preceding results can be generalized to continuous-time models and vector channels. It is remarkable that the entropy, differential entropy, divergence, and mutual information in fairly general settings admit expressions in pure estimation-theoretic quantities. It remains to be seen whether such representations lead to new insights and applications.

VII. Conclusion

This paper reveals that the input–output mutual information and the (noncausal) MMSE in estimating the input given the output determine each other by a simple formula under both discrete- and continuous-time, scalar, and vector Gaussian channel models. A consequence of this relationship is the coupling of the MMSEs achievable by smoothing and filtering with arbitrary signals corrupted by Gaussian noise. Moreover, new expressions in terms of MMSE are found for information measures such as entropy and divergence.

The idea of incremental channels is the underlying basis for the most streamlined proof of the main results and for their interpretation. The white Gaussian nature of the noise is key to this approach: 1) the sum of independent Gaussian variates is Gaussian; and 2) the Wiener process has independent increments. In fact, the relationship between the mutual information and the noncausal estimation error holds in even more general settings of Gaussian channels. In a follow-up to this work, Zakai has recently extended formula (1) to the abstract Wiener space [44], which generalizes the classical m-dimensional Wiener process.

The incremental-channel technique in this paper is relevant for an entire family of channels the noise of which has independent increments, i.e., that is characterized by Lévy processes [59]. A particularly interesting case, reported in [60], is the Poisson channel, where the corresponding mutual information–estimation error relationship involves an error measure quite different from mean-square error.

Applications of the relationships revealed in this paper are abundant. In addition to the application in [30] to multiuser channels, [38] shows applications to key results in EXIT charts for the analysis of sparse-graph codes. Other applications as well as counterparts to non-Gaussian channels will be published in the near future. In all, the relations shown in this paper illuminate intimate connections between information theory and estimation theory.

Appendix I
Verification of (15): Binary Input

Proof: From (17) and (18), it can be checked that

$$2\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}I(\mathsf{snr}) - \mathsf{mmse}(\mathsf{snr}) = 1 - \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\left(1-\frac{y}{\sqrt{\mathsf{snr}}}\right)\tanh(\mathsf{snr}-\sqrt{\mathsf{snr}}\,y)\mathrm{d}y \tag{185}$$

$$= 1 - \frac{1}{\sqrt{\mathsf{snr}}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\sqrt{\mathsf{snr}})^2}z\tanh(\sqrt{\mathsf{snr}}\,z)\mathrm{d}z \tag{186}$$

where from (185) to (186) $\sqrt{\mathsf{snr}}-y$ is replaced by z. The integral in (186) can be regarded as the expectation of $Z\tanh(\sqrt{\mathsf{snr}}Z)$ where $Z \sim \mathcal{N}(\sqrt{\mathsf{snr}},1)$. The expectation remains the same if Z is replaced by $Z' \sim \mathcal{N}(-\sqrt{\mathsf{snr}},1)$ due to symmetry. Hence, the integral can be rewritten by averaging over the two cases as

$$\frac{1}{2}\int_{-\infty}^{\infty}\left[e^{-\frac{1}{2}(z-\sqrt{\mathsf{snr}})^2}+e^{-\frac{1}{2}(z+\sqrt{\mathsf{snr}})^2}\right]z\tanh(\sqrt{\mathsf{snr}}\,z)\frac{\mathrm{d}z}{\sqrt{2\pi}}$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\left[e^{-\frac{1}{2}(z-\sqrt{\mathsf{snr}})^2}-e^{-\frac{1}{2}(z+\sqrt{\mathsf{snr}})^2}\right]z\frac{\mathrm{d}z}{\sqrt{2\pi}} \tag{187}$$

$$= \frac{1}{2}(\mathsf{E}Z - \mathsf{E}Z') \quad (188)$$
$$= \sqrt{\mathsf{snr}}. \quad (189)$$

Therefore, (186) vanishes by (189), and (15) holds. □

APPENDIX II
PROOF OF LEMMA 1

By (64), the mutual information admits the following decomposition:

$$I(Y;Z) = \mathsf{D}(P_{Y|Z}\|P_{Y'}|P_Z) - \mathsf{D}(P_Y\|P_{Y'}) \quad (190)$$

where $Y' \sim \mathcal{N}\left(\mathsf{E}Y, \sigma_Y^2\right)$. Let the variance of Z be denoted by v. The first term on the right-hand side of (190) is equal to a divergence between two Gaussian distributions, which is found as

$$\frac{1}{2}\log(1+\delta v) = \frac{\delta v}{2} + o(\delta) \quad (191)$$

by using the general formula

$$\begin{aligned}&\mathsf{D}\left(\mathcal{N}\left(m_1,\sigma_1^2\right)\|\mathcal{N}\left(m_0,\sigma_0^2\right)\right)\\ &\quad= \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}\left(\frac{(m_1-m_0)^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\right)\log e. \quad (192)\end{aligned}$$

It suffices then to show that

$$\mathsf{D}(P_Y\|P_{Y'}) = \mathsf{E}\left\{\log\frac{p_Y(Y)}{p_{Y'}(Y)}\right\} = o(\delta) \quad (193)$$

which is straightforward to check by plugging in the density functions:

$$\begin{aligned}
&\log\frac{p_Y(y)}{p_{Y'}(y)}\\
&= \log\left[\frac{1}{\sqrt{2\pi}}\mathsf{E}\left\{\exp\left[-\frac{1}{2}(y-\sqrt{\delta}Z)^2\right]\right\}\right]\\
&\quad - \log\left[\frac{1}{\sqrt{2\pi(\delta v+1)}}\exp\left[-\frac{(y-\mathsf{E}Y)^2}{2(\delta v+1)}\right]\right] \quad (194)\\
&= \log\mathsf{E}\left\{\exp\left[\frac{(y-\sqrt{\delta}\mathsf{E}Z)^2}{2(\delta v+1)} - \frac{1}{2}(y-\sqrt{\delta}Z)^2\right]\right\}\\
&\quad + \frac{1}{2}\log(1+\delta v) \quad (195)\\
&= \log\mathsf{E}\Big\{1+\sqrt{\delta}y(Z-\mathsf{E}Z) + \frac{\delta}{2}\big(y^2(Z-\mathsf{E}Z)^2 - vy^2\\
&\qquad - Z^2 + (\mathsf{E}Z)^2\big) + o(\delta)\Big\} + \frac{1}{2}\log(1+\delta v) \quad (196)\\
&= \log\left(1-\frac{\delta v}{2}\right) + \frac{1}{2}\log(1+\delta v) + o(\delta) \quad (197)\\
&= o(\delta). \quad (198)
\end{aligned}$$

The limit and the expectation can be exchanged to obtain (197) as long as $\mathsf{E}\,Z^2 < \infty$ due to the Lebesgue convergence theorem. □

It is interesting to note that the proof relies on the fact that the divergence between the output distributions of a Gaussian channel under different input distributions is sublinear in the SNR when the noise dominates.

APPENDIX III
A FOURTH PROOF OF THEOREM 1

For simplicity, it is assumed that the order of expectation and derivative can be exchanged freely. A rigorous proof is relegated to Appendix IV, where every such assumption is validated in the more general vector model.

Let $q_i(y;\mathsf{snr})$ be defined as in (87). It can be checked that for all i

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,\mathsf{snr}}q_i(y;\mathsf{snr}) &= \frac{1}{2\sqrt{\mathsf{snr}}}yq_{i+1}(y;\mathsf{snr}) - \frac{1}{2}q_{i+2}(y;\mathsf{snr}) \quad (199)\\ &= -\frac{1}{2\sqrt{\mathsf{snr}}}\frac{\mathrm{d}}{\mathrm{d}y}q_{i+1}(y;\mathsf{snr}). \quad (200)\end{aligned}$$

The derivative of the mutual information, expressed as (89), can be obtained as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}I(\mathsf{snr}) &= -\int[\log q_0(y;\mathsf{snr})+1]\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}q_0(y;\mathsf{snr})\mathrm{d}y \quad (201)\\
&= \frac{1}{2\sqrt{\mathsf{snr}}}\int\log q_0(y;\mathsf{snr})\frac{\mathrm{d}}{\mathrm{d}y}q_1(y;\mathsf{snr})\mathrm{d}y \quad (202)\\
&= -\frac{1}{2\sqrt{\mathsf{snr}}}\int\frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\frac{\mathrm{d}}{\mathrm{d}y}q_0(y;\mathsf{snr})\mathrm{d}y \quad (203)\\
&= \frac{1}{2\sqrt{\mathsf{snr}}}\int q_1(y;\mathsf{snr})\left[y - \sqrt{\mathsf{snr}}\frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\right]\mathrm{d}y \quad (204)
\end{aligned}$$

where (203) follows by integrating by parts. Noting that the fraction in (204) is exactly the conditional mean estimate (cf. (88))

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}I(\mathsf{snr}) &= \frac{1}{2\sqrt{\mathsf{snr}}}\mathsf{E}\Big\{\mathsf{E}\{X\mid Y;\mathsf{snr}\}\\
&\qquad\times[Y-\sqrt{\mathsf{snr}}\mathsf{E}\{X\mid Y;\mathsf{snr}\}]\Big\} \quad (205)\\
&= \frac{1}{2}\mathsf{E}\left\{\frac{XY}{\sqrt{\mathsf{snr}}} - (\mathsf{E}\{X\mid Y;\mathsf{snr}\})^2\right\} \quad (206)\\
&= \frac{1}{2}\mathsf{E}\left\{(X-\mathsf{E}\{X\mid Y;\mathsf{snr}\})^2\right\} \quad (207)\\
&= \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \quad (208)
\end{aligned}$$

□

APPENDIX IV
PROOF OF THEOREM 2

Proof: It suffices to prove the theorem assuming $\boldsymbol{H} = \boldsymbol{I}$ since one can always regard $\boldsymbol{HX}$ as the input. The vector channel (19) has a Gaussian conditional density (20). The unconditional density of the channel output is given by (55), which is strictly positive for all $\boldsymbol{y}$. The mutual information can be written as (cf. (89))

$$\begin{aligned}I(\mathsf{snr}) &= -\frac{L}{2}\log(2\pi e)\\ &\quad - \int p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\mathrm{d}\boldsymbol{y}. \quad (209)\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{dsnr}} I(\mathsf{snr}) = -\int [\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1] \times \frac{\mathrm{d}}{\mathrm{dsnr}} p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) \mathrm{d}\boldsymbol{y} \tag{210}$$

$$= -\int [\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1] \times \mathsf{E}\left\{ \frac{\mathrm{d}}{\mathrm{dsnr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}) \right\} \mathrm{d}\boldsymbol{y} \tag{211}$$

where the derivative penetrates the integral in (210) by the Lebesgue convergence theorem, and the order of taking the derivative and expectation in (211) can be exchanged by Lemma 8, which is shown later in this appendix. It can be checked that (cf. (199) and (200))

$$\frac{\mathrm{d}}{\mathrm{dsnr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) = \frac{1}{2\sqrt{\mathsf{snr}}} \boldsymbol{x}^\top (\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{x}) p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) \tag{212}$$

$$= -\frac{1}{2\sqrt{\mathsf{snr}}} \boldsymbol{x}^\top \nabla p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}). \tag{213}$$

Using (213), the right-hand side of (211) can be written as

$$\frac{1}{2\sqrt{\mathsf{snr}}} \mathsf{E}\left\{ \boldsymbol{X}^\top \int [\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1] \times \nabla p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}) \mathrm{d}\boldsymbol{y} \right\}. \tag{214}$$

The integral in (214) can be carried out by parts to obtain

$$-\int p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}) \nabla [\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1] \mathrm{d}\boldsymbol{y} \tag{215}$$

since for all $\boldsymbol{x}$, as $\|\boldsymbol{y}\| \to \infty$

$$p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) [\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1] \to 0. \tag{216}$$

Hence, the expectation in (214) can be further evaluated as

$$-\int \mathsf{E}\left\{ \boldsymbol{X}^\top \frac{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})}{p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})} \right\} \nabla p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) \mathrm{d}\boldsymbol{y} \tag{217}$$

where we have changed the order of the expectation with respect to $\boldsymbol{X}$ and the integral (i.e., expectation with respect to $\boldsymbol{Y}$). By (213) and Lemma 9 (shown below in this appendix), (217) can be written as

$$\int \mathsf{E}\{\boldsymbol{X}^\top \mid \boldsymbol{Y} = \boldsymbol{y}; \mathsf{snr}\} \times \mathsf{E}\{(\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{X}) p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\} \mathrm{d}\boldsymbol{y}. \tag{218}$$

Therefore, (211) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{dsnr}} I(\mathsf{snr}) = \frac{1}{2\sqrt{\mathsf{snr}}} \int \mathsf{E}\{\boldsymbol{X}^\top \mid \boldsymbol{Y} = \boldsymbol{y}; \mathsf{snr}\} \times \mathsf{E}\{\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{X} \big| \boldsymbol{Y} = \boldsymbol{y}; \mathsf{snr}\} p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) \mathrm{d}\boldsymbol{y} \tag{219}$$

$$= \mathsf{E}\left\{ \mathsf{E}\{\boldsymbol{X}^\top \mid \boldsymbol{Y}; \mathsf{snr}\} \times \mathsf{E}\left\{ \frac{\boldsymbol{Y}}{2\sqrt{\mathsf{snr}}} - \frac{1}{2}\boldsymbol{X} \Big| \boldsymbol{Y}; \mathsf{snr} \right\} \right\} \tag{220}$$

$$= \mathsf{E}\left\{ \frac{1}{2}\|\boldsymbol{X}\|^2 - \frac{1}{2}\|\mathsf{E}\{\boldsymbol{X} \mid \boldsymbol{Y}; \mathsf{snr}\}\|^2 \right\} \tag{221}$$

$$= \frac{1}{2}\mathsf{E}\{\|\boldsymbol{X} - \mathsf{E}\{\boldsymbol{X} \mid \boldsymbol{Y}; \mathsf{snr}\}\|^2\}. \tag{222}$$

Hence the proof of Theorem 2. □

The following two lemmas were needed to justify the exchange of derivatives and expectation with respect to $P_{\boldsymbol{X}}$ in the above proof.

Lemma 8: If $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$, then

$$\frac{\mathrm{d}}{\mathrm{dsnr}} \mathsf{E}\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\} = \mathsf{E}\left\{ \frac{\mathrm{d}}{\mathrm{dsnr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}) \right\}. \tag{223}$$

Proof: Let

$$f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = \frac{1}{\delta} \left[p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr} + \delta) - p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}) \right] \tag{224}$$

and

$$f(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = \frac{\mathrm{d}}{\mathrm{dsnr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}). \tag{225}$$

Then, $\forall\, \boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}$

$$f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) \to f(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr})$$

as $\delta \to 0$. Lemma 8 is equivalent to

$$\lim_{\delta \to 0} \int f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) P_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) = \int f(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) P_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}). \tag{226}$$

Suppose we can show that for every $\delta, \boldsymbol{x}, \boldsymbol{y}$ and snr

$$|f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr})| < \|\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}} |\boldsymbol{y}^\top \boldsymbol{x}|. \tag{227}$$

Then (226) holds by the Lebesgue convergence theorem since the right-hand side of (227) is integrable with respect to $P_{\boldsymbol{X}}$ by the assumption in the lemma. Note that

$$f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = (2\pi)^{-\frac{L}{2}} \frac{1}{\delta} \left(\exp\left[-\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr} + \delta}\boldsymbol{x}\|^2 \right] - \exp\left[-\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{x}\|^2 \right] \right). \tag{228}$$

If

$$\frac{1}{\delta} \le \|\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}} |\boldsymbol{y}^\top \boldsymbol{x}| \tag{229}$$

then (227) holds trivially. Otherwise

$$|f_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr})| < \frac{1}{\delta} \left| \exp\left[\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr} + \delta}\boldsymbol{x}\|^2 \right] - 1 \right| \tag{230}$$

$$< \frac{1}{2\delta} [\exp|\delta\|\boldsymbol{x}\|^2 - (\sqrt{\mathsf{snr} + \delta} - \sqrt{\mathsf{snr}})\boldsymbol{y}^\top \boldsymbol{x}| - 1] \tag{231}$$

$$< \frac{1}{2\delta} \left[\exp\left[\delta \left(\|\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}} \|\boldsymbol{y}^\top \boldsymbol{x}\| \right) \right] - 1 \right]. \tag{232}$$

The inequality (227) holds for all $\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}$ due to the fact that

$$e^t - 1 < 2t, \qquad \forall\, 0 \le t < 1. \tag{233}$$

□

Lemma 9: If $\mathsf{E}\boldsymbol{X}$ exists, then for $i = 1, \dots, L$

$$\frac{\partial}{\partial y_i} \mathsf{E}\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr})\} = \mathsf{E}\left\{ \frac{\partial}{\partial y_i} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}) \right\}. \tag{234}$$

Proof: The proof is similar to that of Lemma 8. Let

$$g_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = \frac{1}{\delta}\big[p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}+\delta\boldsymbol{e}_i|\boldsymbol{X};\mathsf{snr}) - p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\big] \quad (235)$$

where $\boldsymbol{e}_i$ is a vector composed of zeros except the ith entry, which is 1. Then, $\forall\, \boldsymbol{x},\boldsymbol{y},\mathsf{snr}$

$$\lim_{\delta\to 0} g_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = \frac{\partial}{\partial y_i} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}). \quad (236)$$

We show that

$$|g_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})| < |y_i| + 1 + \sqrt{\mathsf{snr}}|x_i| \quad (237)$$

so that (234) holds by the Lebesgue convergence theorem (cf. (226)). Note that

$$g_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = (2\pi)^{-\frac{L}{2}}\frac{1}{\delta}\left(\exp\left[-\frac{1}{2}\|\boldsymbol{y}+\delta\boldsymbol{e}_i-\sqrt{\mathsf{snr}}\boldsymbol{x}\|^2\right] - \exp\left[-\frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}}\boldsymbol{x}\|^2\right]\right). \quad (238)$$

If

$$\frac{1}{\delta} \le |y_i| + 1 + \frac{1}{\sqrt{\mathsf{snr}}}|x_i| \quad (239)$$

then (237) holds trivially. Otherwise

$$\begin{aligned} |g_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})| &< \frac{1}{2\delta}\left(\exp\left|\frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}}\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y}+\delta\boldsymbol{e}_i-\sqrt{\mathsf{snr}}\boldsymbol{x}\|^2\right| - 1\right) \quad (240)\\ &= \frac{1}{2\delta}\left(\exp\left|\frac{\delta}{2}(2y_i+\delta-2\sqrt{\mathsf{snr}}x_i)\right| - 1\right) \quad (241)\end{aligned}$$

and (237) holds by (233). □

APPENDIX V
VERIFICATION OF (111): RANDOM TELEGRAPH INPUT

Let $\xi = -\frac{2\nu}{\mathsf{snr}}$ and define

$$f(i,j) = \int_1^\infty u^{\frac{i}{2}}(u-1)^{\frac{j}{2}}e^{\xi u}\mathrm{d}u. \quad (242)$$

It can be checked that

$$f(i,j) = f(i+2,j) - f(i,j+2) \quad (243)$$

$$\frac{\mathrm{d}}{\mathrm{d}\xi}f(i,j) = f(i+2,j) \quad (244)$$

$$-\xi f(i,j) = \frac{i}{2}f(i-2,j) + \frac{j}{2}f(i,j-2) \quad (245)$$

where verifying (245) entails integration by parts. Then (113) can be rewritten as

$$\mathsf{cmmse}(\mathsf{snr}) = f(-1,-1)/f(1,-1) \quad (246)$$

and hence,

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}[\mathsf{snr}\cdot\mathsf{cmmse}(\mathsf{snr})] = [f(-1,-1)f(1,-1) - \xi f^2(1,-1) + \xi f(-1,-1)f(3,-1)]/f^2(1,-1). \quad (247)$$

With the change of variables $t=(1-x^2)^{-1}$ and $u=(1-y^2)^{-1}$, (114) can also be rewritten as

$$\mathsf{mmse}(\mathsf{snr}) = f^{-2}(1,-1)\int_1^\infty\int_1^\infty \frac{e^{(t+u)\xi}}{t+u-1}t^{\frac{1}{2}}u^{\frac{1}{2}}(t-1)^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}}\mathrm{d}t\mathrm{d}u. \quad (248)$$

The denominator in (248) prevents the double integral from being separated. This can be circumvented by taking derivative with respect to ξ. Noting that

$$e^\xi\frac{\mathrm{d}}{\mathrm{d}\xi}[e^{-\xi}f^2(1,-1)\mathsf{mmse}(\mathsf{snr})] = f^2(1,-1) \quad (249)$$

the identity (111) is equivalent to

$$e^\xi\frac{\mathrm{d}}{\mathrm{d}\xi}\left[e^{-\xi}\big(f(-1,-1)f(1,-1) - \xi f^2(1,-1) + \xi f(-1,-1)f(3,-1)\big)\right] = f^2(1,-1) \quad (250)$$

since both sides of (111) tend to 0 as $\xi\to-\infty$. With the help of (243)–(245), verifying (250) is a matter of algebra.

APPENDIX VI
PROOF OF LEMMA 5

Lemma 5 can be regarded as a consequence of Duncan's theorem. The mutual information can be expressed as a time integral of the causal MMSE

$$I\left(Z_0^T;Y_0^T\right) = \frac{\delta}{2}\int_0^T \mathsf{E}\left(Z_t - \mathsf{E}\left\{Z_t \mid Y_0^t;\delta\right\}\right)^2\mathrm{d}t. \quad (251)$$

As the SNR $\delta\to 0$, the observation Y_0^T becomes inconsequential in estimating the input signal. Indeed, the causal MMSE estimate converges to the unconditional mean in mean-square sense

$$\mathsf{E}\left\{Z_t \mid Y_0^t;\delta\right\} \to \mathsf{E}Z_t. \quad (252)$$

Putting (251) and (252) together proves Lemma 5.

In parallel with the proof of Lemma 1, another reasoning of Lemma 5 from first principles without invoking Duncan's theorem is presented in the following. In fact, Lemma 5 is established first in this way so that a more intuitive proof of Duncan's theorem is given in Section III-D using the idea of time-incremental channels.

Proof: [Lemma 5] By definition (98), the mutual information is the expectation of the logarithm of the Radon–Nikodym derivative (99), which can be obtained by the chain rule as

$$\Phi = \frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_Y\mathrm{d}\mu_Z} = \frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}}\left(\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mu_W}\right)^{-1}. \quad (253)$$

First assume that $\{Z_t\}$ is a bounded uniformly stepwise process, i.e., there exists a finite subdivision of $[0,T]$

$$0=t_0<t_1<\cdots<t_n=T$$

and a finite constant M such that

$$Z_t(\omega)=Z_{t_i}(\omega),\qquad t\in[t_i,t_{i+1}),\ i=0,\dots,n-1 \tag{254}$$

and $Z_t(\omega)<M, \forall\, t\in[0,T]$. Let $\boldsymbol{Z}=[Z_{t_0},\dots,Z_{t_n}]$, $\boldsymbol{Y}=[Y_{t_0},\dots,Y_{t_n}]$, and $\boldsymbol{W}=[W_{t_0},\dots,W_{t_n}]$ be $(n+1)$-dimensional vectors formed by the samples of the random processes. Then, the input–output conditional density is Gaussian

$$p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})=\prod_{i=0}^{n-1}\frac{1}{\sqrt{2\pi(t_{i+1}-t_i)}}\times\exp\left[-\frac{(y_{i+1}-y_i-\sqrt{\delta}z_i(t_{i+1}-t_i))^2}{2(t_{i+1}-t_i)}\right]. \tag{255}$$

Easily

$$\frac{p_{\boldsymbol{YZ}}(\boldsymbol{b},\boldsymbol{z})}{p_{\boldsymbol{WZ}}(\boldsymbol{b},\boldsymbol{z})}=\frac{p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{b}|\boldsymbol{z})}{p_{\boldsymbol{W}}(\boldsymbol{b})} \tag{256}$$

$$=\exp\left[\sqrt{\delta}\sum_{i=0}^{n-1}z_i(b_{i+1}-b_i)-\frac{\delta}{2}\sum_{i=0}^{n-1}z_i^2(t_{i+1}-t_i)\right]. \tag{257}$$

Thus the Radon–Nikodym derivative can be established as

$$\frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}}=\exp\left[\sqrt{\delta}\int_0^T Z_t\mathrm{d}W_t-\frac{\delta}{2}\int_0^T Z_t^2\mathrm{d}t\right] \tag{258}$$

using the finite-dimensional likelihood ratios (257). It is clear that $\mu_{YZ}\ll\mu_{WZ}$.

For the case of a general finite-power process (not necessarily bounded) $\{Z_t\}$, a sequence of bounded uniformly stepwise processes which converge to the $\{Z_t\}$ in $L^2(\mathrm{d}t\mathrm{d}P)$ can be obtained. The Radon–Nikodym derivative (258) of the sequence of processes also converges. Absolute continuity is preserved. Therefore, (258) holds for all such processes $\{Z_t\}$.[10]

The derivative (258) can be rewritten as

$$\frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}}=1+\sqrt{\delta}\int_0^T Z_t\mathrm{d}W_t+\frac{\delta}{2}\left[\left(\int_0^T Z_t\mathrm{d}W_t\right)^2-\int_0^T Z_t^2\mathrm{d}t\right]+o(\delta). \tag{259}$$

By the independence of the processes $\{W_t\}$ and $\{Z_t\}$, the measure $\mu_{WZ}=\mu_W\mu_Z$. Thus, integrating on the measure μ_Z gives

$$\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mu_W}=1+\sqrt{\delta}\int_0^T \mathsf{E}Z_t\mathrm{d}W_t+\frac{\delta}{2}\left[\mathsf{E}_{\mu_Z}\left(\int_0^T Z_t\mathrm{d}W_t\right)^2-\int_0^T \mathsf{E}Z_t^2\mathrm{d}t\right]+o(\delta). \tag{260}$$

[10]A shortcut to the proof of (258) is by the Girsanov theorem [47].

Using (259), (260), and the chain rule (253), the Radon–Nikodym derivative Φ exists and is given by

$$\begin{aligned}\Phi=1&+\sqrt{\delta}\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\\&+\frac{\delta}{2}\Bigg[\left(\int_0^T Z_t\mathrm{d}W_t\right)^2-\int_0^T Z_t^2\mathrm{d}t\\&\quad-2\int_0^T \mathsf{E}Z_t\mathrm{d}W_t\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\\&\quad-\mathsf{E}_{\mu_Z}\left(\int_0^T Z_t\mathrm{d}W_t\right)^2+\int_0^T\mathsf{E}Z_t^2\mathrm{d}t\Bigg]+o(\delta)\end{aligned} \tag{261}$$

$$\begin{aligned}=1&+\sqrt{\delta}\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\\&+\frac{\delta}{2}\Bigg[\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\right)^2-\mathsf{E}_{\mu_Z}\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\right)^2\\&\quad-\int_0^T Z_t^2-\mathsf{E}Z_t^2\mathrm{d}t\Bigg]+o(\delta).\end{aligned} \tag{262}$$

Note that the mutual information is an expectation with respect to the measure μ_{YZ}. It can be written as

$$I\left(Z_0^T;Y_0^T\right)=\int\log\Phi'\mathrm{d}\mu_{YZ} \tag{263}$$

where Φ' is obtained from Φ (262) by substituting all occurrences of $\mathrm{d}W_t$ by $\mathrm{d}Y_t=\sqrt{\delta}Z_t+\mathrm{d}W_t$

$$\begin{aligned}\Phi'=1&+\sqrt{\delta}\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}Y_t\\&+\frac{\delta}{2}\Bigg[\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}Y_t\right)^2\\&\quad-\mathsf{E}_{\mu_Z}\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}Y_t\right)^2\\&\quad-\int_0^T Z_t^2-\mathsf{E}Z_t^2\mathrm{d}t\Bigg]+o(\delta)\end{aligned} \tag{264}$$

$$\begin{aligned}=1&+\sqrt{\delta}\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\\&+\frac{\delta}{2}\Bigg[\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\right)^2\\&\quad-\mathsf{E}_{\mu_Z}\left(\int_0^T Z_t-\mathsf{E}Z_t\mathrm{d}W_t\right)^2\\&\quad+\int_0^T(Z_t-\mathsf{E}Z_t)^2\mathrm{d}t\\&\quad+\int_0^T\mathsf{E}(Z_t-\mathsf{E}Z_t)^2\mathrm{d}t\Bigg]+o(\delta)\end{aligned} \tag{265}$$

$$\begin{aligned}=1&+\sqrt{\delta}\int_0^T\tilde{Z}_t\mathrm{d}W_t\\&+\frac{\delta}{2}\Bigg[\left(\int_0^T\tilde{Z}_t\mathrm{d}W_t\right)^2-\mathsf{E}_{\mu_Z}\left(\int_0^T\tilde{Z}_t\mathrm{d}W_t\right)^2\\&\quad+\int_0^T\tilde{Z}_t^2\mathrm{d}t+\int_0^T\mathsf{E}\tilde{Z}_t^2\mathrm{d}t\Bigg]+o(\delta)\end{aligned} \tag{266}$$

where $\tilde{Z}_t = Z_t - \mathsf{E}Z_t$. Hence,

$$\log \Phi' = \sqrt{\delta}\int_0^T \tilde{Z}_t \mathrm{d}W_t + \frac{\delta}{2}\left[-\mathsf{E}_{\mu_Z}\left(\int_0^T \tilde{Z}_t \mathrm{d}W_t\right)^2 + \int_0^T \tilde{Z}_t^2 \mathrm{d}t + \int_0^T \mathsf{E}\tilde{Z}_t^2 \mathrm{d}t\right] + o(\delta). \quad (267)$$

Therefore, the mutual information is

$$\begin{aligned}\mathsf{E}\log\Phi' &= \frac{\delta}{2}\left[2\int_0^T \mathsf{E}\tilde{Z}_t^2\mathrm{d}t - \mathsf{E}\left(\int_0^T \tilde{Z}_t\mathrm{d}W_t\right)^2\right] \\ &\quad + o(\delta) && (268)\\ &= \frac{\delta}{2}\left[2\int_0^T \mathsf{E}\tilde{Z}_t^2\mathrm{d}t - \int_0^T \mathsf{E}\tilde{Z}_t^2\mathrm{d}t\right] + o(\delta) && (269)\\ &= \frac{\delta}{2}\int_0^T \mathsf{E}\tilde{Z}_t^2\mathrm{d}t + o(\delta) && (270)\end{aligned}$$

and the lemma is proved. □

Appendix VII
Proof of Lemma 6

Proof: Let $Y = \sqrt{\mathsf{snr}}g(X) + N$. Since

$$0 \le H(X) - I(X;Y) = H(X|Y) \quad (271)$$

it suffices to show that the uncertainty about X given Y vanishes as $\mathsf{snr} \to \infty$

$$\lim_{\mathsf{snr}\to\infty} H(X|Y) = 0. \quad (272)$$

Assume first that X takes a finite number $(m < \infty)$ of distinct values. Given Y, let $\widehat{X}$ be the decision for X that achieves the minimum probability of error, which is denoted by p. Then

$$H(X|Y) \le H(X|\widehat{X}) \le p\log(m-1) + H_2(p) \quad (273)$$

where $H_2(\cdot)$ stands for the binary entropy function, and the second inequality is due to Fano [29]. Since $p \to 0$ as $\mathsf{snr} \to \infty$, the right-hand side of (273) vanishes and (272) is proved.

In case X takes a countable number of values and that $H(X) < \infty$, for every natural number m, let U_m be an indicator which takes the value of 1 if X takes one of the m most likely values and 0 otherwise. Let $\widehat{X}_m$ be the function of Y which minimizes $\mathsf{P}\left\{X \ne \widehat{X}_m | U_m = 1\right\}$. Then for every m

$$\begin{aligned}H(X|Y) &\le H(X|\widehat{X}_m) && (274)\\ &= H(X, U_m|\widehat{X}_m) && (275)\\ &= H(X|\widehat{X}_m, U_m) + H(U_m|\widehat{X}_m) && (276)\\ &\le \mathsf{P}\{U_m = 1\}H(X|\widehat{X}_m, U_m = 1) \\ &\quad + \mathsf{P}\{U_m = 0\}H(X|\widehat{X}_m, U_m = 0) + H(U_m) && (277)\\ &\le \mathsf{P}\{U_m = 1\}H(X|\widehat{X}_m, U_m = 1) \\ &\quad + \mathsf{P}\{U_m = 0\}H(X) + H_2(\mathsf{P}\{U_m = 0\}). && (278)\end{aligned}$$

The conditional probability of error $\mathsf{P}\{X \ne \widehat{X}_m | U_m = 1\}$ vanishes as $\mathsf{snr} \to \infty$ and so does $H(X|\widehat{X}_m, U_m = 1)$ by Fano's inequality. Therefore, for every m

$$\lim_{\mathsf{snr}\to\infty} H(X|Y) \le \mathsf{P}\{U_m = 0\}H(X) + H_2(\mathsf{P}\{U_m = 0\}). \quad (279)$$

The limit in (279) must be 0 since $\mathsf{P}\{U_m = 0\} \to 0$ as $m \to \infty$. Thus, (272) is also proved in this case.

In case $H(X) = \infty$, $H(X|U_m = 1) \to \infty$ as $m \to \infty$. For every m, the mutual information (expressed in the form of a divergence) converges

$$\lim_{\mathsf{snr}\to\infty} \mathsf{D}(P_{Y|X,U_m=1} \| P_{Y|U_m=1} | P_{X|U_m=1}) = H(X|U_m = 1). \quad (280)$$

Therefore, the mutual information increases without bound as $\mathsf{snr} \to \infty$ by also noticing

$$\begin{aligned}I(X;Y) &\ge I(X;Y|U_m) && (281)\\ &\ge \mathsf{P}\{U_m = 1\} \\ &\quad \times \mathsf{D}(P_{Y|X,U_m=1} | P_{Y|U_m=1} | P_{X|U_m=1}). && (282)\end{aligned}$$

We have thus proved (175) in all cases. □

Acknowledgment

The authors gratefully acknowledge discussions with Prof. Haya Kaspi, Prof. Tsachy Weissman, Prof. Moshe Zakai, and Prof. Ofer Zeitouni.

References

[1] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.
[2] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1118–1134, May 2002.
[3] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1319–1343, Jun. 2002.
[4] V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.
[5] T. Kailath, "A note on least squares estimates from likelihood ratios," *Inf. Contr.*, vol. 13, pp. 534–540, 1968.
[6] ——, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 2, pp. 350–361, May 1969.
[7] ——, "A further note on a general likelihood formula for random signals in Gaussian noise," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 393–396, Jul. 1970.
[8] A. G. Jaffer and S. C. Gupta, "On relations between detection and estimation of discrete time processes," *Inf. Contr.*, vol. 20, pp. 46–54, 1972.
[9] R. Esposito, "On a relation between detection and estimation in decision theory," *Inf. Contr.*, vol. 12, pp. 116–120, 1968.
[10] C. P. Hatsell and L. W. Nolte, "Some geometric properties of the likelihood ratio," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 5, pp. 616–618, Sep. 1971.
[11] R. R. Mazumdar and A. Bagchi, "On the relation between filter maps and correction factors in likelihood ratios," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 833–836, May 1995.
[12] B. D. O. Anderson and S. Chirarattananon, "Smoothing as an improvement on filtering: A universal bound," *Electron. Lett.*, vol. 7, pp. 524–525, Sep. 1971.
[13] T. Kailath and H. V. Poor, "Detection of stochastic processes," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2230–2259, Oct. 1998.
[14] T. E. Duncan, "Evaluation of likelihood functions," *Inf. Contr.*, vol. 13, pp. 62–74, 1968.
[15] ——, "On the calculation of mutual information," *SIAM J. Appl. Math.*, vol. 19, pp. 215–220, Jul. 1970.
[16] T. Kailath, "The innovations approach to detection and estimation theory," *Proc. IEEE*, vol. 58, no. 5, pp. 680–695, May 1970.
[17] T. T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white Gaussian channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 368–371, Jul. 1971.
[18] ——, "The capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 372–378, Jul. 1971.
[19] I. Bar-David and S. Shamai (Shitz), "On information transfer by envelope constrained signals over the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 34, no. 3, pp. 371–379, May 1988.
[20] N. Chayat and S. Shamai (Shitz), "Bounds on the capacity of intertransition-time-restricted binary signaling over an AWGN channel," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1992–2006, Sep. 1999.

[21] R. S. Liptser, "Optimal encoding and decoding for transmission of a Gaussian Markov signal in a noiseless-feedback channel," *Probl. Pered. Inform.*, vol. 10, pp. 3–15, Oct./Dec. 1974.

[22] E. Mayer-Wolf and M. Zakai, "On a formula relating the Shannon information to the Fisher information for the filtering problem," in *Lecture Notes in Control and Information Sciences*. Berlin, Germany: Springer-Verlag, 1983, vol. 61, pp. 164–171.

[23] R. S. Bucy, "Information and filtering," *Inf. Sci.*, vol. 18, pp. 179–187, 1979.

[24] A. B. Shmelev, "Relationship between optimal filtering and interpolation of random signals in observations against a background of white Gaussian noise," *Radiotekh. i Elektron.*, pp. 86–89, 1985.

[25] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, Jul. and Oct. 1948.

[26] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[27] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[28] S. Verdú, "Capacity region of Gaussian CDMA channels: The symbol-synchronous case," in *Proc. 24th Allerton Conf. Communication, Control and Computing*, Monticello, IL, Oct. 1986, pp. 1025–1034.

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

[30] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inf. Theory*, to be published.

[31] D. Guo, "Gaussian channels: Information, estimation and multiuser detection," Ph.D. dissertation, Dept. Elec. Eng., Princeton Univ., Princeton, NJ, 2004.

[32] M. K. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Monterey, CA, Nov. 1997, pp. 1405–1409.

[33] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 622–640, Mar. 1999.

[34] T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 173–187, Jan. 2005.

[35] G. D. Forney Jr, "Shannon meets Wiener II: On MMSE estimation in successive decoding schemes," in *Proc. 42nd Allerton Conf. Commun., Control, and Computing*, Monticello, IL, 2004.

[36] M. Zakai and J. Ziv, "Lower and upper bounds on the optimal filtering error of certain diffusion processes," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 3, pp. 325–331, May 1972.

[37] B. Z. Bobrovsky and M. Zakai, "A lower bound on the estimation error for certain diffusion processes," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 45–52, Jan. 1976.

[38] C. Méasson, R. Urbanke, A. Montanari, and T. Richardson. (2004) Life above threshold: From list decoding to area theorem and MSE. [Online]. Available: http://arxiv. org/abs/cs/0410028

[39] A. Ashikhmin, G. Kramer, and S. ten Brink, "Extrinsic information transfer functions: Model and erasure channel properties," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2657–2673, Nov. 2004.

[40] 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, Oct. 2004.

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

[42] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 6, pp. 751–760, Nov. 1985.

[43] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1994.

[44] M. Zakai, "On mutual information, likelihood-ratios and estimation error for the additive Gaussian channel," *IEEE Trans. Inf. Theory*, to be published.

[45] H. L. Royden, *Real Analysis*, U.K.: New York, 1988.

[46] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[47] B. Øksendal, *Stochastic Differential Equations: An Introduction With Applications*, 6th ed. Berlin, Germany: Springer-Verlag, 2003.

[48] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes I: General Theory*, 2nd ed. Berlin, Germany: Springer-Verlag, 2001.

[49] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IEEE Trans. Inf. Theory*, vol. PGIT-2, no. 3, pp. 102–108, Sep. 1956.

[50] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964.

[51] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[52] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. New York: Wiley, 1942.

[53] M. C. Yovits and J. L. Jackson, "Linear filter optimization with game theory considerations," *Proc. IRE*, vol. 43, pp. 376–376, 1955.

[54] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *J. SIAM Contr., Ser. A*, vol. 2, pp. 347–369, 1965.

[55] Y.-C. Yao, "Estimation of noisy telegraph processes: Nonlinear filtering versus nonlinear smoothing," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 444–446, May 1985.

[56] Y. Steinberg, B. Z. Bobrovsky, and Z. Schuss, "Fixed-point smoothing of scalar diffusions II: The error of the optimal smoother," *SIAM J. Appl. Math*, vol. 61, pp. 1431–1444, 2001.

[57] B. Z. Bobrovsky and M. Zakai, "Asymptotic a priori estimates for the error in the nonlinear filtering problem," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 371–376, Mar. 1982.

[58] G. L. Turin, "Signal design for sequential detection systems with feedback," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 3, pp. 401–408, Jul. 1965.

[59] J. Bertoin, *Lévy Processes*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

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