# Mutual Information and Minimum Mean-square Error in Gaussian Channels

Dongning Guo[†] Shlomo Shamai[‡] Sergio Verdú [†]

[†]Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA
Email: {dguo,verdu}@princeton.edu

[‡]Dept. of Electrical Engineering
Technion-Israel Institute of Technology
32000 Haifa, Israel
Email: sshlomo@ee.technion.ac.il

*Submitted to IEEE Transactions on Information Theory, April 1, 2004.*
*Revised, September 13, 2004.*

**Abstract**

This paper deals with arbitrarily distributed finite-power input signals observed in additive Gaussian noise. 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 (smoothing).

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

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

## 1 Introduction

This paper is centered around two basic quantities that measure the noisiness of channels, namely, the *mutual information* between the input and output, and the *minimum mean-square error* (MMSE) in estimating the input given the output. It is well-known that the MMSE is achieved by conditional mean estimation. The key result of this paper is a new relationship between the mutual information and MMSE that holds regardless of the input distribution, as long as the input and output are related through additive Gaussian noise, and that the input has finite power.

Take for example the simplest scalar Gaussian channel with an arbitrary input. Let the signal-to-noise ratio (SNR) of the channel be denoted by $\mathsf{snr}$. Fix the input distribution. Both the input-output mutual information (nats) and the MMSE are then monotonic 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 go from one monotonic function to another by simply composing the inverse function of one with the other; what is quite surprising here is that the overall transformation is not only strikingly simple but 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 signal processing respectively. In general, the input-output mutual information is an indicator of how much information can be pumped through a channel reliably given a certain input signaling, whereas the MMSE measures how accurately the input can be recovered using the channel output. 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 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, 2, 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 a new idea of "incremental channels", which is to analyze the increase in the mutual information due to an infinitesimal increase in SNR, or equivalently, the decrease in mutual information due to an independent extra Gaussian noise which is infinitesimally small. The change in mutual information is found to be equal to the mutual information of a Gaussian channel whose SNR is infinitesimally small, in which region the mutual information is essentially linear in the estimation error, and hence relates the rate of mutual information increase to the MMSE.

A deeper reasoning of the relationship, however, 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 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 (or equivalently, of joint and product measures of input and output). The likelihood ratio also plays a fundamental role in detection and estimation, e.g., in hypothesis testing, it is compared to a threshold to determine which hypothesis to take. Moreover, the likelihood ratio is central in the connection of detection and estimation, in either continuous setting [5, 6, 7] or discrete one [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]). The relationships between information and estimation have been continuously used to evaluate results in one area taking advantage of known results from the other. This is best exemplified by the classical capacity-rate distortion relations, that have been used to develop lower bounds on estimation errors on one hand [12] and on the other to find achievable bounds for mutual information based on estimation errors associated with linear estimators [13].

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. Let $\mathsf{cmmse}(\mathsf{snr})$ and $\mathsf{mmse}(\mathsf{snr})$ denote the causal and noncausal MM-

SEs as a function of the SNR respectively. Let $I(\mathsf{snr})$ denote now the *mutual information rate*, which measures the average mutual information between the input and output processes per unit time. Formula (1) also holds literally in this continuous-time setting, i.e., the derivative of the mutual information rate is equal to half the noncausal MMSE. Furthermore, the filtering MMSE is equal to the expected value of the smoothing MMSE:

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

where $\Gamma$ is chosen uniformly distributed between 0 and $\mathsf{snr}$. 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 [14], where a bound on the loss due to causality constraint is quantified. Duncan [15, 16], Zakai [17, ref. [53]] and Kadota *et al.* [18] pioneered the investigation of relations between the mutual information and conditional mean filtering [15, 16], which capitalized on earlier research on the "estimator-correlator" principle by Price [19], Kailath [20], and others (see [21]). In particular, Duncan showed that the input-output mutual information can be expressed as a time-integral of the causal MMSE [16].[1] Duncan's relationship is proven to be useful in a wide spectrum of applications in information theory and statistics [18, 22, 23, 24]. There are also a number of other works in this area, most notably those of Liptser [25] and Mayer-Wolf and Zakai [26], 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 [27]. In [28] Shmelev devised a general, albeit complicated, procedure to obtain the optimal smoother from the optimal filter.

The new relationships as well as Duncan's Theorem are proved in this paper using a new idea of "incremental channels", which analyzes the increase in the input-output mutual information due to an infinitesimal increase in either the signal-to-noise ratio or observation time. A counterpart of formula (1) in continuous-time setting is first established. The result connecting filtering and smoothing MMSEs is then proved by also invoking Duncan's theorem [16]. So far, no non-information-theoretic proof is known for (2). In the discrete-time setting, the identity (1) still holds, while the relationship between the mutual information and the causal MMSEs (Duncan's Theorem) doesnot: Instead, the mutual information is lower bounded by the filtering error but upper bounded by the prediction error.

The white Gaussian nature of the noise is key to this approach since: 1) the sum of independent Gaussian variates is Gaussian; and 2) the Wiener process (time-integral of white Gaussian noise) has independent increments. In fact, the relationship between the mutual information and noncausal estimation error holds in even more general settings of Gaussian channels. In a follow-up to this paper, Zakai has recently extended the central formula to the abstract Wiener space [29], which generalizes the classical $m$-dimensional Wiener process.

The newly discovered relationship between the mutual information and MMSE finds its first use in relating the mutual informations of a Gaussian vector channel under joint and separate decoding in the large-system limit [30]. The fact that the mutual information and the (noncausal) 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 using a suboptimal (resp. genie

[1] Duncan's Theorem was independently obtained by Zakai in the more general setting of inputs that may depend causally on the noisy output in a 1969 unpublished Bell Labs Memorandum (see [17]).

aided) estimator. Lower bounds on the MMSE, e.g., [31], may also lead to new lower bounds on the mutual information.

The remainder of this paper is organized as follows. Section 2 deals with random variable/vector channels, while continuous-time channels are considered in Section 3. Interactions between discrete- and continuous-time models are studied in Section 4. Results for information measures and general channels are presented in Section 5.

## 2 Scalar and Vector Channels

Let the input and output of a channel be real-valued random variables $X$ and $Y$ respectively.[2] Upon the observation of $Y$, one would like to infer the information bearing input $X$. Let the input distribution be $P_X$ and the probability density function of $Y$ conditioned on $X$ be $p_{Y|X}$. The *mutual information* between $X$ and $Y$ is a measure of how many distinct input sequences are distinguishable on average by observing the output sequence from repeated and independent use of such a channel, and is obtained as:

$$I(X;Y) = \mathsf{E}\left\{\log \frac{p_{Y|X}(Y|X)}{p_Y(Y)}\right\} \tag{3}$$

where $\mathsf{E}\{\cdot\}$ takes the expectation over the joint distribution of the random variates in the brackets. Here, $p_Y$ is the marginal probability density function of $Y$, i.e.,

$$p_Y(y) = \mathsf{E}\left\{p_{Y|X}(y|X)\right\}, \quad \forall y. \tag{4}$$

Oftentimes, one would also want an estimate of the value of $X$ given $Y$. A strategy, or more precisely, a function of the output, $f(Y)$, is therefore called for. Let the estimation error be measured in mean square sense:

$$\mathsf{E}\left\{(X - f(Y))^2\right\}. \tag{5}$$

It is well-known that the minimum of (5), referred to as the *minimum mean-square error* or MMSE, is achieved by the conditional mean estimator (e.g., [32]):

$$\widehat{X}(Y) = \mathsf{E}\{X \mid Y\}. \tag{6}$$

### 2.1 The Scalar Gaussian-noise Channel

Consider a real-valued scalar Gaussian-noise channel of the form

$$Y = \sqrt{\mathsf{snr}}\, X + N, \tag{7}$$

where $\mathsf{snr}$ denotes the signal-to-noise ratio of the observed signal,[3] and the noise $N \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable independent of the input, $X$. The input-output conditional probability density is described by

$$p_{Y|X;\mathsf{snr}}(y|x;\mathsf{snr}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(y - \sqrt{\mathsf{snr}}\, x\right)^2\right]. \tag{8}$$

[2] Random objects are always denoted by upper-case letters.

[3] If $\mathsf{E}X^2 = 1$ then $\mathsf{snr}$ complies with the usual notion of signal-to-noise power ratio $E_s/\sigma^2$.

Let the distribution of the input be $P_X$, which does not depend on $\mathsf{snr}$. The marginal probability density function of the output exists:

$$p_{Y;\mathsf{snr}}(y;\mathsf{snr}) = \mathsf{E}\left\{p_{Y|X;\mathsf{snr}}(y|X;\mathsf{snr})\right\}, \quad \forall y. \tag{9}$$

Given the channel output, the MMSE in estimating the input is a function of $\mathsf{snr}$:

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

The input-output mutual information of the channel (7) is also a function of $\mathsf{snr}$. Let it be denoted by

$$I(\mathsf{snr}) = I\left(X; \sqrt{\mathsf{snr}}\, X + N\right). \tag{11}$$

To start with, consider the special case when the distribution $P_X$ of the input $X$ is standard Gaussian. The input-output mutual information is then the well-known channel capacity under constrained input power [33]:

$$I(\mathsf{snr}) = \mathsf{C}(\mathsf{snr}) = \frac{1}{2}\log(1+\mathsf{snr}). \tag{12}$$

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

$$\widehat{X}(Y;\mathsf{snr}) = \frac{\sqrt{\mathsf{snr}}}{1+\mathsf{snr}}\, Y, \tag{13}$$

and hence the MMSE is:

$$\mathsf{mmse}(\mathsf{snr}) = \frac{1}{1+\mathsf{snr}}. \tag{14}$$

An immediate observation is

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr})\, \log e. \tag{15}$$

Here the base of logarithm is consistent with the unit of mutual information. From this point on throughout this paper, we assume nats to be the unit of all information measures, and that logarithms have base $e$, so that $\log e = 1$ disappears from (15). It turns out that the above relationship holds not only for Gaussian inputs, but for all inputs of finite power:

**Theorem 1** *For every input distribution $P_X$ that satisfies $\mathsf{E}X^2 < \infty$,*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I\left(X; \sqrt{\mathsf{snr}}\, X + N\right) = \frac{1}{2}\mathsf{mmse}\left(X \mid \sqrt{\mathsf{snr}}\, X + N\right). \tag{16}$$

*Proof:* See Section 2.3. ∎

The identity (16) reveals an intimate and intriguing connection between Shannon's mutual information and optimal estimation in the Gaussian channel (7), namely, the rate of the mutual information increase as the SNR increases is equal to half the minimum mean-square error achieved by the optimal (in general nonlinear) estimator.

Theorem 1 can also be verified for a simple and important input signaling: $\pm 1$ with equal probability. The conditional mean estimate is given by

$$\widehat{X}(Y;\mathsf{snr}) = \tanh\left(\sqrt{\mathsf{snr}}\, Y\right). \tag{17}$$

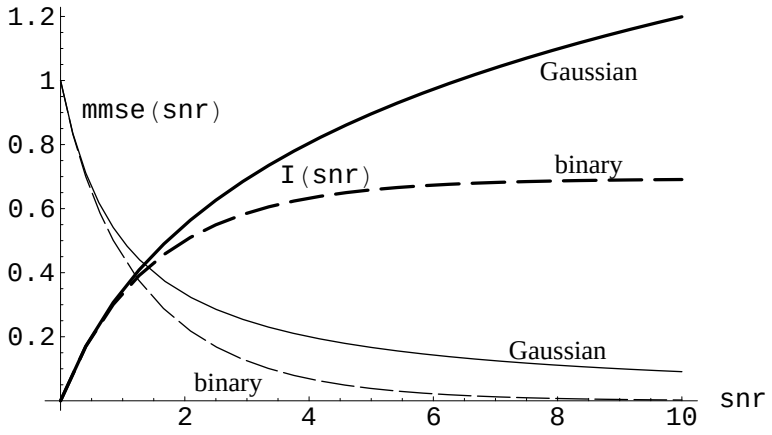


Figure 1: The mutual information (in nats) and MMSE of scalar Gaussian channel with Gaussian and binary inputs, respectively.

The MMSE and the mutual information are obtained as:

$$\mathsf{mmse}(\mathsf{snr}) = 1 - \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} \tanh(\mathsf{snr} - \sqrt{\mathsf{snr}}\, y)\ \mathrm{d}y, \tag{18}$$

and (e.g., [34, p. 274] and [35, Problem 4.22])

$$I(\mathsf{snr}) = \mathsf{snr} - \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} \log\cosh(\mathsf{snr} - \sqrt{\mathsf{snr}}\, y)\ \mathrm{d}y \tag{19}$$

respectively. Verifying (15) or (16) is a matter of algebra [36].

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

## 2.2 A Vector Channel

Consider a multiple-input multiple-output (MIMO) system described by the vector Gaussian channel:

$$\boldsymbol{Y} = \sqrt{\mathsf{snr}}\, \boldsymbol{H}\, \boldsymbol{X} + \boldsymbol{N} \tag{20}$$

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

$$p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) = (2\pi)^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\left\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x}\right\|^2\right], \tag{21}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. Let the weighted MMSE be defined as the minimum error in estimating $\boldsymbol{H}\boldsymbol{X}$:

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{E}\left\{\left\|\boldsymbol{H}\,\boldsymbol{X} - \boldsymbol{H}\,\widehat{\boldsymbol{X}}(\boldsymbol{Y};\mathsf{snr})\right\|^2\right\}, \tag{22}$$

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

**Theorem 2** *Consider the vector model (20). For every $P_{\boldsymbol{X}}$ satisfying $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$,*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I\left(\boldsymbol{X}; \sqrt{\mathsf{snr}}\,\boldsymbol{H}\,\boldsymbol{X} + \boldsymbol{N}\right) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{23}$$

*Proof:* See Section 2.3. ∎

A verification of Theorem 2 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

$$\mathsf{E}\left\{\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)^{\top}\right\} = \left(\boldsymbol{\Sigma}^{-1} + \mathsf{snr}\boldsymbol{H}^{\top}\boldsymbol{H}\right)^{-1}, \tag{24}$$

from which one can calculate the MMSE:

$$\mathsf{E}\left\{\left\|\boldsymbol{H}\,\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)\right\|^2\right\} = \mathsf{tr}\left\{\boldsymbol{H}\left(\boldsymbol{\Sigma}^{-1} + \mathsf{snr}\boldsymbol{H}^{\top}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\top}\right\}. \tag{25}$$

The mutual information is [37]:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\log\det\left(\boldsymbol{I} + \mathsf{snr}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^{\top}\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}\right), \tag{26}$$

where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is the unique positive semi-definite symmetric matrix such that $\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^2 = \boldsymbol{\Sigma}$. The relationship (23) can be checked:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\boldsymbol{X};\boldsymbol{Y}) &= \frac{1}{2}\,\mathrm{tr}\left\{\left(\boldsymbol{I} + \mathsf{snr}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^{\top}\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{H}^{\top}\boldsymbol{H}\boldsymbol{\Sigma}^{\frac{1}{2}}\right\} &(27)\\
&= \frac{1}{2}\,\mathsf{E}\left\{\left\|\boldsymbol{H}\,\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)\right\|^2\right\}. &(28)
\end{aligned}$$

Note that in the special case of independent Gaussian inputs ($\boldsymbol{\Sigma} = \boldsymbol{I}$), the MMSE in estimating $\boldsymbol{H}\,\boldsymbol{X}$ can also be written as a function of the MMSE in estimating $\boldsymbol{X}$:

$$\mathsf{E}\left\{\left\|\boldsymbol{H}\,\boldsymbol{X} - \boldsymbol{H}\,\widehat{\boldsymbol{X}}\right\|^2\right\} = \frac{1}{\mathsf{snr}}\left(K - \mathsf{E}\left\{\left\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\right\|^2\right\}\right). \tag{29}$$

Equation (29) does not hold in general for inputs not consisting of independent Gaussian entries.

The versions of Theorems 1 and 2 for complex-valued channel and signaling hold verbatim if each real/imaginary component of the circularly symmetric Gaussian noise $N$ or $\boldsymbol{N}$ has unit variance, i.e., $\mathsf{E}\{\boldsymbol{N}\boldsymbol{N}^{\mathrm{H}}\} = 2\boldsymbol{I}$. In particular, the factor of 1/2 in (16) and (23) remains intact. However, with the more common definition of $\mathsf{snr}$ in complex valued channels where the complex noise has real and imaginary components with variance 1/2 each, the factor of 1/2 in (16) and (23) disappears.

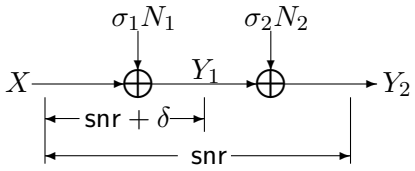


Figure 2: An SNR-incremental Gaussian channel.

### 2.3 SNR-Incremental Channel

The central relationship given by Theorems 1 and 2 can be proved in various, rather different, ways. In fact, five proofs are given in this paper, including two direct proofs by taking derivative of the mutual information and a related divergence respectively, a proof through the de Bruijn identity, and a proof taking advantage of results in the continuous-time domain. However, the most enlightening proof is by considering what we call an "incremental channel" and apply the chain rule for mutual information. A proof of Theorem 1 using this technique is given next, while its generalization to the vector version is omitted but straightforward. The alternative proofs are discussed in Section 2.6.

The key to the incremental-channel proof is to reduce the proof of the relationship for all SNRs to that for the special case of vanishing SNR, in which domain better is known about the mutual information:

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

$$Y = \sqrt{\delta}\, Z + U, \tag{30}$$

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

$$I(Y;Z) = \frac{\delta}{2}\, \mathsf{E}(Z - \mathsf{E}Z)^2 + o(\delta). \tag{31}$$

Essentially, Lemma 1 states that the mutual information is half the SNR times the variance of the input at the vicinity of zero SNR, but insensitive to the shape of the input distribution otherwise. Lemma 1 has been given in [2] and [3] (also implicitly in [1]). A proof is given in Appendix C for completeness. 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 Figure 2:

$$Y_1 = X + \sigma_1 N_1, \tag{32a}$$
$$Y_2 = Y_1 + \sigma_2 N_2, \tag{32b}$$

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

$$\mathsf{snr} + \delta = \frac{1}{\sigma_1^2}, \tag{33a}$$
$$\mathsf{snr} = \frac{1}{\sigma_1^2 + \sigma_2^2}, \tag{33b}$$

so that the signal-to-noise ratio of the first channel (32a) is $\mathsf{snr}+\delta$ and that of the composite channel is $\mathsf{snr}$. Such a channel is referred to as an *SNR-incremental Gaussian channel* since the signal-to-noise ratio increases by $\delta$ from $Y_2$ to $Y_1$. Note that we choose to scale the noise here for obvious reason.

Since the mutual information vanishes trivially at zero SNR, Theorem 1 is equivalent to the following:

$$\begin{aligned} I(X;Y_1) - I(X;Y_2) &= I(\mathsf{snr}+\delta) - I(\mathsf{snr}) && (34)\\ &= \frac{\delta}{2}\,\mathsf{mmse}(\mathsf{snr}) + o(\delta). && (35)\end{aligned}$$

Noting that $X$—$Y_1$—$Y_2$ is a Markov chain,

$$\begin{aligned} I(X;Y_1) - I(X;Y_2) &= I(X;Y_1,Y_2) - I(X;Y_2) && (36)\\ &= I(X;Y_1|Y_2), && (37)\end{aligned}$$

where (37) is by the mutual information chain rule [38]. Given $X$, the outputs $Y_1$ and $Y_2$ are jointly Gaussian. Hence $Y_1$ is Gaussian conditioned on $X$ and $Y_2$. Using (32), it is easy to check that

$$(\mathsf{snr}+\delta)\,Y_1 - \mathsf{snr}\,Y_2 - \delta\,X = \delta\,\sigma_1\,N_1 - \mathsf{snr}\,\sigma_2\,N_2. \tag{38}$$

Let

$$N = \frac{1}{\sqrt{\delta}}\,(\delta\,\sigma_1\,N_1 - \mathsf{snr}\,\sigma_2\,N_2). \tag{39}$$

Then $N$ is a standard Gaussian random variable due to (33). Given $X$, $N$ is independent of $Y_2$ since, by (32) and (33),

$$\mathsf{E}\,\{\,N\,Y_2 \mid X\} = \frac{1}{\sqrt{\delta}}\,\left(\delta\,\sigma_1^2 - \mathsf{snr}\,\sigma_2^2\right) = 0. \tag{40}$$

Therefore, (38) is tantamount to

$$(\mathsf{snr}+\delta)\,Y_1 = \mathsf{snr}\,Y_2 + \delta\,X + \sqrt{\delta}\,N, \tag{41}$$

where $N \sim \mathcal{N}(0,1)$ is independent of $X$ and $Y_2$. Clearly,

$$I(X;Y_1|Y_2) = I\left(X;\delta\,X + \sqrt{\delta}\,N \;\middle|\; Y_2\right). \tag{42}$$

Hence given $Y_2$, (41) is equivalent to a Gaussian channel with its SNR equal to $\delta$ where the input distribution is $P_{X|Y_2}$. Applying Lemma 1 to the Gaussian channel (41) conditioned on $Y_2 = y_2$, one obtains

$$I(X;Y_1|Y_2=y_2) = \frac{\delta}{2}\,\mathsf{E}\left\{\,(X - \mathsf{E}\,\{\,X \mid Y_2\})^2 \;\middle|\; Y_2 = y_2\right\} + o(\delta). \tag{43}$$

Taking the expectation over $Y_2$ on both sides of (43), one has

$$I(X;Y_1|Y_2) = \frac{\delta}{2}\,\mathsf{E}\left\{(X - \mathsf{E}\,\{\,X \mid Y_2\})^2\right\} + o(\delta), \tag{44}$$

which establishes (35) by (36) together with the fact that

$$\mathsf{E}\left\{(X - \mathsf{E}\,\{\,X \mid Y_2\})^2\right\} = \mathsf{mmse}(\mathsf{snr}). \tag{45}$$

Hence the proof of Theorem 1. ∎

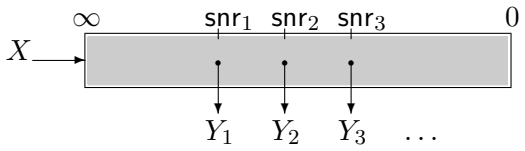


Figure 3: A Gaussian pipe where noise is added gradually.

## 2.4 Discussions

### 2.4.1 Mutual Information Chain Rule

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

$$I(X;Y_1,\ldots,Y_n)=\sum_{i=1}^{n} I(X;Y_i\,|\,Y_{i+1},\ldots,Y_n). \tag{46}$$

In case that $X$—$Y_1$—…—$Y_n$ is a Markov chain, (46) becomes

$$I(X;Y_1)=\sum_{i=1}^{n} I(X;Y_i\,|\,Y_{i+1}), \tag{47}$$

where we let $Y_{n+1}\equiv 0$. This applies to the train of outputs tapped from a Gaussian pipe where noise is added gradually until the SNR vanishes as depicted in Figure 3. The sum in (47) converges to an integral as $Y_i$ becomes a finer and finer sequence of Gaussian channel outputs by noticing from (44) that each conditional mutual information in (47) is that of 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:

$$I(\mathsf{snr})=\frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma)\;\mathrm{d}\gamma. \tag{48}$$

Therefore, the mutual information can be regarded as an accumulation of the MMSE as a function of the SNR, as is illustrated by the curves in Figure 1.

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 henceforth obtain the integral equation (16) in Theorem 1.

### 2.4.2 De Bruijn's Identity

An interesting observation here is that Theorem 2 is equivalent to the (multivariate) de Bruijn identity [39, 40]:

$$\frac{\mathrm{d}}{\mathrm{d}t}h\left(\boldsymbol{H}\boldsymbol{X}+\sqrt{t}\,\boldsymbol{N}\right)=\frac{1}{2}\mathsf{tr}\left\{\boldsymbol{J}\left(\boldsymbol{H}\boldsymbol{X}+\sqrt{t}\,\boldsymbol{N}\right)\right\} \tag{49}$$

where $h(\cdot)$ stands for the differential entropy and $\boldsymbol{J}(\cdot)$ for Fisher's information matrix [32], which is defined as[4]

$$\boldsymbol{J}(\boldsymbol{y})=\mathsf{E}\left\{\left[\nabla\log p_{\boldsymbol{Y}}(\boldsymbol{y})\right]\left[\nabla\log p_{\boldsymbol{Y}}(\boldsymbol{y})\right]^{\top}\right\}. \tag{50}$$

[4]The gradient operator can be regarded as $\nabla=\left[\frac{\partial}{\partial y_1},\cdots,\frac{\partial}{\partial y_L}\right]^{\top}$. For any differentiable function $f:\ \mathcal{R}^L\to\mathcal{R}$, its gradient at any $\boldsymbol{y}$ is a column vector $\nabla f(\boldsymbol{y})=\left[\frac{\partial f}{\partial y_1}(\boldsymbol{y}),\cdots,\frac{\partial f}{\partial y_L}(\boldsymbol{y})\right]^{\top}$.

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

$$h\left(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\,\boldsymbol{N}\right) = h(\boldsymbol{Y}) - \frac{L}{2}\log \mathsf{snr} \tag{51}$$

$$= I(\boldsymbol{X};\boldsymbol{Y}) - \frac{L}{2}\log\frac{\mathsf{snr}}{2\pi e}. \tag{52}$$

In the meantime,

$$\boldsymbol{J}\left(\boldsymbol{H}\boldsymbol{X} + \sqrt{t}\,\boldsymbol{N}\right) = \mathsf{snr}\,\boldsymbol{J}(\boldsymbol{Y}). \tag{53}$$

Note that

$$p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) = \mathsf{E}\left\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right\}, \tag{54}$$

where $p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr})$ is a Gaussian density (21). It can be shown that

$$\nabla \log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) = \sqrt{\mathsf{snr}}\,\boldsymbol{H}\widehat{\boldsymbol{X}}(\boldsymbol{y};\mathsf{snr}) - \boldsymbol{y}. \tag{55}$$

Plugging (55) into (53) and (50) gives

$$\boldsymbol{J}(\boldsymbol{Y}) = \boldsymbol{I} - \mathsf{snr}\,\boldsymbol{H}\,\mathsf{E}\left\{\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)^{\top}\right\}\boldsymbol{H}^{\top}. \tag{56}$$

Now de Bruijn's identity (49) and Theorem 2 prove each other by (52) and (56). Noting this equivalence, the incremental-channel approach offers an intuitive alternative to the conventional proof of de Bruijn's identity obtained by integrating by parts (e.g., [38]).

The inverse of Fisher's information is in general a lower bound on estimation accuracy, a result known as the Cramér-Rao lower bound [32]. For Gaussian channels, Fisher's information matrix and the covariance of conditional mean estimation error determine each other in a simple way (56). In particular, for a scalar channel,

$$J\left(\sqrt{\mathsf{snr}}\,X + N\right) = 1 - \mathsf{snr}\cdot\mathsf{mmse}(\mathsf{snr}). \tag{57}$$

### 2.4.3 Derivative of the Divergence

Consider an input-output pair $(X, Y)$ connected through (7). The mutual information $I(X;Y)$ is the average value over the input $X$ of a divergence:

$$\mathsf{D}\left(P_{Y|X=x}\|P_Y\right) = \int \log\frac{\mathrm{d}P_{Y|X=x}(y)}{\mathrm{d}P_Y(y)}\;\mathrm{d}P_{Y|X=x}(y). \tag{58}$$

Refining Theorem 1, it is possible to directly obtain the derivative of the divergence given any value of the input:

**Theorem 3** *Consider the channel (7). For every input distribution $P_X$ that satisfies $\mathsf{E}X^2 < \infty$,*

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

*where $X'$ is an auxiliary random variable which is i.i.d. with $X$ conditioned on $Y$.*

The auxiliary random variable $X'$ has an interesting physical meaning. It can be regarded as the output of the "retrochannel" [41], which takes $Y$ as the input and generates a random variable according to the posterior probability distribution $p_{X|Y;\mathsf{snr}}$. Using Theorem 3, Theorem 1 can be recovered by taking expectation on both sides of (59). The left hand side becomes the derivative of the mutual information. The right hand side becomes $1/2$ times the following:

$$\frac{1}{\sqrt{\mathsf{snr}}}\mathsf{E}\left\{(X-X')(Y-\sqrt{\mathsf{snr}}X')\right\} = \frac{1}{\sqrt{\mathsf{snr}}}\mathsf{E}\left\{XY-X'Y\right\} + \mathsf{E}\left\{(X')^2 - XX'\right\}. \tag{60}$$

Since conditioned on $Y$, $X'$ and $X$ are i.i.d., (60) can be further written as

$$\begin{aligned}\mathsf{E}\left\{X^2 - XX'\right\} &= \mathsf{E}\left\{X^2 - \mathsf{E}\left\{XX' \,\middle|\, Y;\mathsf{snr}\right\}\right\} &(61)\\ &= \mathsf{E}\left\{X^2 - \left(\mathsf{E}\left\{X \mid Y;\mathsf{snr}\right\}\right)^2\right\}, &(62)\end{aligned}$$

which is the MMSE.

### 2.4.4 Multiuser Channel

A multiuser system in which users may transmit at different signal-to-noise ratios can be better modelled by:

$$\boldsymbol{Y} = \boldsymbol{H}\,\boldsymbol{\Gamma}\,\boldsymbol{X} + \boldsymbol{N} \tag{63}$$

where $\boldsymbol{H}$ is a deterministic $L \times K$ matrix, $\boldsymbol{\Gamma} = \mathrm{diag}\{\sqrt{\mathsf{snr}_1}, \ldots, \sqrt{\mathsf{snr}_K}\}$ consists of the square-root of the SNRs of the $K$ users, and $\boldsymbol{N}$ consists of i.i.d. 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 $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$,*

$$\frac{\partial}{\partial \mathsf{snr}_k} I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\sum_{i=1}^{K}\sqrt{\frac{\mathsf{snr}_i}{\mathsf{snr}_k}}\,[\boldsymbol{H}^\top\boldsymbol{H}]_{ki}\,\mathsf{E}\left\{\mathsf{Cov}\left\{X_k, X_i|\boldsymbol{Y};\boldsymbol{\Gamma}\right\}\right\}, \tag{64}$$

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

*Proof:* The proof follows that of Theorem 2 in Appendix B (see also [36]). ∎

Using Theorem 4, Theorem 1 can be easily recovered by setting $K = 1$ and $\boldsymbol{\Gamma} = \sqrt{\mathsf{snr}}$, since

$$\mathsf{E}\left\{\mathsf{Cov}\left\{X, X|Y;\mathsf{snr}\right\}\right\} = \mathsf{E}\left\{\mathsf{var}\{X|Y;\mathsf{snr}\}\right\} \tag{65}$$

is exactly the MMSE. Theorem 2 can also be recovered by letting $\mathsf{snr}_k = \mathsf{snr}$ for all $k$. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\boldsymbol{X};\boldsymbol{Y}) &= \sum_{k=1}^{K}\frac{\partial}{\partial \mathsf{snr}_k} I(\boldsymbol{X};\boldsymbol{Y}) &(66)\\ &= \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{K}[\boldsymbol{H}^\top\boldsymbol{H}]_{ki}\,\mathsf{E}\left\{\mathsf{Cov}\left\{X_k, X_i|\boldsymbol{Y};\boldsymbol{\Gamma}\right\}\right\} &(67)\\ &= \frac{1}{2}\,\mathsf{E}\left\{\|\boldsymbol{H}\,\boldsymbol{X} - \boldsymbol{H}\,\mathsf{E}\left\{\boldsymbol{X} \mid \boldsymbol{Y};\boldsymbol{\Gamma}\right\}\|^2\right\}. &(68)\end{aligned}$$

## 2.5 Some Applications of Theorems 1 and 2

### 2.5.1 Extremality of Gaussian Inputs

Gaussian inputs are most favorable for Gaussian channels in information-theoretic sense that they maximize mutual information for a given power; on the other hand they are least favorable in estimation-theoretic sense that they maximize MMSE for a given power. 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 input is capacity achieving by showing that the linear estimation upper bound for the MMSE is achieved for Gaussian inputs.

### 2.5.2 Joint and Separate Decoding Capacities

Theorem 1 has been first used in [30] to show a relationship between code-division multiple access (CDMA) channel capacities under joint and separate decoding.

Consider a CDMA channel described by (63) where $\sqrt{L}\,\boldsymbol{H}$ consists of i.i.d. entries with zero mean and unit variance (the scaling is to keep the energy of each user finite). Reference [30] studies the model in the large-system limit where the number of users $K$ and the dimensionality of the channel (i.e., the spreading factor) $L$ both tend to infinity but with a fixed ratio $\beta = K/L$. It is assumed that the signal-to-noise ratios $\mathsf{snr}_k$ are i.i.d. with distribution $P_{\mathsf{snr}}$. Under i.i.d. inputs of arbitrary distribution $P_X$, it is shown in [30] that the MMSE and mutual information converge to deterministic numbers as the system size goes to infinity. In this case, the multiuser channel can be decoupled in the large-system limit, namely, each user experiences a Gaussian single-user channel, where the interference from all other users is summarized as a signal-to-noise ratio degradation factor called the multiuser efficiency. Precisely, the equivalent single-user channel for user $k$ is described by

$$Z_k = \sqrt{\eta\,\mathsf{snr}_k}\,X_k + N_k \tag{69}$$

where $N_k \sim \mathcal{N}(0,1)$ and the multiuser efficiency $\eta$ is a solution to the fixed-point equation:

$$\eta^{-1} = 1 + \beta\,\mathsf{E}\left\{\mathsf{snr}\cdot\mathsf{mmse}(\eta\,\mathsf{snr})\right\} \tag{70}$$

where the expectation is taken over $P_{\mathsf{snr}}$, and $\mathsf{mmse}(\eta\,\mathsf{snr}_k)$ is the MMSE of the estimate of $X_k$ given $Z_k$ in the scalar Gaussian channel (69). The mutual information for user $k$ is then the input-output mutual information of the channel (69), i.e., $I(\eta\,\mathsf{snr}_k)$.

The overall spectral efficiency under suboptimal separate decoding is the sum of the single-user mutual informations divided by the spreading factor, which is simply

$$\mathsf{C}_{\text{sep}}(\beta) = \beta\,\mathsf{E}\left\{I(\eta\,\mathsf{snr})\right\}. \tag{71}$$

The optimal spectral efficiency under joint decoding is greater than that under separate decoding:

$$\mathsf{C}_{\text{joint}}(\beta) = \beta\,\mathsf{E}\left\{I(\eta\,\mathsf{snr})\right\} + \frac{1}{2}(\eta - 1 - \log\eta). \tag{72}$$

The spectral efficiencies under joint and separate decoding were first derived for Gaussian inputs in [42], and then found implicitly in [43] and later explicitly [44] for equal-power users with binary inputs. The general form (71) and (72) were derived in [30] for arbitrary input distributions and received powers.

Using the central formula given by Theorem 1, one can show the following:

**Theorem 5 (Guo and Verdú [30])** *Under every input distribution $P_X$ with $\mathsf{E}X^2 < \infty$,*

$$\mathsf{C}_{\text{joint}}(\beta) = \int_0^\beta \frac{1}{\beta'} \mathsf{C}_{\text{sep}}(\beta') \, \mathrm{d}\beta'. \tag{73}$$

*Proof:* Since $\mathsf{C}_{\text{joint}}(0) = 0$, it suffices to show

$$\beta \frac{\mathrm{d}}{\mathrm{d}\beta} \mathsf{C}_{\text{joint}}(\beta) = \mathsf{C}_{\text{sep}}(\beta). \tag{74}$$

By (71) and (72), it is enough to show

$$\beta \frac{\mathrm{d}}{\mathrm{d}\beta} \mathsf{E}\left\{ I(\eta \, \mathsf{snr}) \right\} + \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}\beta} [\eta - 1 - \log \eta] = 0. \tag{75}$$

Noticing that $\eta$ is a function of $\beta$, (75) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \mathsf{E}\left\{ I(\eta \, \mathsf{snr}) \right\} + \frac{1}{2\beta} \left( 1 - \eta^{-1} \right) = 0. \tag{76}$$

By Theorem 1,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} I(\eta \, \mathsf{snr}) = \frac{\mathsf{snr}}{2} \, \mathsf{mmse}(\eta \, \mathsf{snr}). \tag{77}$$

Thus (76) holds as $\eta$ satisfies the fixed-point equation (70). ∎

The above proof reveals that the relationship (73) hinges on the connection of the mutual information and MMSE, where the multiuser efficiency takes an interesting role of summarizing their interaction in a many-user scenario. Moveover, the integral equation (73) is a consequence of the information chain rule, which holds for all inputs and arbitrary number of users:

$$I(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{H}) = \sum_{k=1}^{K} I(X_k; \boldsymbol{Y} | \boldsymbol{H}, X_{k+1}, \ldots, X_K). \tag{78}$$

The left hand side of (78) is the total mutual information of the multiuser channel. Each summand on the right hand side of (78) is a single-user mutual information over the multiuser channel conditioned on previously decoded users' symbols. In the large-system limit, this single-user mutual information is achievable by canceling the already-decoded symbols and conditional mean estimation against the rest of the users followed by single-user decoding [30], and hence is equivalent to that of a separate detection problem with fewer users, i.e., a smaller load of $k/L < \beta$. The limit of (78) as $K \to \infty$ becomes the integral equation (73). The practical lesson from Theorem 5 is the optimality in the large-system limit of successive single-user decoding with interferences canceled from already decoded users, and MMSE multiuser detection in order to mitigate uncanceled users [45, 46, 47]. In the special case of Gaussian inputs, the optimality is known to hold for arbitrary number of users [48, 49, 50].

## 2.6 Alternative Proofs of Theorems 1 and 2

The incremental-channel proof of Theorem 1 given in Section 2.3 provides much information-theoretic insight into the result. In this subsection, we give an alternative proof of Theorem 2, which is a distilled version of the more general result of Zakai [29] (follow-up to our work) that uses the Malliavin calculus and shows that the central relationship between the mutual

information and estimation error holds 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 alternative proofs. In fact, a third proof using the de Bruijn identity is already evident in Section 2.4.2. A fourth proof of Theorems 1 and 2 by taking the derivative of the mutual information is given in Appendices A and B respectively. A fifth proof taking advantage of results in the continuous-time domain is relegated to Section 4.

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 (20) is represented by the canonical $L$-dimensional Gaussian channel:

$$\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{N}. \tag{79}$$

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

$$I(\boldsymbol{Y};\boldsymbol{Z}) = \mathsf{D}\left(P_{\boldsymbol{Y}|\boldsymbol{Z}} \| P_{\boldsymbol{Y}} | P_{\boldsymbol{Z}}\right) = \mathsf{D}\left(P_{\boldsymbol{Y}|\boldsymbol{Z}} \| P_{\boldsymbol{Y}'} | P_{\boldsymbol{Z}}\right) - \mathsf{D}\left(P_{\boldsymbol{Y}} \| P_{\boldsymbol{Y}'}\right), \tag{80}$$

where $P_{\boldsymbol{Y}'}$ is an arbitrary distribution as long as the two divergences on the right hand side of (80) are well-defined. Choose $\boldsymbol{Y}' = \boldsymbol{N}$. Then the mutual information can be expressed in the divergence between the unconditional output distribution and the noise distribution:

$$I(\boldsymbol{Y};\boldsymbol{Z}) = \frac{1}{2}\mathsf{E}\|\boldsymbol{Z}\|^2 - \mathsf{D}\left(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}\right). \tag{81}$$

Hence Theorem 2 is equivalent to the following:

**Theorem 6** *For every $P_{\boldsymbol{X}}$ satisfying $\mathsf{E}\|\boldsymbol{X}\|^2 < \infty$,*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} \mathsf{D}\left(P_{\sqrt{\mathsf{snr}}\,\boldsymbol{X}+\boldsymbol{N}} \| P_{\boldsymbol{N}}\right) = \frac{1}{2}\mathsf{E}\left\{\left\|\mathsf{E}\left\{\boldsymbol{X} \mid \sqrt{\mathsf{snr}}\,\boldsymbol{X} + \boldsymbol{N}\right\}\right\|^2\right\}. \tag{82}$$

It is clear that, $p_{\boldsymbol{Y}}$, the probability density function for the channel output exists. The likelihood ratio between two hypotheses, one with the input signal $\boldsymbol{Z}$ and the other with zero input, is given by

$$l(\boldsymbol{y}) = \frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{N}}(\boldsymbol{y})}. \tag{83}$$

Theorem 6 can be proved using some geometric properties of the above likelihood ratio. The following lemmas are important steps.

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

$$\nabla \log l(\boldsymbol{y}) = \mathsf{E}\left\{\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\right\}. \tag{84}$$

**Lemma 3 (Hatsell and Nolte [10])** *The log-likelihood ratio satisfies Poisson's equation:*[5]

$$\nabla^2 \log l(\boldsymbol{y}) = \mathsf{E}\left\{\|\boldsymbol{Z}\|^2 \,\middle|\, \boldsymbol{Y} = \boldsymbol{y}\right\} - \left\|\mathsf{E}\left\{\boldsymbol{Z} \mid \boldsymbol{Y} = \boldsymbol{y}\right\}\right\|^2. \tag{85}$$

[5]For any differentiable $\boldsymbol{f}: \mathcal{R}^L \to \mathcal{R}^L$, $\nabla \cdot \boldsymbol{f} = \sum_{l=1}^{L} \frac{\partial f_l}{\partial y_l}$. Also, if $\boldsymbol{f}$ is doubly differentiable, its Laplacian is defined as $\nabla^2 f = \nabla \cdot (\nabla f) = \sum_{l=1}^{L} \frac{\partial^2 f}{\partial y_l^2}$.

From Lemmas 2 and 3,

$$\begin{aligned}\mathsf{E}\left\{\|\boldsymbol{Z}\|^2 \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right\} &= \nabla^2 \log l(\boldsymbol{y}) + \|\nabla \log l(\boldsymbol{y})\|^2 &(86)\\ &= \frac{l(y)\nabla^2 \log l(\boldsymbol{y}) - \|\nabla l(\boldsymbol{y})\|^2 + \|\nabla l(\boldsymbol{y})\|^2}{l^2(\boldsymbol{y})}. &(87)\end{aligned}$$

Thus we have proved

**Lemma 4**

$$\mathsf{E}\left\{\|\boldsymbol{Z}\|^2 \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right\} = \frac{\nabla^2 l(\boldsymbol{y})}{l(\boldsymbol{y})}. \tag{88}$$

A proof of Theorem 6 is obtained by taking the derivative directly.

*Proof:* [Theorem 6] Note that the likelihood ratio can be expressed as

$$\begin{aligned} l(\boldsymbol{y}) &= \frac{\mathsf{E}\left\{p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{X})\right\}}{p_{\boldsymbol{N}}(\boldsymbol{y})} &(89)\\ &= \mathsf{E}\left\{\exp\left[\sqrt{\mathsf{snr}}\,\boldsymbol{y}^\top \boldsymbol{X} - \frac{\mathsf{snr}}{2}\|\boldsymbol{X}\|^2\right]\right\}. &(90)\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} l(\boldsymbol{y}) &= \frac{1}{2}\mathsf{E}\left\{\left(\frac{1}{\sqrt{\mathsf{snr}}}\boldsymbol{y}^\top \boldsymbol{X} - \|\boldsymbol{X}\|^2\right)\cdot \exp\left[\sqrt{\mathsf{snr}}\,\boldsymbol{y}^\top \boldsymbol{X} - \frac{\mathsf{snr}}{2}\|\boldsymbol{X}\|^2\right]\right\} &(91)\\ &= \frac{1}{2} l(\boldsymbol{y})\left[\frac{1}{\sqrt{\mathsf{snr}}}\boldsymbol{y}^\top \mathsf{E}\left\{\boldsymbol{X} \mid \boldsymbol{Y}=\boldsymbol{y}\right\} - \mathsf{E}\left\{\|\boldsymbol{X}\|^2 \,\middle|\, \boldsymbol{Y}=\boldsymbol{y}\right\}\right] &(92)\\ &= \frac{1}{2\mathsf{snr}}\left[l(\boldsymbol{y})\,\boldsymbol{y}^\top \nabla \log l(\boldsymbol{y}) - \nabla^2 \log l(\boldsymbol{y})\right]. &(93)\end{aligned}$$

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. This is essentially guaranteed by Lemma 8 in Appendix B.

The divergence can be written as

$$\begin{aligned}\mathsf{D}\left(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}\right) &= \int p_{\boldsymbol{Y}}(\boldsymbol{y}) \log \frac{p_{\boldsymbol{Y}}(\boldsymbol{y})}{p_{\boldsymbol{N}}(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y} &(94)\\ &= \mathsf{E}\left\{l(\boldsymbol{N}) \log l(\boldsymbol{N})\right\}, &(95)\end{aligned}$$

and its derivative

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\mathsf{D}\left(P_{\boldsymbol{Y}} \| P_{\boldsymbol{N}}\right) = \mathsf{E}\left\{\log l(\boldsymbol{N}) \frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} l(\boldsymbol{N})\right\}. \tag{96}$$

Again, the derivative and expectation can be exchanged in order. By (93), the derivative (96) can be evaluated as

$$\begin{aligned}&\frac{1}{2\mathsf{snr}}\mathsf{E}\left\{l(\boldsymbol{N}) \log l(\boldsymbol{N})\, \boldsymbol{N}\cdot \nabla \log l(\boldsymbol{N})\right\} - \frac{1}{2\mathsf{snr}}\mathsf{E}\left\{\log l(\boldsymbol{N})\, \nabla^2 l(\boldsymbol{N})\right\}\\ =\ &\frac{1}{2\mathsf{snr}}\mathsf{E}\left\{\nabla \cdot [l(\boldsymbol{N}) \log l(\boldsymbol{N}) \nabla \log l(\boldsymbol{N})] - \log l(\boldsymbol{N})\, \nabla^2 l(\boldsymbol{N})\right\} &(97)\\ =\ &\frac{1}{2\mathsf{snr}}\mathsf{E}\left\{l(\boldsymbol{N})\, \|\nabla \log l(\boldsymbol{N})\|^2\right\} &(98)\\ =\ &\frac{1}{2\mathsf{snr}}\mathsf{E}\,\|\nabla \log l(\boldsymbol{Y})\|^2 &(99)\\ =\ &\frac{1}{2}\mathsf{E}\|\mathsf{E}\left\{\boldsymbol{X} \mid \boldsymbol{Y}\right\}\|^2, &(100)\end{aligned}$$

where to write (97) one also needs the following result which can be proved easily by integration by parts:

$$\mathsf{E}\left\{\boldsymbol{N}^{\top}\boldsymbol{f}(\boldsymbol{N})\right\} = \mathsf{E}\left\{\nabla \cdot \boldsymbol{f}(\boldsymbol{N})\right\} \tag{101}$$

for all $\boldsymbol{f} : \mathcal{R}^L \to \mathcal{R}^L$ that satisfies $f_i(\boldsymbol{n})e^{-\frac{1}{2}n_i^2} \to 0$ as $n_i \to \infty$. ∎

## 2.7 Asymptotics of Mutual Information and MMSE

It can be shown that the mutual information and MMSE are both differentiable functions of the signal-to-noise ratio given any finite-power input. In the following, the asymptotics of the mutual information and MMSE at low and high signal-to-noise ratios are studied mainly for the scalar Gaussian channel.

### 2.7.1 Low-SNR Asymptotics

Using the dominated convergence theorem, one can prove continuity of the MMSE estimate:

$$\lim_{\mathsf{snr}\to 0} \mathsf{E}\left\{X \mid Y; \mathsf{snr}\right\} = \mathsf{E}X, \tag{102}$$

and hence

$$\lim_{\mathsf{snr}\to 0} \mathsf{mmse}(\mathsf{snr}) = \mathsf{mmse}(0) = \sigma_X^2 \tag{103}$$

where $\sigma_X^2$ is the input variance. It has been shown in citeVerdu02IT that symmetric (proper-complex in the complex case) signaling is second-order optimal. Indeed, for any real-valued symmetric input with unit variance, the mutual information can be expressed as

$$I(\mathsf{snr}) = \frac{1}{2}\mathsf{snr} - \frac{1}{4}\mathsf{snr}^2 + o(\mathsf{snr}^2). \tag{104}$$

A more refined study of the asymptotics is possible by examining the Taylor expansion of the following:

$$q_i(y; \mathsf{snr}) = \mathsf{E}\left\{X^i\, p_{Y|X;\mathsf{snr}}(y \,|\, X; \mathsf{snr})\right\}, \tag{105}$$

which is well-defined at least for $i = 1, 2$, and in case all moments of the input are finite, it is defined for all $i$. Clearly, the unconditional probability density function is a special case:

$$p_{Y;\mathsf{snr}}(y; \mathsf{snr}) = q_0(y; \mathsf{snr}). \tag{106}$$

As $\mathsf{snr} \to 0$,

$$\begin{aligned} q_i(y; \mathsf{snr}) = &\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, \mathsf{E}\left\{X^i \left[1 + yX\mathsf{snr}^{\frac{1}{2}} + \frac{1}{2}(y^2-1)X^2\mathsf{snr} + \frac{1}{6}(y^2-3)yX^3\mathsf{snr}^{\frac{3}{2}}\right.\right. \\ &\left.\left. + \frac{1}{24}(y^4 - 6y^2 + 3)X^4\mathsf{snr}^2 + \mathcal{O}\left(\mathsf{snr}^{\frac{5}{2}}\right)\right]\right\}. \end{aligned} \tag{107}$$

Without loss of generality, it is assumed that the input has zero mean and unit variance. For convenience, it is also assumed that the input distribution is symmetric, i.e., $X$ and $-X$ are identically distributed. In this case, the odd moments of $X$ vanishes and by (107),

$$p_{Y;\mathsf{snr}}(y; \mathsf{snr}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[1 + \frac{1}{2}(y^2-1)\mathsf{snr} + \frac{1}{24}(y^4 - 6y^2 + 3)\mathsf{E}X^4\mathsf{snr}^2 + \mathcal{O}\left(\mathsf{snr}^{\frac{5}{2}}\right)\right], \tag{108}$$

and

$$q_1(y;\mathsf{snr}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} y \left[ \mathsf{snr}^{\frac{1}{2}} + \frac{1}{6}(y^2-3)\mathsf{E}X^4 \mathsf{snr}^{\frac{3}{2}} + \mathcal{O}\left(\mathsf{snr}^{\frac{5}{2}}\right)\right]. \tag{109}$$

Thus, the conditional mean estimate is

$$\begin{aligned} \mathsf{E}\left\{X \mid Y = y;\mathsf{snr}\right\} &= \frac{q_1(y;\mathsf{snr})}{p_{Y;\mathsf{snr}}(y;\mathsf{snr})} &(110)\\ &= \sqrt{\mathsf{snr}}\, y \left[1 + \left(\frac{1}{2} - \frac{1}{2}\mathsf{E}X^4 - \frac{1}{2}y^2 + \frac{1}{6}y^2\mathsf{E}X^4\right)\mathsf{snr} + \mathcal{O}(\mathsf{snr}^2)\right]. &(111)\end{aligned}$$

Using (111), a finer characterization of the MMSE than (103) is obtained as

$$\mathsf{mmse}(\mathsf{snr}) = 1 - \mathsf{snr} + \left(3 - \frac{2}{3}\mathsf{E}X^4\right)\mathsf{snr}^2 + \mathcal{O}\left(\mathsf{snr}^3\right). \tag{112}$$

Note that the expression (104) for the mutual information can also be refined either by noting that

$$I(\mathsf{snr}) = -\frac{1}{2}\log(2\pi e) - \mathsf{E}\left\{\log p_{Y;\mathsf{snr}}(Y;\mathsf{snr})\right\}, \tag{113}$$

and using (108), or integrating both sides of (112) and invoking Theorem 1:

$$I(\mathsf{snr}) = \frac{1}{2}\mathsf{snr} - \frac{1}{4}\mathsf{snr}^2 + \left(\frac{1}{2} - \frac{1}{9}\mathsf{E}X^4\right)\mathsf{snr}^3 + \mathcal{O}\left(\mathsf{snr}^4\right). \tag{114}$$

The smoothness of the mutual information and MMSE carries over to the vector channel model (20) for finite-power inputs. The asymptotics also have their counterparts. The MMSE of a real-valued vector channel is obtained as:

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\right\} - \mathsf{snr}\cdot\mathsf{tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\right\} + \mathcal{O}(\mathsf{snr}^2) \tag{115}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the input vector. The input-output mutual information is (see [4]):

$$I(\boldsymbol{X};\sqrt{\mathsf{snr}}\,\boldsymbol{H}\,\boldsymbol{X} + \boldsymbol{N}) = \frac{\mathsf{snr}}{2}\mathsf{tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\right\} - \frac{\mathsf{snr}^2}{4}\mathsf{tr}\left\{\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{H}^\top\right\} + \mathcal{O}(\mathsf{snr}^3). \tag{116}$$

The asymptotics can be refined to any order of the signal-to-noise ratio following the above analysis.

### 2.7.2 High-SNR Asymptotics

At high signal-to-noise ratios, the mutual information does not grow without bound for finite-alphabet inputs such as the binary one (19), whereas it can increase at the speed of $\frac{1}{2}\log\mathsf{snr}$ for Gaussian inputs. Using the entropy power inequality [38], the mutual information of the scalar channel given any symmetric input distribution with a density is shown to be bounded:

$$\frac{1}{2}\log(1+\alpha\,\mathsf{snr}) \le I(\mathsf{snr}) \le \frac{1}{2}\log(1+\mathsf{snr}), \tag{117}$$

for some $\alpha \in (0,1]$.

The MMSE behavior at high SNR depends on the input distribution. The decay can be as low as $1/\mathsf{snr}$ for Gaussian input, whereas for binary input, the MMSE can also be easily shown

to be exponentially small. In fact, for binary equiprobable inputs, the MMSE given by (18) allows another representation:

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{E}\left\{\frac{2}{\exp\left[2(\mathsf{snr} - \sqrt{\mathsf{snr}}\,Y)\right] + 1}\right\} \tag{118}$$

where $Y \sim \mathcal{N}(0,1)$. The MMSE can then be upper bounded by Jensen's inequality and lower bounded by considering only negative values of $Y$:

$$\frac{1}{e^{2\mathsf{snr}}+1} < \mathsf{mmse}(\mathsf{snr}) < \frac{2}{e^{2\mathsf{snr}}+1}, \tag{119}$$

and hence

$$\lim_{\mathsf{snr}\to\infty} \frac{1}{\mathsf{snr}} \log \mathsf{mmse}(\mathsf{snr}) = -2. \tag{120}$$

If the inputs are not equiprobable, then it is possible to have an even faster decay of MMSE as $\mathsf{snr} \to \infty$. For example, using a special input of the type (similar to flash signaling [3])

$$X = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{w.p. } p, \\ -\sqrt{\frac{p}{1-p}} & \text{w.p. } 1-p, \end{cases} \tag{121}$$

it can be shown that in this case

$$\mathsf{mmse}(\mathsf{snr}) \le \frac{1}{2p(1-p)}\, e^{-\frac{\mathsf{snr}}{4p(1-p)}}. \tag{122}$$

Hence the MMSE can be made to decay faster than any given exponential by choosing a small enough $p$.

# 3 Continuous-time Channels

The success in the random variable/vector Gaussian channel setting in Section 2 can be extended to more sophisticated continuous-time models. Consider the following continuous-time Gaussian channel:

$$R_t = \sqrt{\mathsf{snr}}\, X_t + N_t, \quad t \in [0,T], \tag{123}$$

where $\{X_t\}$ is the input process, $\{N_t\}$ a white Gaussian noise with a flat double-sided spectrum of unit height, and $\mathsf{snr}$ denotes the signal-to-noise ratio. Since $\{N_t\}$ is not second-order, it is mathematically more convenient to study an equivalent model obtained by integrating the observations in (123). In a concise form, the input and output processes are related by a standard Wiener process $\{W_t\}$ (also known as the Brownian motion) independent of the input:

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, X_t\, \mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,T]. \tag{124}$$

Note that instead of scaling the Brownian motion as is ubiquitous in the literature, we choose to scale the input process so as to minimize notation in the analysis and results. The additive Brownian motion model is fundamental in many applications and is central in many textbooks (see e.g. [51]).

We are concerned with three quantities associated with the model (124), 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 $X_\tau^t$ denote the process $\{X_t\}$ in the interval $[\tau, t]$. Also, let $\mu_X$ denote the probability measure induced by $\{X_t\}$ in the interval of interest. The input-output mutual information is defined by [52, 53]:

$$I\left(X_0^T; Y_0^T\right) = \int \log \Phi \, \mathrm{d}\mu_{XY} \tag{125}$$

if the Radon-Nikodym derivative

$$\Phi = \frac{\mathrm{d}\mu_{XY}}{\mathrm{d}\mu_X \, \mathrm{d}\mu_Y} \tag{126}$$

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

$$\mathsf{cmmse}(t, \mathsf{snr}) = \mathsf{E}\left\{\left(X_t - \mathsf{E}\left\{X_t \mid Y_0^t; \mathsf{snr}\right\}\right)^2\right\}, \tag{127}$$

and

$$\mathsf{mmse}(t, \mathsf{snr}) = \mathsf{E}\left\{\left(X_t - \mathsf{E}\left\{X_t \mid Y_0^T; \mathsf{snr}\right\}\right)^2\right\}. \tag{128}$$

### 3.1 Mutual Information and MMSEs

Recall the *mutual information rate* (mutual information per unit time) defined in the usual way:

$$I(\mathsf{snr}) = \lim_{T\to\infty} \frac{1}{T} I\left(X_0^T; Y_0^T\right). \tag{129}$$

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

$$\mathsf{cmmse}(\mathsf{snr}) = \frac{1}{T}\int_0^T \mathsf{cmmse}(t, \mathsf{snr}) \, \mathrm{d}t \tag{130}$$

and

$$\mathsf{mmse}(\mathsf{snr}) = \frac{1}{T}\int_0^T \mathsf{mmse}(t, \mathsf{snr}) \, \mathrm{d}t \tag{131}$$

respectively.

To start with, let $T \to \infty$ and assume that the input to the continuous-time model (124) is stationary[6] Gaussian process with power spectrum $S_X(\omega)$. The mutual information rate was obtained by Shannon [54]:

$$I(\mathsf{snr}) = \frac{1}{2}\int_{-\infty}^{\infty} \log\left(1 + \mathsf{snr}\, S_X(\omega)\right) \frac{\mathrm{d}\omega}{2\pi}. \tag{132}$$

In this case optimal filtering and smoothing are both linear. The noncausal MMSE is due to Wiener [55],

$$\mathsf{mmse}(\mathsf{snr}) = \int_{-\infty}^{\infty} \frac{S_X(\omega)}{1 + \mathsf{snr}\, S_X(\omega)} \frac{\mathrm{d}\omega}{2\pi}, \tag{133}$$

and the causal MMSE is due to Yovits and Jackson [56]:

$$\mathsf{cmmse}(\mathsf{snr}) = \frac{1}{\mathsf{snr}}\int_{-\infty}^{\infty} \log\left(1 + \mathsf{snr}\, S_X(\omega)\right) \frac{\mathrm{d}\omega}{2\pi}. \tag{134}$$

[6]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.

From (132) and (133), 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 for the random variable channel (Theorem 1) holds literally in case of continuous-time Gaussian input process. Moreover, (132) and (134) 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 [56].

In fact, these relationships are true not only for Gaussian inputs. Theorem 1 can be generalized to the continuous-time model with an arbitrary input process:

**Theorem 7** *If the input process $\{X_t\}$ to the Gaussian channel (124) has finite average power, i.e.,*

$$\int_0^T \mathsf{E}X_t^2 \, \mathrm{d}t < \infty, \tag{135}$$

*then*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{136}$$

*Proof:* See Section 3.2. ∎

What is special for the continuous-time model is the relationship between the mutual information rate and the causal MMSE due to Duncan [16], which is put into a more concise form here:

**Theorem 8 (Duncan [16])** *For any input process with finite average power,*

$$I(\mathsf{snr}) = \frac{\mathsf{snr}}{2}\mathsf{cmmse}(\mathsf{snr}). \tag{137}$$

Together, Theorems 7 and 8 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 signal-to-noise ratio:

**Theorem 9** *For any input process with finite average power,*

$$\mathsf{cmmse}(\mathsf{snr}) = \frac{1}{\mathsf{snr}} \int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma) \, \mathrm{d}\gamma. \tag{138}$$

Equality (138) is a surprising new relationship between causal and noncausal MMSEs. It is quite remarkable considering the fact that nonlinear filtering is usually a hard problem and few special case analytical expressions are known for the optimal estimation errors in continuous-time problems.

Note that, the equality can be rewritten as

$$\mathsf{cmmse}(\mathsf{snr}) - \mathsf{mmse}(\mathsf{snr}) = -\mathsf{snr}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\mathsf{cmmse}(\mathsf{snr}), \tag{139}$$

which quantifies the increase of the minimum estimation error due to the causality constraint. It is interesting to point out that for stationary inputs the anti-causal MMSE is equal to the causal MMSE. The reason is that the noncausal MMSE remains the same in reversed time and white Gaussian noise is reversible. Note that in general the optimal anti-causal filter is different from the optimal causal filter.

It is worth pointing out that Theorems 7–9 are still valid if the time averages in (129)–(131) are replaced by their limits as $T \to \infty$. This is particularly relevant to the case of stationary inputs.

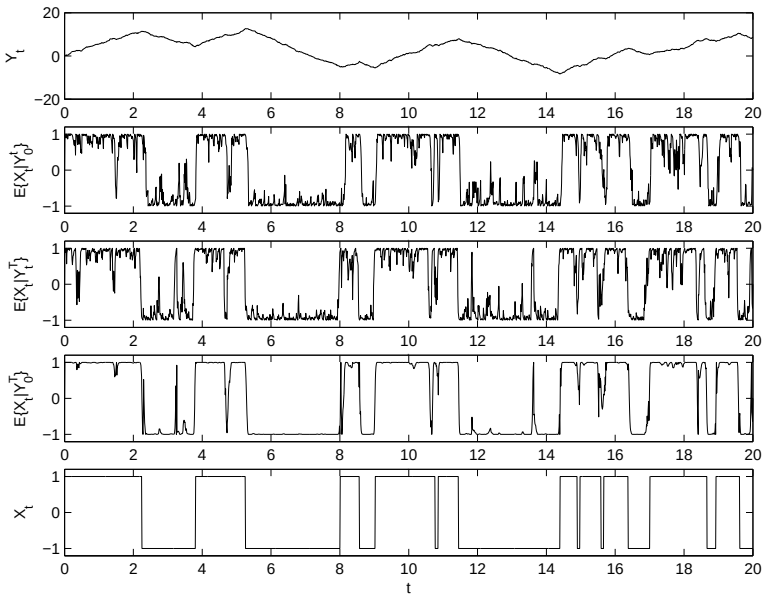


Figure 4: Sample paths of the input and output process of an additive white Gaussian noise channel, the output of the optimal forward and backward filters, as well as the output of the optimal smoother. The input $\{X_t\}$ is a random telegraph waveform with unit transition rate. The signal-to-noise ratio is 15 dB.

### 3.1.1 Random Telegraph Input

Besides Gaussian inputs, another example of the relation in Theorem 9 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 Figure 4 for an illustration. Assume that the transition rate of the input Markov process is $\nu$, i.e., for sufficiently small $h$,

$$\mathsf{P}\{X_{t+h} = X_t\} = 1 - \nu h + o(h), \tag{140}$$

the expressions for the MMSEs achieved by optimal filtering and smoothing are obtained as [57, 58]:

$$\mathsf{cmmse}(\mathsf{snr}) = \frac{\int_1^\infty u^{-\frac{1}{2}}(u-1)^{-\frac{1}{2}} e^{-\frac{2\nu u}{\mathsf{snr}}}\,\mathrm{d}u}{\int_1^\infty u^{\frac{1}{2}}(u-1)^{-\frac{1}{2}} e^{-\frac{2\nu u}{\mathsf{snr}}}\,\mathrm{d}u}, \tag{141}$$

and

$$\mathsf{mmse}(\mathsf{snr}) = \frac{\int_{-1}^1 \int_{-1}^1 \frac{(1+xy)\exp\left[-\frac{2\nu}{\mathsf{snr}}\left(\frac{1}{1-x^2}+\frac{1}{1-y^2}\right)\right]}{-(1-x)^3(1-y)^3(1+x)(1+y)}\,\mathrm{d}x\,\mathrm{d}y}{\left[\int_1^\infty u^{\frac{1}{2}}(u-1)^{-\frac{1}{2}} e^{-\frac{2\nu u}{\mathsf{snr}}}\,\mathrm{d}u\right]^2} \tag{142}$$

respectively. The relationship (138) can be verified by algebra [36]. The MMSEs are plotted in Figure 5 as functions of the SNR for unit transition rate.

Figure 4 shows experimental results of the filtering and smoothing of the random telegraph signal corrupted by additive white Gaussian noise. The forward filter follows Wonham [57]:

$$\mathrm{d}\widehat{X}_t = -\left[2\nu\widehat{X}_t + \mathsf{snr}\,\widehat{X}_t\left(1-\widehat{X}_t^2\right)\right]\mathrm{d}t + \sqrt{\mathsf{snr}}\left(1-\widehat{X}_t^2\right)\mathrm{d}Y_t, \tag{143}$$

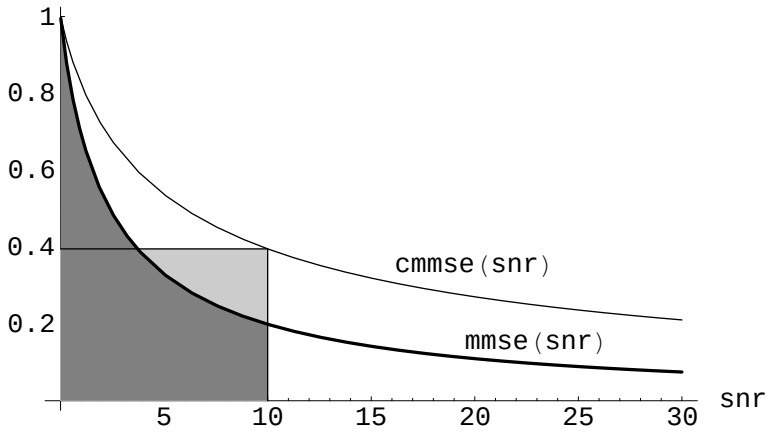


Figure 5: The causal and noncausal MMSEs of continuous-time Gaussian channel with the random telegraph waveform input. The rate $\nu = 1$. The two shaded regions have the same area due to Theorem 9.

where

$$\widehat{X}_t = \mathsf{E}\left\{ X_t \mid Y_0^t \right\}. \tag{144}$$

This is in fact resulted from a representation theorem of Doob's [59]. The backward filter is merely a time reversal of the filter of the same type. The smoother is due to Yao [58]:

$$\mathsf{E}\left\{ X_t \mid Y_0^T \right\} = \frac{\mathsf{E}\left\{ X_t \mid Y_0^t \right\} + \mathsf{E}\left\{ X_t \mid Y_t^T \right\}}{1 + \mathsf{E}\left\{ X_t \mid Y_0^t \right\} \mathsf{E}\left\{ X_t \mid Y_t^T \right\}}. \tag{145}$$

The smoother results in better MMSE of course. Numerical values of the MMSEs in Figure 4 are consistent with the curves in Figure 5.

### 3.1.2 Low- and High-SNR Asymptotics

Based on Theorem 9, one can study the asymptotics of the mutual information and MMSE under low signal-to-noise ratios. The causal and noncausal MMSE relationship implies that

$$\lim_{\mathsf{snr} \to 0} \frac{\mathsf{mmse}(0) - \mathsf{mmse}(\mathsf{snr})}{\mathsf{cmmse}(0) - \mathsf{cmmse}(\mathsf{snr})} = 2 \tag{146}$$

where

$$\mathsf{cmmse}(0) = \mathsf{mmse}(0) = \mathsf{E}X_t^2. \tag{147}$$

Hence the rate of decrease (with $\mathsf{snr}$) of the noncausal MMSE is twice that of the causal MMSE at low signal-to-noise ratios.

In the high signal-to-noise ratio regime, there exist inputs that make the MMSE exponentially small. However, in case of Gauss-Markov input processes, Steinberg *et al.* [60] observed that the causal MMSE is asymptotically twice the noncausal MMSE, as long as the input-output relationship is described by

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, h(X_t)\, \mathrm{d}t + \mathrm{d}W_t \tag{148}$$

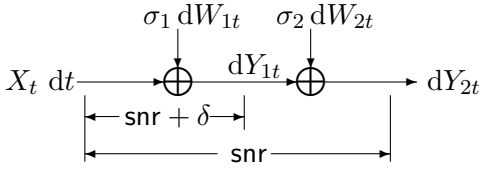


Figure 6: A continuous-time incremental Gaussian channel.

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 [61]:

$$\mathsf{mmse}(\mathsf{snr}) = \frac{c}{\sqrt{\mathsf{snr}}} + o\left(\frac{1}{\mathsf{snr}}\right), \tag{149}$$

which implies according to (138) that

$$\lim_{\mathsf{snr}\to\infty} \frac{\mathsf{cmmse}(\mathsf{snr})}{\mathsf{mmse}(\mathsf{snr})} = 2. \tag{150}$$

Unlike the universal factor of 2 result in (146) for the low signal-to-noise ratio regime, the factor of 2 result in (150) for the high signal-to-noise ratio regime fails to hold in general. For example, for the random telegraph waveform input, the causality penalty increases in the order of $\log \mathsf{snr}$ [58].

### 3.2 The SNR-Incremental Channel

Theorem 7 can be proved using the SNR-incremental channel approach developed in Section 2. Consider a cascade of two Gaussian channels with independent noise processes as depicted in Figure 6:

$$\mathrm{d}Y_{1t} = X_t\,\mathrm{d}t + \sigma_1\,\mathrm{d}W_{1t}, \tag{151a}$$

$$\mathrm{d}Y_{2t} = \mathrm{d}Y_{1t} + \sigma_2\,\mathrm{d}W_{2t}, \tag{151b}$$

where $\{W_{1t}\}$ and $\{W_{2t}\}$ are independent standard Wiener processes also independent of $\{X_t\}$, and $\sigma_1$ and $\sigma_2$ satisfy (33) so that the signal-to-noise ratio of the first channel and the composite channel is $\mathsf{snr}+\delta$ and $\mathsf{snr}$ respectively. Given $\{X_t\}$, $\{Y_{1t}\}$ and $\{Y_{2t}\}$ are jointly Gaussian processes. Following steps similar to those that lead to (41), it can be shown that

$$(\mathsf{snr}+\delta)\,\mathrm{d}Y_{1t} = \mathsf{snr}\,\mathrm{d}Y_{2t} + \delta\,X_t\,\mathrm{d}t + \sqrt{\delta}\,\mathrm{d}W_t, \tag{152}$$

where $\{W_t\}$ is a standard Wiener process independent of $\{X_t\}$ and $\{Y_{2t}\}$. Hence conditioned on the process $\{Y_{2t}\}$ in $[0,T]$, (152) can be regarded as a Gaussian channel with an SNR of $\delta$. Similar to Lemma 1, the following result holds.

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

$$\mathrm{d}Y_t = \sqrt{\delta}\,Z_t\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,T], \tag{153}$$

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

$$\int_0^T \mathsf{E}Z_t^2\,\mathrm{d}t < \infty, \tag{154}$$

*is given by the following:*

$$\lim_{\delta\to 0}\frac{1}{\delta}I\left(Z_0^T;Y_0^T\right)=\frac{1}{2}\int_0^T \mathsf{E}\left(Z_t-\mathsf{E}Z_t\right)^2\ \mathrm{d}t. \tag{155}$$

*Proof:* See Appendix D. ∎

Applying Lemma 5 to the Gaussian channel (152) conditioned on $\{Y_{2t}\}$ in $[0,T]$, one has

$$I\left(X_0^T;Y_{1,0}^T|Y_{2,0}^T\right)=\frac{\delta}{2}\int_0^T \mathsf{E}\left\{\left(X_t-\mathsf{E}\left\{X_t\mid Y_{2,0}^T\right\}\right)^2\right\}\mathrm{d}t+o(\delta). \tag{156}$$

Since $\{X_t\}$—$\{Y_{1t}\}$—$\{Y_{2t}\}$ is a Markov chain, the left hand side of (156) is recognized as the mutual information increase:

$$\begin{aligned} I\left(X_0^T;Y_{1,0}^T\,|\,Y_{2,0}^T\right) &= I\left(X_0^T;Y_{1,0}^T\right)-I\left(X_0^T;Y_{2,0}^T\right) && (157)\\ &= T\left[I(\mathsf{snr}+\delta)-I(\mathsf{snr})\right]. && (158)\end{aligned}$$

By the definition of the noncausal MMSE (128), (156) can be rewritten as

$$I(\mathsf{snr}+\delta)-I(\mathsf{snr}) = \frac{\delta}{2T}\int_0^T \mathsf{mmse}(t,\mathsf{snr})\ \mathrm{d}t+o(\delta). \tag{159}$$

Hence the proof of Theorem 7.

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 7. Indeed, by the SNR-incremental channel it has been shown that the mutual information at a given SNR is an accumulation of the MMSEs of degraded channels due to the fact that an infinitesimal increase in the signal-to-noise ratio adds to the total mutual information an increase proportional to the MMSE.

### 3.3 The Time-Incremental Channel

Note Duncan's Theorem (Theorem 8) that links the mutual information and the causal MMSE is yet another integral equation, although inexplicit, where the integral is with respect to time on the right hand side of (137). Analogous to the SNR-incremental channel, one can investigate the mutual information increase due to an infinitesimal extra time duration of observation of the channel output. This leads to a new proof of Theorem 8 in the following, which is more intuitive than Duncan's original one [16].

Theorem 8 is equivalent to

$$I\left(X_0^{t+\delta};Y_0^{t+\delta}\right)-I\left(X_0^t;Y_0^t\right)=\delta\,\frac{\mathsf{snr}}{2}\,\mathsf{E}\left\{\left(X_t-\mathsf{E}\left\{X_t\mid Y_0^t\right\}\right)^2\right\}+o(\delta), \tag{160}$$

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 (160) can be written as

$$\begin{aligned} & I\left(X_0^{t+\delta};Y_0^{t+\delta}\right)-I\left(X_0^t;Y_0^t\right)\\ =\ & I\left(X_0^t,X_t^{t+\delta};Y_0^t,Y_t^{t+\delta}\right)-I\left(X_0^t;Y_0^t\right) && (161)\\ =\ & I\left(X_t^{t+\delta};Y_t^{t+\delta}\,|\,Y_0^t\right)+I\left(X_0^t;Y_t^{t+\delta}\,|\,X_t^{t+\delta},Y_0^t\right)+I\left(X_0^t,X_t^{t+\delta};Y_0^t\right)-I\left(X_0^t;Y_0^t\right) && (162)\\ =\ & I\left(X_t^{t+\delta};Y_t^{t+\delta}\,|\,Y_0^t\right)+I\left(X_0^t;Y_t^{t+\delta}\,|\,X_t^{t+\delta},Y_0^t\right)+I\left(X_t^{t+\delta};Y_0^t\,|\,X_0^t\right). && (163)\end{aligned}$$

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 (163) vanish due to conditional independence. Therefore,

$$I\left(X_0^{t+\delta};Y_0^{t+\delta}\right)-I\left(X_0^t;Y_0^t\right)=I\left(X_t^{t+\delta};Y_t^{t+\delta}\,|\,Y_0^t\right), \tag{164}$$

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. This can be understood easily by considering a conceptual "time-incremental channel". Note that conditioned on $Y_0^t$, the channel in $(t,t+\delta)$ remains the same but with a different input distribution due to conditioning on $Y_0^t$. Let us denote this new channel by

$$\mathrm{d}\tilde{Y}_t=\sqrt{\mathsf{snr}}\,\tilde{X}_t\,\mathrm{d}t+\,\mathrm{d}W_t,\quad t\in[0,\delta], \tag{165}$$

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

$$\sqrt{\delta}\,W_{t/\delta} \tag{166}$$

is also a standard Wiener process, the channel (165) in $[0,\delta]$ is equivalent to a new channel described by

$$\mathrm{d}\tilde{\tilde{Y}}_\tau=\sqrt{\delta\,\mathsf{snr}}\,\tilde{\tilde{X}}_\tau\,\mathrm{d}\tau+\,\mathrm{d}W'_\tau,\quad \tau\in[0,1], \tag{167}$$

where $\tilde{\tilde{X}}_\tau=\tilde{X}_{\tau\delta}$, and $\{W'_t\}$ is a standard Wiener process. The channel (167) is of (fixed) unit duration but a diminishing signal-to-noise ratio of $\delta\,\mathsf{snr}$. It is interesting to note here that the trick here performs a "time-SNR" transform. By Lemma 5, the mutual information is

$$\begin{aligned}
I\left(X_t^{t+\delta};Y_t^{t+\delta}|Y_0^t\right) &= I\left(\tilde{\tilde{X}}_0^1;\tilde{\tilde{Y}}_0^1\right) && (168)\\
&= \frac{\delta\,\mathsf{snr}}{2}\int_0^1 \mathsf{E}(\tilde{\tilde{X}}_\tau-\mathsf{E}\tilde{\tilde{X}}_\tau)^2\,\mathrm{d}\tau+o(\delta) && (169)\\
&= \frac{\delta\,\mathsf{snr}}{2}\int_0^1 \mathsf{E}\left\{\left(X_{t+\tau\delta}-\mathsf{E}\left\{X_{t+\tau\delta}\,|\,Y_0^t;\mathsf{snr}\right\}\right)^2\right\}\,\mathrm{d}\tau+o(\delta) && (170)\\
&= \frac{\delta\,\mathsf{snr}}{2}\mathsf{E}\left\{\left(X_t-\mathsf{E}\left\{X_t\,|\,Y_0^t;\mathsf{snr}\right\}\right)^2\right\}+o(\delta), && (171)
\end{aligned}$$

where (171) is justified by the continuity of the MMSE. The relation (160) is then established due to (164) and (171), and hence the proof of Theorem 8.

Similar to the discussion in Section 2.4.1, the integral equations in Theorems 7 and 8 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 property both SNR-wise and time-wise is quintessential in establishing the results.

# 4 Discrete- vs. Continuous-time

In Sections 2 and 3, the mutual information and the estimation errors have been shown to satisfy very similar relations in both the random variable/vector and the continuous-time random process models. This section bridges these results for different models under certain circumstances. Moreover, discrete-time models can be analyzed by considering piecewise constant inputs to the continuous-time channel.

### 4.1 A Fifth Proof of Theorem 1

Besides the direct and incremental-channel approaches, 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 we prove Theorem 1 using Theorem 8. The proof can be easily modified to show Theorem 2, using the vector version of Duncan's Theorem [16].

A continuous-time counterpart of the model (7) can be constructed by letting $X_t \equiv X$ for $t \in [0,1]$ where $X$ is a random variable independent of $t$:

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, X \, \mathrm{d}t + \mathrm{d}W_t. \tag{172}$$

For every $u \in [0,1]$, $Y_u$ is a sufficient statistic of the observation $Y_0^u$ for $X$ and hence also for $X_0^u$. Therefore, the input-output mutual information of the scalar channel (7) is equal to the mutual information of the continuous-time channel (172):

$$I(\mathsf{snr}) = I(X;Y_1) = I\left(X_0^1; Y_0^1\right). \tag{173}$$

Integrating both sides of (172), one has

$$Y_u = \sqrt{\mathsf{snr}}\, u\, X + W_u, \quad u \in [0,1], \tag{174}$$

where $W_u \sim \mathcal{N}(0,u)$. Note that (174) is exactly a scalar Gaussian channel with a signal-to-noise ratio of $u\,\mathsf{snr}$. Clearly, the MMSE of the continuous-time model given the observation $Y_0^u$, i.e., the causal MMSE at time $u$ with a signal-to-noise ratio of $\mathsf{snr}$, is equal to the MMSE of a scalar Gaussian channel with a signal-to-noise ratio of $u\,\mathsf{snr}$:

$$\mathsf{cmmse}(u, \mathsf{snr}) = \mathsf{mmse}(u\,\mathsf{snr}). \tag{175}$$

By Theorem 8, the mutual information can be written as

$$\begin{aligned} I(X_0^1; Y_0^1) &= \frac{\mathsf{snr}}{2} \int_0^1 \mathsf{cmmse}(u, \mathsf{snr}) \, \mathrm{d}u & (176)\\ &= \frac{\mathsf{snr}}{2} \int_0^1 \mathsf{mmse}(u\,\mathsf{snr}) \, \mathrm{d}u & (177)\\ &= \frac{1}{2} \int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma) \, \mathrm{d}\gamma. & (178) \end{aligned}$$

Thus Theorem 1 follows by also noticing (173).

Note also that in this setting, the MMSE at any time $t$ of a continuous-time Gaussian channel with a signal-to-noise ratio of $u\,\mathsf{snr}$ is equal to the MMSE of a scalar Gaussian channel at the same SNR:

$$\mathsf{mmse}(t, u\,\mathsf{snr}) = \mathsf{mmse}(u\,\mathsf{snr}), \quad \forall t \in [0,T]. \tag{179}$$

Together, (175) and (179) yield (138) for this special input by taking average over time $u$.

Indeed, for an observation time duration $[0,u]$ of the continuous-time channel output, the corresponding signal-to-noise ratio is $u\,\mathsf{snr}$ in the equivalent scalar channel model; or in other words, the useful signal energy is accumulated over time. The integral over time in (137) and the integral over signal-to-noise ratio are interchangeable in this case. This is clearly another example of the "time-SNR" transform which is also used in Section 3.3.

In retrospect of the above proof, the time-invariant input can be replaced by a general form of $X\,h(t)$, where $h(t)$ is a deterministic signal.

### 4.2 Discrete-time Channels

Consider the case where the input is a discrete-time process and the channel is

$$Y_i = \sqrt{\mathsf{snr}}\, X_i + N_i, \quad i = 1, 2, \ldots, \tag{180}$$

where the noise $N_i$ is a sequence of i.i.d. standard Gaussian random variables. Given that we have already treated the case of a finite-dimensional vector channel, an advantageous analysis of (180) consists of treating the finite-horizon case $i = 1, \ldots, n$ and then taking the limit as $n \to \infty$.

Let $\boldsymbol{X}^n$ denote a column vector formed by the sequence $X_1, \ldots, X_n$. Putting the finite-horizon version of (180) in a vector form results in a MIMO channel of the form (20) with $\boldsymbol{H}$ being the identity matrix. Therefore the relation (23) between the mutual information and the MMSE holds also in this case:

**Theorem 10** *If* $\sum_{i=1}^n \mathsf{E} X_i^2 < \infty$, *then*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I\left(\boldsymbol{X}^n; \sqrt{\mathsf{snr}}\, \boldsymbol{X}^n + \boldsymbol{N}^n\right) = \frac{1}{2} \sum_{i=1}^n \mathsf{mmse}(i, \mathsf{snr}), \tag{181}$$

*where*

$$\mathsf{mmse}(i, \mathsf{snr}) = \mathsf{E}\left\{ \left(X_i - \mathsf{E}\left\{ X_i \mid \boldsymbol{Y}^n; \mathsf{snr} \right\}\right)^2 \right\} \tag{182}$$

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

It is important to note that the MMSE in this case is noncausal since the estimate is obtained through optimal smoothing. It is also interesting to consider optimal filtering and prediction in this setting. Let the MMSE of optimal filtering be defined as

$$\mathsf{cmmse}(i, \mathsf{snr}) = \mathsf{E}\left\{ \left(X_i - \mathsf{E}\left\{ X_i \mid \boldsymbol{Y}^i; \mathsf{snr} \right\}\right)^2 \right\}, \tag{183}$$

and the MMSE of optimal one-step prediction as

$$\mathsf{pmmse}(i, \mathsf{snr}) = \mathsf{E}\left\{ \left(X_i - \mathsf{E}\left\{ X_i \mid \boldsymbol{Y}^{i-1}; \mathsf{snr} \right\}\right)^2 \right\}. \tag{184}$$

**Theorem 11** *The input-output mutual information satisfies:*

$$\frac{\mathsf{snr}}{2} \sum_{i=1}^n \mathsf{cmmse}(i, \mathsf{snr}) \le I\left(\boldsymbol{X}^n; \boldsymbol{Y}^n\right) \le \frac{\mathsf{snr}}{2} \sum_{i=1}^n \mathsf{pmmse}(i, \mathsf{snr}). \tag{185}$$

*Proof:* Consider the discrete-time model (180) and its piecewise constant continuous-time counterpart:

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, X_{\lceil t \rceil}\, \mathrm{d}t + \mathrm{d}W_t, \quad t \in [0, \infty). \tag{186}$$

It is clear that in the time interval $(i-1, i]$ the input to the continuous-time model is equal to the random variable $X_i$. Note the delicacy in notation. $Y_0^n$ stands for a sample path of the continuous-time random process $\{Y_t, t \in [0, n]\}$, $\boldsymbol{Y}^n$ stands for a discrete-time process $\{Y_1, \ldots, Y_n\}$, or the vector consisting of samples of $\{Y_t\}$ at integer times, whereas $Y_i$ is either the $i$-th point of $\boldsymbol{Y}^n$ or the sample of $\{Y_t\}$ at $t = i$ depending on the context. It is easy to see that the samples of $\{Y_t\}$ at natural numbers are sufficient statistics for the input process $\boldsymbol{X}^n$. Hence

$$I\left(\boldsymbol{X}^n; \boldsymbol{Y}^n\right) = I\left(\boldsymbol{X}^n; Y_0^n\right). \tag{187}$$

Note that the causal MMSE of the continuous-time model takes the same value as the causal MMSE of the discrete-time model at integer values $i$. Thus it suffices to use $\mathsf{cmmse}(\cdot, \mathsf{snr})$ to denote the causal MMSE under both discrete- and continuous-time models. Here, $\mathsf{cmmse}(i, \mathsf{snr})$ is the MMSE of the estimation of $X_i$ given the observation $\boldsymbol{Y}^i$ which is a sufficient statistic of $Y_0^i$, while $\mathsf{pmmse}(i, \mathsf{snr})$ is the MMSE of the estimation of $X_i$ given the observation $\boldsymbol{Y}^{i-1}$ which is a sufficient statistic of $Y_0^{i-1}$. Suppose that $t \in (i-1, i]$. Since the filtration generated by $Y_0^i$ (or $\boldsymbol{Y}^i$) contains more information about $X_i$ than the filtration generated by $Y_0^t$, which in turn contains more information about $X_i$ than $Y_0^{i-1}$, one has

$$\mathsf{cmmse}(\lceil t \rceil, \mathsf{snr}) \le \mathsf{cmmse}(t, \mathsf{snr}) \le \mathsf{pmmse}(\lceil t \rceil, \mathsf{snr}). \tag{188}$$

Integrating (188) over $t$ establishes Theorem 11 by noting also that

$$I\left(\boldsymbol{X}^n; \boldsymbol{Y}^n\right) = \frac{\mathsf{snr}}{2} \int_0^n \mathsf{cmmse}(t, \mathsf{snr}) \, \mathrm{d}t \tag{189}$$

due to Theorem 8. ∎

The above analysis can also be reversed to prove the continuous-time results (Theorems 7 and 8) starting from the discrete-time ones (Theorems 10 and 11) through piecewise constant process approximations at least for continuous input processes. In particular, let $\boldsymbol{X}^n$ be the samples of $X_t$ equally spaced in $[0, T]$. Letting $n \to \infty$ allows Theorem 8 to be recovered from Theorem 11, since the sum on both sides of (185) (divided by $n$) converge to integrals and the prediction MMSE converges to the causal MMSE due to continuity.

# 5 Generalizations and Observations

## 5.1 General Additive-noise Channel

Theorems 1 and 2 show the relationship between the mutual information and the MMSE as long as the mutual information is between the observation and the signal observed embedded in Gaussian noise. Let us now consider the more general setting where the input is preprocessed arbitrarily before contamination by additive Gaussian noise as depicted in Figure 7. Let $X$ be a random message jointly distributed with a real-valued random variable $Z$. The channel output is expressed as

$$Y = \sqrt{\mathsf{snr}}\, Z + N, \tag{190}$$

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,

$$I(X; Y) = I(Z; Y) - I(Z; Y \,|\, X). \tag{191}$$

Note that given $(X, Z)$, the channel output $Y$ is Gaussian. Two applications of Theorem 1 to the right side of (191) give the following:

**Theorem 12** *Let $X$—$Z$—$Y$ be a Markov chain and $Z$ and $Y$ be connected through (190). If $\mathsf{E} Z^2 < \infty$, then*

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X; Y) = \frac{1}{2} \mathsf{E}\left\{ \left( Z - \mathsf{E}\left\{ Z \,|\, Y; \mathsf{snr} \right\} \right)^2 \right\} - \frac{1}{2} \mathsf{E}\left\{ \left( Z - \mathsf{E}\left\{ Z \,|\, Y, X; \mathsf{snr} \right\} \right)^2 \right\}. \tag{192}$$

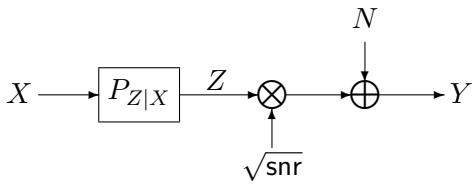


Figure 7: General additive-noise channel.

The special case of this result for zero SNR is given by Theorem 1 of [4]. As a simple illustration of Theorem 12, 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 $\sigma^2$. Then straightforward calculations yield

$$I(X;Y) = \frac{1}{2} \log \left( 1 + \frac{\mathsf{snr}\, \sigma_X^2}{1 + \mathsf{snr}\, \sigma^2} \right), \tag{193}$$

and

$$\mathsf{E}\left\{ \left( Z - \mathsf{E}\left\{ Z \mid Y; \mathsf{snr} \right\} \right)^2 \right\} = \frac{\sigma_X^2 + \sigma^2}{1 + \mathsf{snr}\left( \sigma_X^2 + \sigma^2 \right)}, \tag{194}$$

$$\mathsf{E}\left\{ \left( Z - \mathsf{E}\left\{ Z \mid Y, X; \mathsf{snr} \right\} \right)^2 \right\} = \frac{\sigma^2}{1 + \mathsf{snr}\, \sigma^2}. \tag{195}$$

The relationship (192) is easy to check.

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 (192) vanishes. Note also that since $I(X;Y) = I(g(X);Y)$ in this case, one has

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(X; \sqrt{\mathsf{snr}}\, g(X) + N) = \frac{1}{2} \mathsf{E}\left\{ \left( g(X) - \mathsf{E}\left\{ g(X) \mid Y; \mathsf{snr} \right\} \right)^2 \right\}. \tag{196}$$

Hence (16) 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 $(X_{n-1}, X_{n-2}, \ldots)$. Theorem 12 can possibly be used to calculate (or bound) the mutual information given a certain input distribution. Another domain of applications of Theorem 12 is the case of fading channels known or unknown at the receiver.

Using similar arguments as in the above, nothing prevents us from generalizing the continuous-time results in Section 3 to a much broader family of models:

$$\mathrm{d}Y_t = \sqrt{\mathsf{snr}}\, Z_t \, \mathrm{d}t + \mathrm{d}W_t, \tag{197}$$

where $\{Z_t\}$ is a random process jointly distributed with the random message $X$, and $\{W_t\}$ is a Wiener process independent of $X$ and $\{Z_t\}$. The following is straightforward in view of Theorem 12.

**Theorem 13** *As long as $\{Z_t\}$ to the channel (197) has finite average power,*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I\left(X;Y_0^T\right) =& \frac{1}{2T}\int_0^T \mathsf{E}\left\{\left(Z_t - \mathsf{E}\left\{Z_t \mid Y_0^T;\mathsf{snr}\right\}\right)^2\right\} \\ & - \mathsf{E}\left\{\left(Z_t - \mathsf{E}\left\{Z_t \mid Y_0^T, X;\mathsf{snr}\right\}\right)^2\right\}\,\mathrm{d}t.\end{aligned} \tag{198}$$

In case $Z_t = g_t(X)$, where $g_t(\,)$ is an arbitrary time-varying mapping, Theorems 7-9 hold verbatim except that the finite-power requirement now applies to $g_t(X)$, and the MMSEs in this case refer to the minimum mean-square errors in estimating $g_t(X)$.

Needless to say, extension of the results to the case of colored Gaussian noise is straightforward by filtering the observation to whiten the noise and recover the canonical model of the form (190).

## 5.2 New Representation of Information Measures

Consider a discrete random variable $X$. The mutual information between $X$ and its observation through a Gaussian channel converges to the entropy of $X$ as the signal-to-noise ratio of the channel goes to infinity.

**Lemma 6** *For any discrete real-valued random variable $X$,*

$$H(X) = \lim_{\mathsf{snr}\to\infty} I\left(X;\sqrt{\mathsf{snr}}\,X+N\right). \tag{199}$$

*Proof:* See Appendix E. ∎

Note that if $H(X)$ is infinity then the mutual information in (199) 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 (196), the following theorem is immediate.

**Theorem 14** *For any discrete random variable $X$ and one-to-one mapping $g(\cdot)$ that maps $X$ to real numbers, the entropy in nats can be obtained as*

$$H(X) = \frac{1}{2}\int_0^\infty \mathsf{E}\left\{\left(g(X) - \mathsf{E}\left\{g(X)\mid \sqrt{\mathsf{snr}}\,g(X)+N\right\}\right)^2\right\}\,\mathrm{d}\mathsf{snr}. \tag{200}$$

It is interesting to note that the integral on the right hand side of (200) is not dependent on the choice of $g(\cdot)$, which is not evident from estimation-theoretic properties alone. It is possible, however, to check this in special cases.

Other than for discrete random variables, the entropy is not defined and the input-output mutual information is in general unbounded as SNR increases. One may consider the divergence between the input distribution and a Gaussian distribution with the same mean and variance.

**Lemma 7** *For any real-valued random variable $X$. Let $X'$ 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 (7) with $X$ and $X'$ as the input respectively. Then*

$$\mathsf{D}\left(P_X\|P_{X'}\right) = \lim_{\mathsf{snr}\to\infty}\mathsf{D}\left(P_Y\|P_{Y'}\right). \tag{201}$$

The lemma can be proved using 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

$$\mathsf{D}\left(P_Y \| P_{Y'}\right) = I(X';Y') - I(X;Y), \tag{202}$$

the following theorem is straightforward by applying Theorem 1.

**Theorem 15** *For any random variable $X$ with $\sigma_X^2 < \infty$,*

$$\mathsf{D}\left(P_X \| \mathcal{N}(\mathsf{E}X, \sigma_X^2)\right) = \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\mathsf{snr}\,\sigma_X^2} - \mathsf{mmse}\left(X|\sqrt{\mathsf{snr}}\,X+N\right)\ \mathsf{dsnr}. \tag{203}$$

Note that the integrand in (203) is always positive since Gaussian inputs maximizes the MMSE. Also, Theorem 15 holds even if the divergence is infinity, for example in the case that $X$ is not a continuous random variable.

In view of Theorem 15, the differential entropy of $X$ can also be expressed as a function of the MMSE:

$$\begin{aligned} h(X) &= \frac{1}{2}\log\left(2\pi e\,\sigma_X^2\right) - \mathsf{D}\left(P_X \| P_{X'}\right) && (204)\\ &= \frac{1}{2}\log\left(2\pi e\,\sigma_X^2\right) - \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\mathsf{snr}\,\sigma_X^2} - \mathsf{mmse}\left(X|\sqrt{\mathsf{snr}}\,X+N\right)\ \mathsf{dsnr}. && (205)\end{aligned}$$

Theorem 12 provides an apparently new means of representing the mutual information between an arbitrary random variable $X$ and a real-valued random variable $Z$:

$$I(X;Z) = \frac{1}{2}\int_0^\infty \mathsf{E}\left\{\left(\mathsf{E}\left\{Z \mid \sqrt{\mathsf{snr}}\,Z+N, X\right\}\right)^2 - \left(\mathsf{E}\left\{Z \mid \sqrt{\mathsf{snr}}\,Z+N\right\}\right)^2\right\}\ \mathsf{dsnr}, \tag{206}$$

where $N$ is standard Gaussian.

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 find any application.

### 5.3 Generalization to Vector Models and Beyond

Just as that Theorem 1 obtained under a scalar model has its counterpart (Theorem 2) under a vector model, all the results in Sections 3 and 4 are generalizable to vector models, under both discrete-time and continuous-time settings. For example, the vector continuous-time model takes the form of

$$\mathrm{d}\boldsymbol{Y}_t = \sqrt{\mathsf{snr}}\,\boldsymbol{X}_t\,\mathrm{d}t + \mathrm{d}\boldsymbol{W}_t, \tag{207}$$

where $\{\boldsymbol{W}_t\}$ is an $m$-dimensional Wiener process, and $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$ are $m$-dimensional random processes. Theorem 7 holds literally, while the mutual information rate, estimation errors, and power are now defined with respect to the vector signals and their Euclidean norms. Note also that Duncan's Theorem was originally given in vector form [16]. 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 5.1 and 5.2 also extend straightforwardly to vector models.

# 6 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 differential 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 input-output mutual information of a general channel with real/complex-valued output. Asymptotics of the mutual information and MMSE are studied in both the low- and high-SNR domains.

The idea of incremental channels is the underlying basis for the most streamlined proof of the main result and for its interpretation. The results are obtained for Gaussian noise, thanks to its infinite divisibility and independent increments. The techniques in this paper are relevant for an entire family of channels the noise of which has independent increments, i.e., that is characterized by Lévy processes [62]. A particular interesting case, which will be reported elsewhere, 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. The fact that the mutual information and the (noncausal) MMSE determine each other also provides a new means to calculate or bound one quantity using the other. In all, the relations shown in this paper illuminate intimate connections between information theory and estimation theory.


# Acknowledgements

We gratefully acknowledge suggestions by Professors Tsachy Weissman, Moshe Zakai, Ofer Zeitouni, and Haya Kaspi.


# A Proof of Theorem 1

*Proof:* For simplicity, it is assumed that the order of expectation and derivative can be exchanged freely. A rigorous proof is relegated to Appendix B where every such assumption is validated in the case of more general vector model. The input-output conditional probability density function is given by (8). Let us define (same as (105))

$$q_i(y;\mathsf{snr}) = \mathsf{E}\left\{X^i\, p_{Y|X;\mathsf{snr}}(y\,|\,X;\mathsf{snr})\right\}. \tag{208}$$

Then

$$\begin{aligned} I(\mathsf{snr}) &= \mathsf{E}\left\{\log \frac{p_{Y|X;\mathsf{snr}}(Y|X;\mathsf{snr})}{p_{Y;\mathsf{snr}}(Y;\mathsf{snr})}\right\} && (209)\\ &= -\frac{1}{2}\log(2\pi e) - \int q_0(y;\mathsf{snr})\log q_0(y;\mathsf{snr})\ \mathrm{d}y. && (210)\end{aligned}$$

It is easy to check that

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} q_i(y;\mathsf{snr}) = \frac{1}{2\sqrt{\mathsf{snr}}}\, y\, q_{i+1}(y;\mathsf{snr}) - \frac{1}{2}\, q_{i+2}(y;\mathsf{snr}) = -\frac{1}{2\sqrt{\mathsf{snr}}}\frac{\mathrm{d}}{\mathrm{d}y} q_{i+1}(y;\mathsf{snr}) \tag{211}$$

as long as $q_{i+2}(y;\mathsf{snr})$ is well-defined. Therefore,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) &= -\int \left[\log q_0(y;\mathsf{snr}) + 1\right] \frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} q_0(y;\mathsf{snr})\,\mathrm{d}y && (212)\\
&= \frac{1}{2\sqrt{\mathsf{snr}}} \int \log q_0(y;\mathsf{snr})\, \frac{\mathrm{d}}{\mathrm{d}y} q_1(y;\mathsf{snr})\,\mathrm{d}y && (213)\\
&= -\frac{1}{2\sqrt{\mathsf{snr}}} \int \frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\, \frac{\mathrm{d}}{\mathrm{d}y} q_0(y;\mathsf{snr})\,\mathrm{d}y && (214)\\
&= -\frac{1}{2\sqrt{\mathsf{snr}}} \int \frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\, \left[\sqrt{\mathsf{snr}}\, q_1(y;\mathsf{snr}) - y\, q_0(y;\mathsf{snr})\right]\,\mathrm{d}y && (215)\\
&= \frac{1}{2\sqrt{\mathsf{snr}}} \int \frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\, \left[y - \sqrt{\mathsf{snr}}\, \frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}\right]\, q_0(y;\mathsf{snr})\,\mathrm{d}y, && (216)
\end{aligned}$$

where (214) is by integrating by parts. Note that the fraction in (216) is exactly the conditional mean estimate:

$$\widehat{X}(y;\mathsf{snr}) = \frac{q_1(y;\mathsf{snr})}{q_0(y;\mathsf{snr})}. \tag{217}$$

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) &= \frac{1}{2\sqrt{\mathsf{snr}}} \mathsf{E}\left\{\mathsf{E}\left\{X \mid Y;\mathsf{snr}\right\}\, \left[Y - \sqrt{\mathsf{snr}}\,\mathsf{E}\left\{X \mid Y;\mathsf{snr}\right\}\right]\right\} && (218)\\
&= \frac{1}{2}\,\mathsf{E}\left\{X^2 - \left(\mathsf{E}\left\{X \mid Y;\mathsf{snr}\right\}\right)^2\right\} && (219)\\
&= \frac{1}{2}\,\mathsf{mmse}(\mathsf{snr}). && (220)
\end{aligned}$$

■

Using the above techniques, it is not difficult to find the derivative of the conditional mean estimate $\widehat{X}(y;\mathsf{snr})$ (217) with respect to the signal-to-noise ratio. In fact, one can find any derivative of the mutual information in this way.

# B Proof of Theorem 2

*Proof:* The vector channel (20) has a Gaussian conditional density (21). The unconditional density of the channel output is given by (54), which is strictly positive for all $\boldsymbol{y}$. The mutual information can be written as

$$I(\mathsf{snr}) = -\frac{L}{2}\log(2\pi e) - \int p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\, \log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\, \mathrm{d}\boldsymbol{y}. \tag{221}$$

Hence,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) &= -\int \left[\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1\right]\, \frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\, \mathrm{d}\boldsymbol{y} && (222)\\
&= -\int \left[\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr}) + 1\right]\, \mathsf{E}\left\{\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right\}\, \mathrm{d}\boldsymbol{y}, && (223)
\end{aligned}$$

where the order of taking the derivative and expectation in (223) can be exchanged by Lemma 8, which is shown below in this Appendix. It is easy to check that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) &= \frac{1}{2\sqrt{\mathsf{snr}}}\,(\boldsymbol{H}\boldsymbol{x})^{\top}\left(\boldsymbol{y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{x}\right)\, p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) &(224)\\ &= -\frac{1}{2\sqrt{\mathsf{snr}}}\,(\boldsymbol{H}\boldsymbol{x})^{\top}\nabla p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}). &(225)\end{aligned}$$

Using (225), the right hand side of (223) can be written as

$$\frac{1}{2\sqrt{\mathsf{snr}}}\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\int\left[\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})+1\right]\,\nabla p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\,\mathrm{d}\boldsymbol{y}\right\}. \tag{226}$$

The integral in (226) can be carried out by parts to obtain

$$-\int p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\,\nabla\left[\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})+1\right]\,\mathrm{d}\boldsymbol{y}, \tag{227}$$

since for $\forall \boldsymbol{x}$,

$$p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr})\,\left[\log p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})+1\right]\to 0 \quad \text{as } \|\boldsymbol{y}\|\to\infty. \tag{228}$$

Hence, (226) can be further evaluated as

$$-\frac{1}{2\sqrt{\mathsf{snr}}}\int\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\frac{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})}{p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})}\right\}\,\nabla p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\,\mathrm{d}\boldsymbol{y} \tag{229}$$

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 (225) and Lemma 9 (shown below in this Appendix), (229) can be further written as

$$\frac{1}{2\sqrt{\mathsf{snr}}}\int\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\,\middle|\,\boldsymbol{Y}=\boldsymbol{y};\mathsf{snr}\right\}\,\mathsf{E}\left\{\left(\boldsymbol{y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{X}\right)\,p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right\}\,\mathrm{d}\boldsymbol{y}. \tag{230}$$

Therefore, (223) can be rewritten as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) &= \frac{1}{2\sqrt{\mathsf{snr}}}\int\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\,\middle|\,\boldsymbol{Y}=\boldsymbol{y};\mathsf{snr}\right\}\\ &\qquad\times\mathsf{E}\left\{\boldsymbol{y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{X}\,\middle|\,\boldsymbol{Y}=\boldsymbol{y};\mathsf{snr}\right\}\,p_{\boldsymbol{Y};\mathsf{snr}}(\boldsymbol{y};\mathsf{snr})\,\mathrm{d}\boldsymbol{y} &(231)\\ &= \frac{1}{2\sqrt{\mathsf{snr}}}\mathsf{E}\left\{\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\,\middle|\,\boldsymbol{Y};\mathsf{snr}\right\}\mathsf{E}\left\{\boldsymbol{Y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{X}\,\middle|\,\boldsymbol{Y};\mathsf{snr}\right\}\right\} &(232)\\ &= \frac{1}{2\sqrt{\mathsf{snr}}}\mathsf{E}\left\{(\boldsymbol{H}\boldsymbol{X})^{\top}\boldsymbol{Y}\right\}-\mathsf{E}\left\{\left\|\mathsf{E}\left\{\boldsymbol{H}\boldsymbol{X}\mid\boldsymbol{Y};\mathsf{snr}\right\}\right\|^2\right\} &(233)\\ &= \frac{1}{2}\mathsf{E}\left\{\|\boldsymbol{H}\boldsymbol{X}\|^2\right\}-\frac{1}{2}\mathsf{E}\left\{\left\|\mathsf{E}\left\{\boldsymbol{H}\boldsymbol{X}\mid\boldsymbol{Y};\mathsf{snr}\right\}\right\|^2\right\} &(234)\\ &= \frac{1}{2}\mathsf{E}\left\{\left\|\boldsymbol{H}\,\boldsymbol{X}-\boldsymbol{H}\,\mathsf{E}\left\{\boldsymbol{X}\mid\boldsymbol{Y};\mathsf{snr}\right\}\right\|^2\right\}. &(235)\end{aligned}$$

■

The following two lemmas were needed to justify the exchange of expectation with respect to $P_{\boldsymbol{X}}$ and derivatives in the above proof of Theorem 2.

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

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}\mathsf{E}\left\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right\} = \mathsf{E}\left\{\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right\}. \tag{236}$$

*Proof:* Let

$$f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = \frac{1}{\delta}\left[p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr}+\delta) - p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right] \tag{237}$$

and

$$f(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = \frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}}p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}). \tag{238}$$

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

$$\lim_{\delta\to 0} f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) = f(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}). \tag{239}$$

Lemma 8 is equivalent to

$$\lim_{\delta\to 0}\int f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})P_{\boldsymbol{X}}(\,\mathrm{d}\boldsymbol{x}) = \int f(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})P_{\boldsymbol{X}}(\,\mathrm{d}\boldsymbol{x}). \tag{240}$$

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

$$|f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})| < \|\boldsymbol{H}\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}}|\boldsymbol{y}^\top\boldsymbol{H}\boldsymbol{x}|. \tag{241}$$

Since the right hand side of (241) is integrable with respect to $P_{\boldsymbol{X}}$ by the assumption in the lemma, (240) holds by the Lebesgue Convergence Theorem [63]. Note that

$$\begin{aligned} f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr}) =& (2\pi)^{-\frac{L}{2}}\frac{1}{\delta}\left\{\exp\left[-\frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}+\delta}\,\boldsymbol{H}\boldsymbol{x}\|^2\right]\right. \\ &\left. -\exp\left[-\frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{x}\|^2\right]\right\}. \end{aligned} \tag{242}$$

If

$$\frac{1}{\delta} \le \|\boldsymbol{H}\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}}|\boldsymbol{y}^\top\boldsymbol{H}\boldsymbol{x}|, \tag{243}$$

then (241) holds trivially. Otherwise,

$$|f_\delta(\boldsymbol{x},\boldsymbol{y},\mathsf{snr})| < \frac{1}{\delta}\left|\exp\left[\frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}}\,\boldsymbol{H}\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y}-\sqrt{\mathsf{snr}+\delta}\,\boldsymbol{H}\boldsymbol{x}\|^2\right] - 1\right| \tag{244}$$

$$< \frac{1}{2\delta}\left(\exp\left|\delta\,\|\boldsymbol{H}\boldsymbol{x}\|^2 - \left(\sqrt{\mathsf{snr}+\delta}-\sqrt{\mathsf{snr}}\right)\boldsymbol{y}^\top\boldsymbol{H}\boldsymbol{x}\right| - 1\right) \tag{245}$$

$$< \frac{1}{2\delta}\left(\exp\left[\delta\,\left(\|\boldsymbol{H}\boldsymbol{x}\|^2 + \frac{1}{\sqrt{\mathsf{snr}}}|\boldsymbol{y}^\top\boldsymbol{H}\boldsymbol{x}|\right)\right] - 1\right). \tag{246}$$

Using the fact

$$e^t - 1 < 2t, \quad \forall\, 0 \le t < 1, \tag{247}$$

the inequality (241) holds for all $\boldsymbol{x},\boldsymbol{y},\mathsf{snr}$. ∎

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

$$\frac{\partial}{\partial y_i}\mathsf{E}\left\{p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr})\right\} = \mathsf{E}\left\{\frac{\partial}{\partial y_i}p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr})\right\}. \tag{248}$$

*Proof:* Let

$$g(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = \frac{\partial}{\partial y_i} p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{x};\mathsf{snr}) \tag{249}$$

and

$$g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = \frac{1}{\delta}\left[p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y} + \delta\, \boldsymbol{e}_i|\boldsymbol{X};\mathsf{snr}) - p_{\boldsymbol{Y}|\boldsymbol{X};\mathsf{snr}}(\boldsymbol{y}|\boldsymbol{X};\mathsf{snr})\right] \tag{250}$$

where $\boldsymbol{e}_i$ is a vector with all zero except on the $i^{\text{th}}$ entry, which is 1. Then, $\forall \boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}$,

$$\lim_{\delta \to 0} g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) = g(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}). \tag{251}$$

Lemma 9 is equivalent to

$$\lim_{\delta \to 0} \int g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) P_{\boldsymbol{X}}(\,\mathrm{d}\boldsymbol{x}) = \int g(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) P_{\boldsymbol{X}}(\,\mathrm{d}\boldsymbol{x}). \tag{252}$$

If one can show that

$$|g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr})| < |y_i| + 1 + \sqrt{\mathsf{snr}}\,|(\boldsymbol{H}\boldsymbol{x})_i|, \tag{253}$$

then (252) holds by the Lebesgue Convergence Theorem since the right hand side of (253) is integrable with respect to $P_{\boldsymbol{X}}$ by assumption. Note that

$$\begin{aligned} g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr}) =& (2\pi)^{-\frac{L}{2}} \frac{1}{\delta} \left\{ \exp\left[-\frac{1}{2}\|\boldsymbol{y} + \delta\, \boldsymbol{e}_i - \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x}\|^2\right] \right. \\ & \left. - \exp\left[-\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x}\|^2\right] \right\}. \end{aligned} \tag{254}$$

If

$$\frac{1}{\delta} \le |y_i| + 1 + \frac{1}{\sqrt{\mathsf{snr}}} |(\boldsymbol{H}\boldsymbol{x})_i|, \tag{255}$$

then (253) holds trivially. Otherwise,

$$\begin{aligned} |g_\delta(\boldsymbol{x}, \boldsymbol{y}, \mathsf{snr})| &< \frac{1}{2\delta}\left(\exp\left|\frac{1}{2}\|\boldsymbol{y} - \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y} + \delta\, \boldsymbol{e}_i - \sqrt{\mathsf{snr}}\, \boldsymbol{H}\boldsymbol{x}\|^2\right| - 1\right) && (256) \\ &= \frac{1}{2\delta}\left(\exp\left|\frac{\delta}{2}\left(2y_i + \delta - 2\sqrt{\mathsf{snr}}\,(\boldsymbol{H}\boldsymbol{x})_i\right)\right| - 1\right) && (257) \\ &< \frac{1}{2\delta}\left(\exp\left[\delta\,\left(|y_i| + 1 + \sqrt{\mathsf{snr}}\,|(\boldsymbol{H}\boldsymbol{x})_i|\right)\right] - 1\right) && (258) \\ &< |y_i| + 1 + \sqrt{\mathsf{snr}}\,|(\boldsymbol{H}\boldsymbol{x})_i|. && (259) \end{aligned}$$

■

# C Proof of Lemma 1

*Proof:* By (80), the mutual information admits the following decomposition:

$$I(Y; Z) = \mathsf{D}\left(P_{Y|Z} \| P_{Y'} | P_Z\right) - \mathsf{D}\left(P_Y \| P_{Y'}\right), \tag{260}$$

where $Y' \sim \mathcal{N}\left(\mathsf{E}Y, \sigma_Y^2\right)$. Let the variance of $Z$ be denoted by $v$. The probability density function associated with $Y'$ is

$$p_{Y'}(y) = \frac{1}{\sqrt{2\pi(\delta v + 1)}} \exp\left[-\frac{(y - \mathsf{E}Y)^2}{2(\delta v + 1)}\right]. \tag{261}$$

The first term on the right hand side of (80) is a divergence between two Gaussian distributions. Using a general formula [1]

$$\mathsf{D}\left(\mathcal{N}\left(m_1,\sigma_1^2\right)\|\mathcal{N}\left(m_0,\sigma_0^2\right)\right)=\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}+\frac{1}{2}\left(\frac{(m_1-m_0)^2}{\sigma_0^2}+\frac{\sigma_1^2}{\sigma_0^2}-1\right)\log e, \tag{262}$$

the interested divergence can be easily found as

$$\frac{1}{2}\log(1+\delta v)=\frac{\delta v}{2}+o(\delta). \tag{263}$$

The unconditional output distribution can be expressed as

$$p_Y(y)=\frac{1}{\sqrt{2\pi}}\mathsf{E}\left\{\exp\left[-\frac{1}{2}\left(y-\sqrt{\delta}\,Z\right)^2\right]\right\}. \tag{264}$$

By (261) and (264),

$$\begin{aligned}
&\log\frac{p_Y(y)}{p_{Y'}(y)}\\
=\;&\frac{1}{2}\log(1+\delta v)+\log\mathsf{E}\left\{\exp\left[\frac{(y-\sqrt{\delta}\,\mathsf{E}Z)^2}{2(\delta v+1)}-\frac{1}{2}(y-\sqrt{\delta}\,Z)^2\right]\right\} &(265)\\
=\;&\frac{1}{2}\log(1+\delta v)+\log\mathsf{E}\left\{\exp\left[\sqrt{\delta}\,y(Z-\mathsf{E}Z)-\frac{\delta}{2}\left(vy^2+Z^2-(\mathsf{E}Z)^2\right)+o(\delta)\right]\right\} &(266)\\
=\;&\frac{1}{2}\log(1+\delta v)\\
&+\log\mathsf{E}\left\{1+\sqrt{\delta}\,y(Z-\mathsf{E}Z)+\frac{\delta}{2}\left(y^2(Z-\mathsf{E}Z)^2-vy^2-Z^2+(\mathsf{E}Z)^2\right)+o(\delta)\right\} &(267)\\
=\;&\frac{1}{2}\log(1+\delta v)+\log\left(1-\frac{\delta v}{2}\right)+o(\delta) &(268)\\
=\;&o(\delta), &(269)
\end{aligned}$$

where the limit $\delta\to 0$ and the expectation can be exchanged in (268) as long as $\mathsf{E}Z^2<\infty$ due to Lebesgue Convergence Theorem [63]. Therefore, the second divergence on the right hand side of (80) is $o(\delta)$. Lemma 1 is immediate:

$$I(Y;Z)=\frac{\delta v}{2}+o(\delta). \tag{270}$$

■

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.

# D Proof of Lemma 5

Lemma 5 can be regarded as a consequence of Duncan's Theorem (Theorem 8). Consider the interval $[0,T]$. The mutual information can be expressed as a time-integral of the causal MMSE:

$$I\left(Z_0^T;Y_0^T\right)=\frac{\delta}{2}\int_0^T\mathsf{E}\left(Z_t-\mathsf{E}\left\{Z_t\,|\,Y_0^t;\delta\right\}\right)^2\,\mathrm{d}t, \tag{271}$$

Notice that as the signal-to-noise ratio $\delta \to 0$, the observed signal $Y_0^T$ becomes inconsequential in estimating the input signal. Indeed, the causal MMSE estimate converges to the unconditional mean in mean square sense:

$$\mathsf{E}\left\{ Z_t \mid Y_0^t; \delta \right\} \to \mathsf{E} Z_t. \tag{272}$$

Putting (271) and (272) together proves Lemma 5.

In parallel with the development in Theorem 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 paper so that a more intuitive proof of Duncan's Theorem is given in Section 3.3 using the idea of time-incremental channels.

*Proof:* [Lemma 5] By definition (125), the mutual information is the expectation of the logarithm of the Radon-Nikodym derivative (126), which can be obtained by the chain rule as

$$\Phi = \frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_Y \, \mathrm{d}\mu_Z} = \frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}} \left( \frac{\mathrm{d}\mu_Y}{\mathrm{d}\mu_W} \right)^{-1}. \tag{273}$$

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

$$Z_t(\omega) = Z_{t_i}(\omega), \quad t \in [t_i, t_{i+1}],\ i = 0, \ldots, n-1, \tag{274}$$

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

$$p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi(t_{i+1}-t_i)}} \exp\left[ -\frac{\left( y_{i+1} - y_i - \sqrt{\delta} z_i (t_{i+1} - t_i) \right)^2}{2(t_{i+1}-t_i)} \right]. \tag{275}$$

Easily,

$$\begin{aligned} \frac{p_{\boldsymbol{YZ}}(\boldsymbol{b},\boldsymbol{z})}{p_{\boldsymbol{WZ}}(\boldsymbol{b},\boldsymbol{z})} &= \frac{p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{b}|\boldsymbol{z})}{p_{\boldsymbol{W}}(\boldsymbol{b})} && (276) \\ &= \exp\left[ \sqrt{\delta} \sum_{i=0}^{n-1} z_i (b_{i+1} - b_i) - \frac{\delta}{2} \sum_{i=0}^{n-1} z_i^2 (t_{i+1} - t_i) \right]. && (277) \end{aligned}$$

Thus the Radon-Nikodym derivative can be established as

$$\frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}} = \exp\left[ \sqrt{\delta} \int_0^T Z_t \, \mathrm{d}W_t - \frac{\delta}{2} \int_0^T Z_t^2 \, \mathrm{d}t \right] \tag{278}$$

using the finite-dimensional likelihood ratios (277). 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 (278) of the sequence of processes also converges. Absolutely continuity is preserved. Therefore, (278) holds for all such processes $\{Z_t\}$.

The derivative (278) can be re-written as

$$\frac{\mathrm{d}\mu_{YZ}}{\mathrm{d}\mu_{WZ}} = 1 + \sqrt{\delta} \int_0^T Z_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \left( \int_0^T Z_t \, \mathrm{d}W_t \right)^2 - \int_0^T Z_t^2 \, \mathrm{d}t \right] + o(\delta). \tag{279}$$

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

$$\frac{\mathrm{d}\mu_Y}{\mathrm{d}\mu_W} = 1 + \sqrt{\delta} \int_0^T \mathsf{E} Z_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \mathsf{E}_{\mu_Z} \left( \int_0^T Z_t \, \mathrm{d}W_t \right)^2 - \int_0^T \mathsf{E} Z_t^2 \, \mathrm{d}t \right] + o(\delta). \tag{280}$$

Clearly, $\mu_Y \ll \mu_W$. Using (279), (280) and the chain rule (273), the Radon-Nikodym derivative $\Phi$ exists and is given by

$$\begin{aligned}
\Phi &= 1 + \sqrt{\delta} \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \left( \int_0^T Z_t \, \mathrm{d}W_t \right)^2 - \int_0^T Z_t^2 \, \mathrm{d}t \right. \\
&\quad \left. -2 \int_0^T \mathsf{E} Z_t \, \mathrm{d}W_t \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t - \mathsf{E}_{\mu_Z} \left( \int_0^T Z_t \, \mathrm{d}W_t \right)^2 + \int_0^T \mathsf{E} Z_t^2 \, \mathrm{d}t \right] + o(\delta) \qquad (281) \\
&= 1 + \sqrt{\delta} \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t \right)^2 \right. \\
&\quad \left. - \mathsf{E}_{\mu_Z} \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t \right)^2 - \int_0^T Z_t^2 - \mathsf{E} Z_t^2 \, \mathrm{d}t \right] + o(\delta). \qquad (282)
\end{aligned}$$

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

$$I\left(Z_0^T; Y_0^T\right) = \int \log \Phi' \, \mathrm{d}\mu_{YZ} \tag{283}$$

where $\Phi'$ is obtained from $\Phi$ (282) by substitute all occurrences of $\mathrm{d}W_t$ by $\mathrm{d}Y_t = \sqrt{\delta}\, Z_t + \mathrm{d}W_t$:

$$\begin{aligned}
\Phi' &= 1 + \sqrt{\delta} \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}Y_t + \frac{\delta}{2} \left[ \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}Y_t \right)^2 \right. \\
&\quad \left. - \mathsf{E}_{\mu_Z} \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}Y_t \right)^2 - \int_0^T Z_t^2 - \mathsf{E} Z_t^2 \, \mathrm{d}t \right] + o(\delta) \qquad (284) \\
&= 1 + \sqrt{\delta} \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t \right)^2 - \mathsf{E}_{\mu_Z} \left( \int_0^T Z_t - \mathsf{E} Z_t \, \mathrm{d}W_t \right)^2 \right. \\
&\quad \left. + \int_0^T (Z_t - \mathsf{E} Z_t)^2 \, \mathrm{d}t + \int_0^T \mathsf{E} (Z_t - \mathsf{E} Z_t)^2 \, \mathrm{d}t \right] + o(\delta) \qquad (285) \\
&= 1 + \sqrt{\delta} \int_0^T \tilde{Z}_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ \left( \int_0^T \tilde{Z}_t \, \mathrm{d}W_t \right)^2 \right. \\
&\quad \left. - \mathsf{E}_{\mu_Z} \left( \int_0^T \tilde{Z}_t \, \mathrm{d}W_t \right)^2 + \int_0^T \tilde{Z}_t^2 \, \mathrm{d}t + \int_0^T \mathsf{E} \tilde{Z}_t^2 \, \mathrm{d}t \right] + o(\delta) \qquad (286)
\end{aligned}$$

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

$$\log \Phi' = \sqrt{\delta} \int_0^T \tilde{Z}_t \, \mathrm{d}W_t + \frac{\delta}{2} \left[ -\mathsf{E}_{\mu_Z} \left( \int_0^T \tilde{Z}_t \, \mathrm{d}W_t \right)^2 + \int_0^T \tilde{Z}_t^2 \, \mathrm{d}t + \int_0^T \mathsf{E} \tilde{Z}_t^2 \, \mathrm{d}t \right] + o(\delta). \tag{287}$$

Therefore, the mutual information is

$$\begin{aligned}
\mathsf{E}\log\Phi' &= \frac{\delta}{2}\left[-\mathsf{E}\left(\int_0^T \tilde{Z}_t\ \mathrm{d}W_t\right)^2 + 2\int_0^T \mathsf{E}\tilde{Z}_t^2\ \mathrm{d}t\right] + o(\delta) && (288)\\
&= \frac{\delta}{2}\left[-\int_0^T \mathsf{E}\tilde{Z}_t^2\ \mathrm{d}t + 2\int_0^T \mathsf{E}\tilde{Z}_t^2\ \mathrm{d}t\right] + o(\delta) && (289)\\
&= \frac{\delta}{2}\int_0^T \mathsf{E}\tilde{Z}_t^2\ \mathrm{d}t + o(\delta), && (290)
\end{aligned}$$

and the lemma is proved. ∎

# E Proof of Lemma 6

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

$$0 \le H(X) - I(X;Y) = H(X|Y), \tag{291}$$

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

$$\lim_{\mathsf{snr}\to\infty} H(X|Y) = 0. \tag{292}$$

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

$$H(X|Y) \le H(X|\widehat{X}) \le p\log(m-1) + H_2(p), \tag{293}$$

where $H_2(\cdot)$ stands for the binary entropy function, and the second inequality is due to Fano [38]. Since $p \to 0$ as $\mathsf{snr} \to \infty$, the right hand side of (293) vanishes and (292) 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 $m$,

$$\begin{aligned}
H(X|Y) &\le H(X|\widehat{X}_m) && (294)\\
&= H(X, U_m|\widehat{X}_m) && (295)\\
&= H(X|\widehat{X}_m, U_m) + H(U_m|\widehat{X}) && (296)\\
&\le \mathsf{P}\{U_m = 1\}H(X|\widehat{X}, U_m = 1) + \mathsf{P}\{U_m = 0\}H(X|\widehat{X}, U_m = 0) + H(U_m) && (297)\\
&\le \mathsf{P}\{U_m = 1\}H(X|\widehat{X}, U_m = 1) + \mathsf{P}\{U_m = 0\}H(X) + H_2(\mathsf{P}\{U_m = 0\}). && (298)
\end{aligned}$$

Conditioned on $U_m = 1$, the probability of error $\mathsf{P}\left\{X \ne \widehat{X}_m | U_m = 1\right\}$ vanishes as $\mathsf{snr} \to \infty$ by Fano's inequality. Therefore, for every $m$,

$$\lim_{\mathsf{snr}\to\infty} H(X|Y) \le \mathsf{P}\{U_m = 0\}H(X) + H_2(\mathsf{P}\{U_m = 0\}). \tag{299}$$

The limit in (299) must be 0 since $\lim_{m\to\infty} \mathsf{P}\{U_m = 0\} = 0$. Thus (292) 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:

$$\lim_{\mathsf{snr}\to\infty} \mathsf{D}\left(P_{Y|X,U_m=1} \| P_{Y|U_m=1} | P_{X|U_m=1}\right) = H(X|U_m = 1). \tag{300}$$

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

$$I(X;Y) \geq I(X;Y|U_m) \geq \mathsf{P}\{U_m = 1\}\mathsf{D}\left(P_{Y|X,U_m=1} \| P_{Y|U_m=1} | P_{X|U_m=1}\right). \tag{301}$$

We have thus proved (199) in all cases. ∎

# References


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

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

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

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

[5] T. Kailath, "A note on least squares estimates from likelihood ratios," *Information and Control*, vol. 13, pp. 534–540, 1968.

[6] T. Kailath, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 15, pp. 350–361, May 1969.

[7] T. Kailath, "A further note on a general likelihood formula for random signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 16, pp. 393–396, July 1970.

[8] A. G. Jaffer and S. C. Gupta, "On relations between detection and estimation of discrete time processes," *Information and Control*, vol. 20, pp. 46–54, 1972.

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

[10] C. P. Hatsell and L. W. Nolte, "Some geometric properties of the likelihood ratio," *IEEE Trans. Inform. Theory*, vol. 17, pp. 616–618, 1971.

[11] R. R. Mazumdar and A. Bagchi, "On the relation between filter maps and correction factors in likelihood ratios," *IEEE Trans. Inform. Theory*, vol. 41, pp. 833–836, May 1995.

[12] M. Zakai and J. Ziv, "Lower and upper bounds on the optimal filtering error of certain diffusion processes," *IEEE Trans. Inform. Theory*, vol. 18, pp. 325–331, May 1972.

[13] M. Gastpar and M. Vetterli, "On the capacity of wireless networks: The relay case," in *Proceedings 2002 IEEE Infocom*, New York, 2002.

[14] B. D. O. Anderson and S. Chirarattananon, "Smoothing as an improvement on filtering: A universal bound," *Electronics Letters*, vol. 7, pp. 524–525, Sept. 1971.

[15] T. E. Duncan, "Evaluation of likelihood functions," *Information and Control*, vol. 13, pp. 62–74, 1968.

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

[17] T. Kailath, "The innovations approach to detection and estimation theory," *Proceedings of the IEEE*, vol. 58, pp. 680–695, May 1970.

[18] T. T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white Gaussian channel with and without feedback," *IEEE Trans. Inform. Theory*, vol. 17, pp. 368–371, July 1971.

[19] R. Price, "Optimum detection of random signals in noise, with application to scatter-multipath communication–I," *IRE Trans. Inform. Theory*, vol. 2, pp. 125–135, Dec. 1956.

[20] T. Kailath, "Adaptive matched filters," in *Mathematical Optimization Techniques* (R. Bellman, ed.), Univ. of California Press, 1963.

[21] T. Kailath and H. V. Poor, "Detection of stochastic processes," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2230–2259, Oct. 1998.

[22] T. T. Kadota, M. Zakai, and J. Ziv, "The capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inform. Theory*, vol. 17, pp. 372–378, July 1971.

[23] I. Bar-David and S. Shamai, "On information transfer by envelop constrained signals over the AWGN channel," *IEEE Trans. Inform. Theory*, vol. 34, pp. 371–379, May 1988.

[24] N. Chayat and S. Shamai, "Bounds on the capacity of intertransition-time-restricted binary signaling over an AWGN channel," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1992–2006, Sept. 1999.

[25] R. S. Liptser, "Optimal encoding and decoding for transmission of a Gaussian Markov signal in a noiseless-feedback channel," *Problemy Peredachi Informatsii*, vol. 10, pp. 3–15, October-December 1974.

[26] 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*, vol. 61, pp. 164–171, Springer-Verlag, New York, 1983.

[27] R. S. Bucy, "Information and filtering," *Information Sciences*, vol. 18, pp. 179–187, 1979.

[28] A. B. Shmelev, "Relationship between optimal filtering and interpolation of random signals in observations against a background of white Gaussian noise," *Radiotekhnika I Elektronika*, no. 1, pp. 86–89, 1985.

[29] M. Zakai, "On mutual information, likelihood-ratios and estimation error for the additive Gaussian channel," *preprint*, 2004.

[30] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, 2004. Submitted for publication.

[31] B. Z. Bobrovsky and M. Zakai, "A lower bound on the estimation error for certain diffusion processes," *IEEE Trans. Inform. Theory*, vol. 22, pp. 45–52, Jan. 1976.

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

[33] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

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

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

[36] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in estimation, filtering and smoothing," tech. rep., Princeton University, 2004.

[37] S. Verdú, "Capacity region of Gaussian CDMA channels: The symbol-synchronous case," in *Proceedings 24th Allerton Conference on Communication, Control and Computing*, pp. 1025–1034, Oct. 1986.

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

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

[40] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inform. Theory*, vol. 31, pp. 751–760, Nov. 1985.

[41] D. Guo and S. Verdú, "Multiuser detection and statistical mechanics," in *Communications, Information and Network Security* (V. Bhargava, H. V. Poor, V. Tarokh, and S. Yoon, eds.), ch. 13, pp. 229–277, Kluwer Academic Publishers, 2002.

[42] S. Shamai and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1302–1327, May 2001.

[43] T. Tanaka, "A statistical mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2888–2910, Nov. 2002.

[44] R. R. Müller and W. Gerstacker, "On the capacity loss due to separation of detection and decoding in large CDMA systems," in *Proceedings IEEE Information Theory Workshop*, p. 222, Bangalore, India, Oct. 2002.

[45] P. Rapajic, M. Honig, and G. Woodward, "Multiuser decision-feedback detection: Performance bounds and adaptive algorithms," in *Proceedings of 1998 IEEE International Symposium on Information Theory*, p. 34, Cambridge, MA USA, Aug. 1998.

[46] S. L. Ariyavisitakul, "Turbo space-time processing to improve wireless channel capacity," *IEEE Trans. Commun.*, vol. 48, pp. 1347–1359, Aug. 2000.

[47] R. R. Müller, "Multiuser receivers for randomly spread signals: Fundamental limits with and without decision-feedback," *IEEE Trans. Inform. Theory*, vol. 47, pp. 268–283, Jan. 2001.

[48] 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. on Signals, Systems and Computers*, pp. 1405–1409, Monterey, CA, Nov. 1997.

[49] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 622–640, March 1999.

[50] T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," *IEEE Trans. Inform. Theory*, 2004. To appear.

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

[52] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IEEE Trans. Inform. Theory*, vol. 2, pp. 102–108, Sept. 1956.

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

[54] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.

[55] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. New York: Wiley, 1949. (Originally published in Feb. 1942, as a classified Nat. Defense Res. Council rep.).

[56] M. C. Yovits and J. L. Jackson, "Linear filter optimization with game theory considerations," *Proc. IRE*, vol. 43, no. 3, p. 376, 1955.

[57] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *J. SIAM Control, Ser. A*, vol. 2, no. 3, pp. 347–369, 1965.

[58] Y.-C. Yao, "Estimation of noisy telegraph processes: Nonlinear filtering versus nonlinear smoothing," *IEEE Trans. Inform. Theory*, vol. 31, pp. 444–446, May 1985.

[59] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1990.

[60] 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.

[61] B. Z. Bobrovsky and M. Zakai, "Asymptotic a priori estimates for the error in the nonlinear filtering problem," *IEEE Trans. Inform. Theory*, vol. 28, pp. 371–376, March 1982.

[62] J. Bertoin, *Lévy Processes*. Cambridge University Press, 1996.

[63] H. L. Royden, *Real Analysis*. Macmillan, 1988.