Extensions of the I-MMSE Relation *

Guangyue Han
The University of Hong Kong
email: ghan@hku.hk

Jian Song
The University of Hong Kong
email: txjsong@hku.hk

June 28, 2015

Abstract

Unveiling a fundamental link between information theory and estimation theory, the I-MMSE relation by Guo, Shamai and Verdu [14], together with its numerous extensions, has great theoretical significance and various practical applications. On the other hand, its influences to date have been restricted to channels without feedback or memory, due to the absence of its extensions to such channels. In this paper, we propose extensions of the I-MMSE relation to discrete-time and continuous-time Gaussian channels with feedback and/or memory. Our approach is based on a very simple observation, which can be applied to other scenarios, such as a simple and direct proof of the classical de Bruijn's identity.

Index Terms: *mutual information, minimum mean-square error, the I-MMSE relation, information theory, estimation theory, feedback channel, memory channel*

1 Introduction

Consider the following discrete-time memoryless Gaussian channel

$$Y = \sqrt{snr}X + Z, \tag{1}$$

where snr denotes the signal-to-noise ratio of the channel, X and Y denote the input and output of the channel, respectively, and the standard normally distributed noise Z is independent of X. An interesting recent result by Guo, Shamai and Verdu [14] states that for any channel input X with $E[X^2] < \infty$,

$$\frac{d}{dsnr} I(X;Y) = \frac{1}{2}\mathbb{E}[(X - \mathbb{E}[X|Y])^2], \tag{2}$$

where the left hand side is the derivative of $I(X;Y)$ with respect to snr, and the right-hand side is half of the so-called minimum mean-square error (MMSE), which corresponds to the best estimation of X given the observation Y. The I-MMSE relation (2) carries over verbatim to linear vector Gaussian channels and has been widely extended to continuous-time Gaussian

*A preliminary version of this paper has been presented in IEEE ISIT 2015 [20].

channels [14], abstract Gaussian channels [48], additive channels [15], arbitrary channels [32], derivatives with respect to arbitrary parameterizations [31], higher order derivatives [33], and so on.

Unveiling an important link between information theory and estimation theory, the I-MMSE relation as above and its numerous extensions are of fundamental significance to relevant areas in these two fields and have been exerting far-reaching influences over a wide-range of topics. Representative applications include, but not limited to, power allocation of parallel Gaussian channels [27], analysis of extrinsic information of code ensembles [35], Gaussian broadcast channels [17], Gaussian wiretap channels [17, 3], Gaussian interference channels [4], interference alignment [47], a simple proof of the classical entropy power inequality [43]. For a comprehensive reference to the applications of the I-MMSE relation and its extensions, we refer to [38] .

On the other hand, all the applications of the I-MMSE relation to date have been restricted to channels without feedback or memory, due to the lack of extensions of the I-MMSE relation to such channels. In this regard, a "plain" generalization of the original I-MMSE relation to feedback channels should not be expected, which has been noted in [14], where an example is given to show that the exact I-MMSE relation fails to hold for some continuous-time feedback channel. In this paper, we remedy the situations with some explicit correctional terms (which vanish if the channel does not have feedback or memory) and extend the I-MMSE relation to channels with feedback or memory. Despite the fact that the I-MMSE relation have been examined from a number of perspectives (see its multiple proofs in [14]), our approach is still novel and powerful. As a matter of fact, other than recovering and extending the I-MMSE relation, our approach can be applied elsewhere, such as yielding a simple and direct proof of the classical de Bruijn's identity [39, 5]; see Section 2.2.

Our approach is based on a surprisingly simple idea, which can be roughly stated as follows: before taking derivative of an information-theoretic quantity with respect to certain parameters, we represent it as an expectation with respect to a probability space independent of the parameters. For illustrative purpose, in what follows, we consider the discrete-time Gaussian channel in (1) and review a "conventional" proof of (2) in [14] and compare it with ours.

First, note that for the channel in (1), taking derivative of $I(X;Y)$ is equivalent to that of $H(Y)$, which can be written as the expectation of $-\log f_Y(Y)$:

$$H(Y) = -\mathbb{E}[\log f_Y(Y)].$$

In their fourth proof of (2), the authors of [14] choose the probability space, with respect to which the expectation as above is taken, to be the sample space of Y (with naturally induced measure), which obviously depends on snr. With respect to this probability space, $H(Y)$ is naturally expressed as:

$$H(Y) = -\int_{\mathbb{R}} f_Y(y) \log f_Y(y) dy.$$

Then, under some mild assumptions, the derivative of $H(Y)$ with respect to snr can penetrate into the integral, and then (2) follows from integration by parts and other straightforward computations.

Under our approach, we would rather choose a probability space independent of snr. For example, choosing the probability space to be the sample space of (X, Z), we will express

$H(Y)$ as

$$H(Y) = -\int_{\mathbb{R}}\int_{\mathbb{R}} f_X(x) f_Z(z) \log f_Y(\sqrt{snr}x + z) dx dz.$$

It turns out such a seemingly innocent shift of viewpoint will render the follow-up computations rather simple and direct before reaching (2); and most importantly, when applied to channels with feedback or memory, it naturally leads to extensions of the I-MMSE relation. For instance, consider the discrete-time Gaussian channel with feedback:

$$Y_i = \sqrt{snr} X_i(M, Y_1^{i-1}) + Z_i, \quad i = 1, 2, \ldots, n,$$

where the channel input X_i depends on the message M and the previous channel outputs Y_1^{i-1}. Using the above-mentioned approach, we will obtain the following extension (see Remark 3.4) of the I-MMSE relation:

$$\frac{d}{dsnr} I(X_1^n \to Y_1^n) = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mathbb{E}[X_i|Y_1^n])^2\right] + snr \sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mathbb{E}[X_i|Y_1^n]) \frac{d}{dsnr} X_i\right], \quad (3)$$

where X_i is the abbreviated form of $X_i(M, Y_1^{i-1})$ and $I(X_1^n \to Y_1^n)$ is the directed information between X_1^n and Y_1^n. Directed information is a notion generalized from mutual information for feedback channels, and the second term in the right hand side of (3) is a correctional term, which vanishes when X_i does not depend on Y_1^{i-1} (*i.e.*, there is no feedback), so (3) is indeed an extension of the I-MMSE relation in (2) to discrete-time Gaussian channels with feedback. As elaborated later, the I-MMSE relation can also be extended to Gaussian channels, in either discrete-time or continuous-time, with feedback and/or memory.

The remainder of the paper is organized as follows. In Section 2, based on the proposed approach, we give a new proof of the I-MMSE relation for discrete-time Gaussian channels, and a new proof of the classical de Bruijn's identity. We will present our extensions of the I-MMSE relation, the main results in this paper, in Sections 3 and 4, which will be followed by an outlook for some promising future directions in Section 5.

2 New Proofs of Existing Results

In this section, to further illustrate the idea of our approach, we give new proofs of some existing results: the original I-MMSE relation in (2) and the classical de Bruijn's identity. To enhance the readability and emphasize the main idea, here and throughout the paper, we omit some technical details of checking the conditions required for the interchanges of differentiation and integration, which will be provided in the Appendices.

2.1 A new proof of the I-MMSE relation

In this section, we consider the Gaussian channel specified in (1) and give a new proof of (2). Here and throughout the paper, we replace $\sqrt{snr}$ with ρ to avoid notational cumbersomeness during the computation; the derivative with respect to snr can be readily obtained with an application of the chain rule. Then, under the new notation, the channel (1) becomes

$$Y = \rho X + Z,$$

where $\rho \in \mathbb{R}_+$, and we only have to prove that

$$\frac{d}{d\rho} I(X;Y) = \rho \mathbb{E}[(X - \mathbb{E}[X|Y])^2]. \tag{4}$$

Obviously, the conditional density of Y given $X = x$ by $f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} e^{-(y-\rho x)^2/2}$, and the density function of Y can be computed as

$$f_Y(y) = \int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x) dx.$$

It follows from the assumption that the channel is memoryless that

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(Z),$$

which, together with the fact that Z does not depend on ρ, implies that

$$\frac{d}{d\rho} I(X;Y) = -\frac{d}{d\rho} \mathbb{E}[\log f_Y(Y)] = -\mathbb{E}\left[\frac{1}{f_Y(Y)} \frac{d}{d\rho} f_Y(Y)\right].$$

Now, some straightforward computations yield

$$\begin{aligned}
\frac{d}{d\rho} f_Y(Y) &= \frac{d}{d\rho} \int_{\mathbb{R}} f_{Y|X}(Y|x) f_X(x) dx \\
&= -\int_{\mathbb{R}} (\rho X + Z - \rho x)(X - x) f_{Y|X}(Y|x) f_X(x) dx \\
&= -f_Y(Y) \int_{\mathbb{R}} (\rho X + Z - \rho x)(X - x) f_{X|Y}(x|Y) dx.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\frac{d}{d\rho} I(X;Y) &= \mathbb{E}\left[\int_{\mathbb{R}} (Y - \rho x)(X - x) f_{X|Y}(x|Y) dx\right] \\
&= \mathbb{E}[YX - Y\mathbb{E}[X|Y] - \rho X \mathbb{E}[X|Y] + \rho \mathbb{E}[X^2|Y]] \\
&= \mathbb{E}[YX] - \mathbb{E}[YX] - \mathbb{E}[\rho \mathbb{E}^2[X|Y]] + \mathbb{E}[\rho \mathbb{E}[X^2|Y]] \\
&= \rho \mathbb{E}[X^2 - \mathbb{E}^2[X|Y]] \\
&= \rho \mathbb{E}[(X - \mathbb{E}[X|Y])^2],
\end{aligned}$$

as desired.

2.2 A new proof of de Bruijn's identity.

The following de Bruijn's identity is a fundamental relationship between the differential entropy and the Fisher information. Based on the proposed approach, we will give a new proof of this classical result.

Theorem 2.1. *Let X be any random variable with a finite variance and let Z be an independent standard normally distributed random variable. Then, for any $t > 0$,*

$$\frac{d}{dt} H(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z), \tag{5}$$

where $J(\cdot)$ is the Fisher information.

Proof. First of all, define

$$Y = X + \sqrt{t}Z,$$

whose density function can be computed as

$$f_Y(y) = \int_{\mathbb{R}} f_X(x) f_{Y|X}(y|x)dx = \int_{\mathbb{R}} \frac{f_X(x)}{\sqrt{2\pi t}} e^{-(y-x)^2/(2t)} dx.$$

Immediately, we have

$$f_Y(Y) = f_Y(X + \sqrt{t}Z) = \int_{\mathbb{R}} \frac{f_X(x)}{\sqrt{2\pi t}} e^{-(X+\sqrt{t}Z-x)^2/(2t)} dx.$$

Now, taking the derivative with respect to t, we obtain

$$\begin{aligned}
\frac{d}{dt} f_Y(Y) &= \int_{\mathbb{R}} \frac{f_X(x)}{\sqrt{2\pi t}} e^{-(X+\sqrt{t}Z-x)^2/(2t)} \left(\frac{(X-x)(X+\sqrt{t}Z-x)}{2t^2} - \frac{1}{2t} \right) dx \\
&= \int_{\mathbb{R}} \left(\frac{(X-x)(X+\sqrt{t}Z-x)}{2t^2} - \frac{1}{2t} \right) f_{Y|X}(Y|x) f_X(x) dx \\
&= f_Y(Y) \int_{\mathbb{R}} \left(\frac{(X-x)(Y-x)}{2t^2} + \frac{1}{2t} \right) f_{X|Y}(x|Y) dx.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\frac{d}{dt} H(Y) &= -\frac{d}{dt} \mathbb{E}\left[\log f_Y(Y)\right] = -\mathbb{E}\left[\frac{1}{f_Y(Y)} \frac{d}{dt} f_Y(Y) \right] \\
&= \mathbb{E}\left[\int_{\mathbb{R}} \left(-\frac{(X-x)(Y-x)}{2t^2} + \frac{1}{2t} \right) f_{X|Y}(x|Y) dx \right] \\
&= \frac{\mathbb{E}[-XY + (X+Y)\mathbb{E}[X|Y] - \mathbb{E}[X^2|Y]]}{2t^2} + \frac{1}{2t} \\
&= \frac{-\mathbb{E}[X^2] + \mathbb{E}[\mathbb{E}^2[X|Y]]}{2t^2} + \frac{1}{2t}.
\end{aligned} \tag{6}$$

On the other hand, similarly as above,

$$f_Y'(Y) = \int_{\mathbb{R}} \frac{f_X(x)}{\sqrt{2\pi t}} e^{-(Y-x)^2/(2t)} \frac{x-Y}{t} dx = f_Y(Y) \int_{\mathbb{R}} \frac{x-Y}{t} f_{X|Y}(x|Y) dx,$$

It then follows that the right hand side of (5) can be computed as

$$\begin{aligned}
J(Y) &= \mathbb{E}\left[\left(\frac{f_Y'(Y)}{f_Y(Y)} \right)^2 \right] \\
&= \frac{\mathbb{E}[\mathbb{E}^2[X|Y] + Y^2 - 2\mathbb{E}[X|Y]Y]}{t^2} \\
&= \frac{\mathbb{E}[\mathbb{E}^2[X|Y]] + \mathbb{E}[Y^2] - 2\mathbb{E}[XY]}{t^2},
\end{aligned}$$

which, by the fact that $t = \mathbb{E}[(X-Y)^2]$, is equal to (6), the left hand side of (5). The theorem then immediately follows. $\square$

Remark 2.2. The new proof of de Bruijn's identity actually further reveals that

$$\frac{d}{dt}H(X+\sqrt{t}Z) = \frac{1}{2}J(X+\sqrt{t}Z) = \frac{1}{t^2}\mathbb{E}[(Y-\mathbb{E}[X|Y])^2].$$

3 The Extended I-MMSE Relation in Discrete Time

In this section, using the ideas and techniques illustrated in Section 2, we give extensions of the I-MMSE relation (2) to channels with feedback or memory.

We start with the following general theorem on a discrete-time system:

Theorem 3.1. *Consider the following discrete-time system*

$$Y_i = \rho g_i(W_1^i, Y_1^{i-1}) + Z_i, \quad i = 1, \dots, n, \tag{7}$$

where $\rho \in \mathbb{R}_+$, all W_i are independent of all Z_i, which are i.i.d. standard normal random variables and each $g_i(\cdot,\cdot)$ is a continuous function differentiable in its second parameter. Assume that for any i and any compact subset $K \subset \mathbb{R}_+$,

$$\mathbb{E}\left[\sup_{\rho \in K} g_i^2(W_1^i, Y_1^{i-1})\right] < \infty, \quad \mathbb{E}\left[\sup_{\rho \in K}\left(\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})\right)^2\right] < \infty. \tag{8}$$

Then we have

$$\frac{d}{d\rho}I(W_1^n; Y_1^n) = \rho\sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}[g_i|Y_1^n])^2\right] + \rho^2 \sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}[g_i|Y_1^n])\frac{d}{d\rho}g_i\right], \tag{9}$$

where we have written $g_i(W_1^i, Y_1^{i-1})$ simply as g_i.

Proof. Note that

$$I(W_1^n; Y_1^n) = H(Y_1^n) - \sum_{i=1}^n H(Y_i|W_1^n, Y_1^{i-1}) = H(Y_1^n) - nH(Z_1),$$

which immediately implies

$$\frac{d}{d\rho}I(W_1^n; Y_1^n) = -\mathbb{E}\left[\frac{d}{d\rho}\log f_{Y_1^n}(Y_1^n)\right] = -\mathbb{E}\left[\frac{1}{f_{Y_1^n}(Y_1^n)}\frac{d}{d\rho}f_{Y_1^n}(Y_1^n)\right].$$

In the remainder of the proof, we will omit the subscripts of the density functions. For instance, $f(y_1^n)$ means the density function of Y_1^n, $f(Y_1^n)$ means the density function of Y_1^n evaluated at Y_1^n, $f(y_1^n|w_1^n)$ means the conditional density function of Y_1^n given $W_1^n = w_1^n$.

Under the system assumptions, we have

$$f(y_1^n|w_1^n) = \prod_{i=1}^n f(y_i|y_1^{i-1}, w_1^n) = \frac{1}{(\sqrt{2\pi})^n}\prod_{i=1}^n \exp\{-(y_i - \rho g_i(w_1^i, y_1^{i-1}))^2/2\},$$

and furthermore,

$$\begin{aligned}\frac{d}{d\rho}f(Y_1^n|w_1^n) &= \frac{1}{(\sqrt{2\pi})^n}\frac{d}{d\rho}\prod_{i=1}^n \exp\{-(Y_i-\rho g_i(w_i,Y_1^{i-1}))^2/2\}\\ &= \frac{1}{(\sqrt{2\pi})^n}\frac{d}{d\rho}\prod_{i=1}^n \exp\{-(\rho g_i(W_1^i,Y_1^{i-1})-\rho g_i(w_1^i,Y_1^{i-1})+Z_i)^2/2\}\\ &= -f(Y_1^n|w_1^n)\sum_{i=1}^n (Y_i-\rho g_i(w_1^i,Y_1^{i-1}))\bigg(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1})\\ &\qquad +\rho\frac{d}{d\rho}(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1}))\bigg).\end{aligned}$$

It then follows that

$$\begin{aligned}\frac{d}{d\rho}f(Y_1^n) &= \frac{d}{d\rho}\int_{\mathbb{R}^n} f(Y_1^n|w_1^n)f(w_1^n)dw_1^n\\ &= \int_{\mathbb{R}^n}\frac{d}{d\rho}f(Y_1^n|w_1^n)f(w_1^n)dw_1^n\\ &= -\int_{\mathbb{R}^n}\sum_{i=1}^n (Y_i-\rho g_i(w_1^i,Y_1^{i-1}))\bigg(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1})\\ &\qquad +\rho\frac{d}{d\rho}(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1}))\bigg)f(Y_1^n|w_1^n)f(w_1^n)dw_1^n\\ &= -f(Y_1^n)\int_{\mathbb{R}^n}\sum_{i=1}^n (Y_i-\rho g_i(w_1^i,Y_1^{i-1})\bigg(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1})\\ &\qquad +\rho\frac{d}{d\rho}(g_i(W_1^i,Y_1^{i-1})-g_i(w_1^i,Y_1^{i-1}))\bigg)f(w_1^n|Y_1^n)dw_1^n.\end{aligned}$$

Writing $g_i(W_1^i,Y_1^{i-1}), g_i(w_1^i,Y_1^{i-1})$ as $g_i, \tilde{g}_i$, respectively, and using the fact that for any measurable function φ,

$$\int_{\mathbb{R}^n}\varphi(w_1^n,Y_1^n)f(w_1^n|Y_1^n)dw_1^n = \mathbb{E}[\varphi(W_1^n,Y_1^n)|Y_1^n],$$

we further compute

$$\begin{aligned}\frac{d}{d\rho}f(Y_1^n) &= -f(Y_1^n)\sum_{i=1}^n\int_{\mathbb{R}^n}(Y_i-\rho\tilde{g}_i)\left((g_i+\rho\frac{d}{d\rho}g_i)-(\tilde{g}_i+\rho\frac{d}{d\rho}\tilde{g}_i)\right)f(w_1^n|Y_1^n)dw_1^n\\ &= -f(Y_1^n)\sum_{i=1}^n\left((g_i+\rho\frac{d}{d\rho}g_i)\mathbb{E}\left[(Y_i-\rho g_i)|\,Y_1^n\right]-\mathbb{E}\left[(g_i+\rho\frac{d}{d\rho}g_i)(Y_i-\rho g_i)\middle|\,Y_1^n\right]\right).\end{aligned}$$

Similarly continue as in the proof of (4), we eventually obtain

$$\begin{aligned}\frac{d}{d\rho}I(W_1^n;Y_1^n) &= \sum_{i=1}^n\left(\mathbb{E}\left[(g_i+\rho\frac{d}{d\rho}g_i)(Y_i-\rho\mathbb{E}\left[g_i|\,Y_1^n\right])\right]-\mathbb{E}\left[(g_i+\rho\frac{d}{d\rho}g_i)(Y_i-\rho g_i)\right]\right)\\ &= \rho\sum_{i=1}^n\mathbb{E}\left[(g_i-\mathbb{E}(g_i|Y_1^n))^2\right]+\rho^2\sum_{i=1}^n\mathbb{E}\left[(g_i-\mathbb{E}(g_i|Y_1^n))\frac{d}{d\rho}g_i\right],\end{aligned}$$

as desired. $\square$

Remark 3.2. Theorem 3.1 still holds if each g_i is a Lebesgue measurable function (again, differentiable in its second parameter) instead, which, however, is less relevant to practical engineering applications.

Remark 3.3. It can readily checked that

$$\mathbb{E}\left[(g_i - \mathbb{E}(g_i|Y_1^n))\,\frac{d}{d\rho}g_i\right] = \mathbb{E}\left[(g_i - \mathbb{E}(g_i|Y_1^n))\left(\frac{d}{d\rho}g_i - E\left[\frac{d}{d\rho}g_i\middle|\,Y_1^n\right]\right)\right],$$

which means that (9) can be rewritten in the following more symmetric form:

$$\frac{d}{d\rho}I(W_1^n;Y_1^n) = \rho\sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}(g_i|Y_1^n))^2\right] + \rho^2\sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}(g_i|Y_1^n))\left(\frac{d}{d\rho}g_i - E\left[\frac{d}{d\rho}g_i\middle|\,Y_1^n\right]\right)\right].$$

Remark 3.4. Consider the discrete-time system as in (7). Rewriting all W_i as M and each g_i as X_i, we then have the following discrete-time Gaussian channel with feedback:

$$Y_i = \sqrt{snr}X_i(M, Y_1^{i-1}) + Z_i, \quad i = 1, 2, \ldots, n$$

where M is interpreted as the message be transmitted and X_i, Y_i are the channel inputs, outputs, respectively. It is well known that for such a feedback channel,

$$I(X_1^n \to Y_1^n) = I(M; Y_1^n),$$

where $I(X_1^n \to Y_1^n)$ is the directed information between X_1^n and Y_1^n. Then, applying Theorem 3.1 and the chain rule for taking derivative, we have

$$\frac{d}{dsnr}I(X_1^n \to Y_1^n) = \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i|Y_1^n])^2\right] + snr\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i|Y_1^n])\,\frac{d}{dsnr}X_i\right], \tag{10}$$

where $X_i = X_i(M, Y_1^{i-1})$. This yields an extension of the I-MMSE relation to discrete-time Gaussian channels with feedback.

Remark 3.5. Alternatively, rewriting each W_i as X_i, we will have the following discrete-time Gaussian channel with input and output memory (it is observed that such a channel is suitable for modeling some storage systems, such as flash memories [1]):

$$Y_i = \sqrt{snr}g_i(X_1^i, Y_1^{i-1}) + Z_i, \quad i = 1, 2, \ldots, n$$

where g_i is interpreted as "part" of the channel and X_i, Y_i are the channel inputs, outputs, respectively. Then, by Theorem 3.1 and the chain rule, we obtain

$$\frac{d}{dsnr}I(X_1^n;Y_1^n) = \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[g_i - \mathbb{E}[g_i|Y_1^n])^2\right] + snr\sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}[g_i|Y_1^n])\,\frac{d}{dsnr}g_i\right], \tag{11}$$

where $g_i = g_i(X_1^i, Y_1^{i-1})$. This yields an extension of the I-MMSE relation to discrete-time Gaussian channels with input and output memory.

4 The Extended I-MMSE Relation in Continuous Time

As elaborated in the following theorem, the continuous-time I-MMSE relation, the continuous-time analog of (2), has been established in [14].

Theorem 4.1 (Theorem 6 of [14]). *Consider the following continuous-time Gaussian channel*

$$Y(t) = \sqrt{snr} \int_0^t X(s) ds + B(t), \quad t \in [0, T],$$

where $\{X(s)\}$ is the channel input satisfying the power constraint

$$\int_0^T \mathbb{E}[X^2(s)] ds < \infty \tag{12}$$

and $\{B(t)\}$ is the standard Brownian motion. Then, we have

$$\frac{d}{dsnr} I(W_0^T; Y_0^T) = \frac{1}{2} \int_0^T \mathbb{E}[\big(X(s) - \mathbb{E}[X(s)|Y_0^T]\big)^2] ds. \tag{13}$$

In this section, using the ideas and techniques illustrated in Section 2, we give extensions of the continuous-time I-MMSE relation to channels with feedback or memory.

We start with a general theorem on a continuous-time system:

Theorem 4.2. *Consider a continuous-time system characterized by the following stochastic differential equation:*

$$Y(t) = \rho \int_0^t g(s, W_0^s, Y_0^s) ds + B(t), \quad t \in [0, T], \tag{14}$$

where $\rho \in \mathbb{R}_+$, the continuous random process $\{W(t)\}$ is independent of the standard Brownian motion $\{B(t)\}$, and $g(\cdot, \cdot, \cdot)$ is a deterministic function. Assume that

(a) $g(s, \gamma_0^s, \phi_0^s)$ is defined for all $\gamma(\cdot), \phi(\cdot) \in C[0, T]$, the set of all continuous functions over $[0, T]$, and is itself a continuous function in s, $s \in [0, T]$;

(b) the solution $\{Y(t)\}$ to the stochastic differential equation (14) uniquely exists;

(c) for any $s \in [0, T]$, $g(s, W_0^s, Y_0^s)$ is continuously differentiable with respect to ρ with probability 1;

(d) for any compact subset $K \subset \mathbb{R}_+$, we have

$$\int_0^T \mathbb{E}\left[\sup_{\rho \in K} g^2(s, W_0^s, Y_0^s)\right] ds < \infty, \quad \int_0^T \mathbb{E}\left[\sup_{\rho \in K} \left(\frac{d}{d\rho} g(s, W_0^s, Y_0^s)\right)^2\right] ds < \infty;$$

(e) $g(s, \gamma_0^s, \phi_0^s)$ is uniformly bounded over all $s \in [0, T]$ and all $\gamma(\cdot), \phi(\cdot) \in C[0, T]$.

Then, we have

$$\frac{d}{d\rho}I(W_0^T;Y_0^T)=\rho\int_0^T\mathbb{E}[\big(g(s)-\mathbb{E}[g(s)|Y_0^T]\big)^2]ds+\rho^2\int_0^T\mathbb{E}\left[\big(g(s)-\mathbb{E}\left[g(s)|\,Y_0^T\right]\big)\frac{d}{d\rho}g(s)\right]ds,$$

where we have written $g(s,W_0^s,Y_0^s)$ *simply as* $g(s)$.

Strictly speaking, Theorem 4.2 is not a generalization of Theorem 4.1: Condition (e) is stronger than the square integrability condition (12), as one can easily find g satisfying the latter but not the former. As elaborated in the following theorem, at the expense of an extra yet mild condition (see (f) in the following theorem), Condition (e) can be relaxed to an integrability condition (see (g)).

Theorem 4.3. *Consider the continuous-time system (14) satisfying Conditions (a), (b), (c), (d) and the following conditions:*

(f) for any $a>0$ *and any* $t\in[0,T]$,

$$P\left(\int_0^t g^2(s,W_0^s,Y_0^s)ds=a\right)=0;$$

(g) with probability 1, *we have (note that the third parameter in the following* g *function is* B_0^s, *rather than* Y_0^s*)*

$$\int_0^T g^2(s,W(s),B_0^s)ds<\infty.$$

Then, we have

$$\frac{d}{d\rho}I(W_0^T;Y_0^T)=\rho\int_0^T\mathbb{E}[\big(g(s)-\mathbb{E}[g(s)|Y_0^T]\big)^2]ds+\rho^2\int_0^T\mathbb{E}\left[\big(g(s)-\mathbb{E}\left[g(s)|\,Y_0^T\right]\big)\frac{d}{d\rho}g(s)\right]ds, \tag{15}$$

where we have written $g(s,W_0^s,Y_0^s)$ *simply as* $g(s)$.

Remark 4.4. Similarly as in Remark 3.3, we can obtain the following more symmetric formula:

$$\frac{d}{d\rho}I(W_0^T;Y_0^T)=\rho\int_0^T\mathbb{E}[(g(s)-\mathbb{E}[g(s)|Y_0^T])^2]ds+\rho^2\int_0^T\mathbb{E}\left[(g(s)-\mathbb{E}[g(s)|Y_0^T])\left(\frac{d}{d\rho}g(s)-E\left[\frac{d}{d\rho}g(s)\middle|\,Y_0^T\right]\right)\right]ds.$$

Remark 4.5. Parallel to Remarks 3.4, the continuous-time system in (14) can be interpreted as the following continuous-time Gaussian channel with feedback:

$$Y(t)=\sqrt{snr}\int_0^t X(s,M,Y_0^s)ds+B(t),\quad t\in[0,T].$$

An application of Theorem 4.2 then yields

$$\frac{d}{dsnr}I(M;Y_0^T)=\frac{1}{2}\int_0^T\mathbb{E}[\big(X(s)-\mathbb{E}[X(s)|Y_0^T]\big)^2]ds+snr\int_0^T\mathbb{E}\left[\big(X(s)-\mathbb{E}\left[X(s)|\,Y_0^T\right]\big)\frac{d}{dsnr}X(s)\right]ds, \tag{16}$$

where $X(s)$ is the abbreviated form of $X(s, M, Y_0^s)$. This gives an extension of the I-MMSE relation to continuous-time Gaussian channels with feedback.

Parallel to Remark 3.5, it can be also interpreted as the following continuous-time Gaussian channel with input and output memory:

$$Y(t) = \sqrt{snr} \int_0^t g(s, X_0^s, Y_0^s) ds + B(t), \quad t \in [0, T].$$

An application of Theorem 4.2 then yields

$$\frac{d}{dsnr} I(X_0^T; Y_0^T) = \frac{1}{2} \int_0^T \mathbb{E}[\left(g(s) - \mathbb{E}[g(s)|Y_0^T]\right)^2] ds + snr \int_0^T \mathbb{E}\left[\left(g(s) - \mathbb{E}\left[g(s)|\, Y_0^T\right]\right) \frac{d}{dsnr} g(s)\right] ds, \tag{17}$$

where $g(s)$ is the abbreviated form of $g(s, X_0^s, Y_0^s)$. This gives an extension of the I-MMSE relation to continuous-time Gaussian channels with input and output memory.

Remark 4.6. It can be readily verified that Theorem 4.3, when interpreted as in the previous remark, includes Theorem 4.1 as a special case; see more detailed explanations in Remark 4.8.

4.1 Properties of the solution to (14)

In this section, we will give certain sufficient conditions that will guarantee the solution Y to (14) uniquely exists (Condition (b) in Theorem 4.2), and moreover, $g(s, W_0^s, Y_0^s)$ is differentiable with respect to ρ (Condition (c) in Theorem 4.2). More precisely, we have the following proposition.

Proposition 4.7. *Under the following conditions:*

- *$Dg(s, \gamma_0^s, \phi_0^s)$, the Frechet derivative of g with respect to its third parameter $\phi(\cdot)$, exists for any $s \in [0, T]$ and any $\gamma(\cdot), \phi(\cdot) \in C[0, T]$;*
- *(extended uniform Lipschitz conditions) There exists a constant C such that for all $s \in [0, T]$ and all $\gamma(\cdot), \phi(\cdot), \psi(\cdot) \in C[0, T]$, we have*

$$|g(s, \gamma_0^s, \phi_0^s) - g(s, \gamma_0^s, \psi_0^s)| \le C\|\phi_0^s - \psi_0^s\|_\infty,$$

and

$$\|Dg(s, \gamma_0^s, \phi_0^s) - Dg(s, \gamma_0^s, \psi_0^s)\| \le C\|\phi_0^s - \psi_0^s\|_\infty;$$

- *(extended linear growth conditions) There exists a constant C such that for all $s \in [0, T]$ and all $\gamma(\cdot), \phi(\cdot) \in C[0, T]$, we have*

$$g^2(s, \gamma_0^s, \phi_0^s) \le C(1 + \|\gamma_0^s\|_\infty^2 + \|\phi_0^s\|_\infty^2),$$

and

$$\|Dg(s, \gamma_0^s, \phi_0^s)\|^2 \le C(1 + \|\gamma_0^s\|_\infty^2 + \|\phi_0^s\|_\infty^2),$$

the solution Y to the continuous-time system (14) uniquely exists, and moreover, with probability 1, $g(s, W_0^s, Y_0^s)$ is differentiable with respect to ρ.

Proof. We only sketch the proof, as it is essentially the standard argument for the existence and uniqueness of the solution to a stochastic differential equation with the well-known uniform Lipschitz and linear growth conditions; see, e.g., the proof of Theorem 2.2 in Chapter 5 of [29].

Consider the following Picard's iteration:

$$Y_{(0)}(t) \equiv 0, \quad Y_{(n+1)}(t) = \int_0^t g(s, W_0^s, Y_{(n),0}^s)ds + B(t), \quad t \in [0, T].$$

It can be easily verified that, for any n and any $t \in [0, T]$, $Y_{(n)}(t)$ is differentiable with respect to ρ. Letting $Z_{(n)}(t) = \frac{d}{d\rho} Y_{(n)}(t)$ for all n, we have

$$Z_{(0)}(t) \equiv 0, \quad Z_{(n+1)}(t) = \int_0^t g(s, W_0^s, Y_{(n),0}^s)ds + \rho \int_0^t Dg(s, W_0^s, Y_{(n),0}^s)(Z_{(n),0}^s)ds, \quad t \in [0, T].$$

Now, applying the standard argument for the existence and uniqueness of the solution to a stochastic differential equation, we deduce that there exists a stochastic process $\{Y(t), t \in [0, T]\}$ such that for any compact set $K \subset \mathbb{R}_+$,

$$\lim_{n \to \infty} \sup_{\rho \in K, t \in [0,T]} |Y_n(t) - Y(t)| = 0, \quad \text{a.s.}$$

and furthermore, there exists a stochastic process $Z(t), t \in [0, T]$ such that for any compact set $K \subset \mathbb{R}_+$,

$$\lim_{n \to \infty} \sup_{\rho \in K, t \in [0,T]} |Z_n(t) - Z(t)| = 0, \quad \text{a.s.}$$

It then follows that $Y(t)$ is differentiable with respect to ρ, and $\frac{d}{d\rho} Y(t) = Z(t)$ with probability 1, and consequently, $g(s, W_0^s, Y_0^s)$ is differentiable with respect to ρ. □

4.2 Proof of Theorem 4.2

Fix $W = w$ and let $Y_{|w)}$ be such that

$$Y_{|w)}(t) = \rho \int_0^t g(s, w_0^s, Y_{|w),0}^s)ds + B(t), \quad t \in [0, T].$$

Then, by Theorem 7.1 of [25] (it can be checked that its assumptions are implied by Condition (e)), we observe that $\mu_{Y_{|w)}} \sim \mu_B \sim \mu_Y$, where "$\sim$" means "equivalent", and furthermore,

$$\frac{d\mu_{Y_{|w)}|W}}{d\mu_B}(Y_{|w),0}^T | w_0^T) = \exp\left\{\rho \int_0^T g(s, w_0^s, Y_{|w),0}^s)dY_{|w)}(s) - \frac{\rho^2}{2} \int_0^T g^2(s, w_0^s, Y_{|w),0}^s)ds\right\}.$$

It then follows from Lemma 4.10 in [25] that

$$\frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T | w_0^T) = \exp\left\{\rho \int_0^T g(s, w_0^s, Y_0^s)dY(s) - \frac{\rho^2}{2} \int_0^T g^2(s, w_0^s, Y_0^s)ds\right\}.$$

Note that, by definition, we have

$$
\begin{aligned}
I(W_0^T; Y_0^T) &= \mathbb{E}\left[\log \frac{d\mu_{WY}}{d(\mu_W \times \mu_Y)}(W_0^T, Y_0^T)\right] \\
&= \mathbb{E}\left[\log \frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T|W_0^T)\right] - \mathbb{E}\left[\log \frac{d\mu_Y}{d\mu_B}(Y_0^T)\right] \\
&= \frac{\rho^2}{2}\int_0^T \mathbb{E}[g^2(s)]ds - \mathbb{E}\left[\log \frac{d\mu_Y}{d\mu_B}(Y_0^T)\right].
\end{aligned}
$$

Taking derivative with respect to ρ then yields

$$
\begin{aligned}
\frac{d}{d\rho} I(W_0^T; Y_0^T) &= \rho \int_0^T \mathbb{E}[g^2(s)]ds + \frac{\rho^2}{2}\frac{d}{d\rho}\int_0^T \mathbb{E}[g^2(s)]ds - \frac{d}{d\rho}\mathbb{E}\left[\log \frac{d\mu_Y}{d\mu_B}(Y_0^T)\right] \\
&= \rho \int_0^T \mathbb{E}[g^2(s)]ds + \rho^2 \int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\right] ds - \frac{d}{d\rho}\mathbb{E}\left[\log \frac{d\mu_Y}{d\mu_B}(Y_0^T)\right].
\end{aligned}
$$

Writing $g(s, w_0^s, Y_0^s)$ as $\tilde{g}(s)$, we have

$$
\begin{aligned}
\frac{d}{d\rho}\left(\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right) &= \frac{d}{d\rho}\int \frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T|w_0^T)\mu_W(dw) \\
&= \frac{d}{d\rho}\int \exp\left\{\rho\int_0^T \tilde{g}(s)dY(s) - \frac{\rho^2}{2}\int_0^T \tilde{g}^2(s)ds\right\}\mu_W(dw) \\
&= \frac{d}{d\rho}\int \exp\left\{\rho^2\int_0^T \tilde{g}(s)g(s)ds + \rho\int_0^T \tilde{g}(s)dB(s) - \frac{\rho^2}{2}\int_0^T \tilde{g}^2(s)ds\right\}\mu_W(dw) \\
&= \int \Bigg(\int_0^T \tilde{g}(s)dY(s) + \rho\int_0^T \frac{d}{d\rho}\tilde{g}(s)dY(s) + \rho\int_0^T \tilde{g}(s)(g(s)-\tilde{g}(s))ds \\
&\qquad + \rho^2\int_0^T \tilde{g}(s)\frac{d}{d\rho}\left(g(s)-\tilde{g}(s)\right)ds\Bigg)\frac{d\mu_{WY}}{d\mu_B}(dw, Y_0^T) \\
&= \frac{d\mu_Y}{d\mu_B}(Y_0^T)\int \Bigg(\int_0^T \tilde{g}(s)dY(s) + \rho\int_0^T \frac{d}{d\rho}\tilde{g}(s)dY(s) + \rho\int_0^T \tilde{g}(s)(g(s)-\tilde{g}(s))ds \\
&\qquad + \rho^2\int_0^T \tilde{g}(s)\frac{d}{d\rho}\left(g(s)-\tilde{g}(s)\right)ds\Bigg)\mu_{W|Y}(dw|Y_0^T) \\
&= \frac{d\mu_Y}{d\mu_B}(Y_0^T)\Bigg(\mathbb{E}\left[\int_0^T g(s)dY(s)\middle| Y_0^T\right] + \rho\mathbb{E}\left[\int_0^T \frac{d}{d\rho}g(s)dY(s)\middle| Y_0^T\right] \\
&\quad + \rho\int_0^T (\mathbb{E}[g(s)|Y_0^T]g(s) - \mathbb{E}[g^2(s)|Y_0^T])ds \\
&\quad + \rho^2\int_0^T \left(\frac{d}{d\rho}g(s)\mathbb{E}[g(s)|Y_0^T] - \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\middle| Y_0^T\right]\right)ds\Bigg).
\end{aligned}
$$

Note that by the properties of conditional expectation and Itô integral, we have

$$
\mathbb{E}\left[\mathbb{E}\left[\int_0^T g(s)dY(s)\middle| Y_0^T\right]\right] = \mathbb{E}\left[\int_0^T g(s)dY(s)\right] = \rho\int_0^T \mathbb{E}[g^2(s)]ds,
$$

and similarly,

$$\mathbb{E}\left[\int_0^T \mathbb{E}[g^2(s)|Y_0^T]ds\right] = \int_0^T \mathbb{E}[g^2(s)]ds,$$

and

$$\rho\mathbb{E}\left[\mathbb{E}\left[\int_0^T \frac{d}{d\rho}g(s)dY(s)\middle| Y_0^T\right]\right] = \rho^2\int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\right]ds = \rho^2\mathbb{E}\left[\int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\middle| Y_0^T\right]ds\right].$$

It then follows that

$$\begin{aligned}
\frac{d}{d\rho}\mathbb{E}\left[\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right] &= \mathbb{E}\left[\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right] \\
&= \mathbb{E}\left[\frac{d}{d\rho}\left(\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right)\Big/\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right] \\
&= \mathbb{E}\Bigg[\mathbb{E}\left[\int_0^T g(s)dY(s)\middle| Y_0^T\right] + \rho\mathbb{E}\left[\int_0^T \frac{d}{d\rho}g(s)dY(s)\middle| Y_0^T\right] \\
&\quad + \rho\int_0^T(\mathbb{E}[g(s)|Y_0^T]g(s) - \mathbb{E}[g^2(s)|Y_0^T])ds + \rho^2\int_0^T\left(\frac{d}{d\rho}g(s)\mathbb{E}[g(s)|Y_0^T] - \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\middle| Y_0^T\right]\right)ds\Bigg] \\
&= \rho\int_0^T \mathbb{E}[\mathbb{E}[g(s)|Y_0^T]g(s)]ds + \rho^2\int_0^T \mathbb{E}\left[\mathbb{E}[g(s)|Y_0^T]\frac{d}{d\rho}g(s)\right]ds.
\end{aligned}$$

So we have

$$\begin{aligned}
\frac{d}{d\rho}I(W_0^T;Y_0^T) &= \rho\int_0^T \mathbb{E}[g^2(s)]ds + \rho^2\int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\right]ds \\
&\quad - \rho\int_0^T \mathbb{E}[\mathbb{E}[g(s)|Y_0^T]g(s)]ds - \rho^2\int_0^T \mathbb{E}\left[\mathbb{E}[g(s)|Y_0^T]\frac{d}{d\rho}g(s)\right]ds \\
&= \rho\int_0^T \mathbb{E}[(g(s) - \mathbb{E}[g(s)|Y_0^T])^2]ds + \rho^2\int_0^T \mathbb{E}\left[(g(s) - \mathbb{E}[g(s)|Y_0^T])\frac{d}{d\rho}g(s)\right]ds,
\end{aligned}$$

as desired.

4.3 Proof of Theorem 4.3

The proof consists of the following 6 steps:

Step 1. First of all, for any fixed $W = w$, by Theorem 7.7 of [25], $\mu_{Y|W=w} \sim \mu_B$ with

$$\frac{d\mu_{Y|W=w}}{d\mu_B}(B_0^T) = \exp\left(\int_0^T g(s, w_0^s, B_0^s)dB(s) - \frac{1}{2}\int_0^T g^2(s, w_0^s, B_0^s)ds\right),$$

where we have used Conditions (d) and (g) before invoking Theorem 7.7. Moreover, by Condition (d), it follows from Theorem 7.2 that $\mu_Y \ll \mu_B$ with

$$\begin{aligned}
\frac{d\mu_Y}{d\mu_B}(B_0^T) &= \int \frac{d\mu_{Y|W=w}}{d\mu_B}(B_0^T)d\mu_W(w) \\
&= \int \exp\left(\int_0^T g(s, w_0^s, B_0^s)dB(s) - \frac{1}{2}\int_0^T g^2(s, w_0^s, B_0^s)ds\right)d\mu_W(w),
\end{aligned}$$

which is obviously positive with probability 1. It then follows from Lemma 6.8 of [25] that $\mu_B \ll \mu_Y$. So, in this step, we have shown that under the conditions specified in theorem, we have $\mu_Y \sim \mu_{Y|W=w} \sim \mu_B$.

Step 2. For any n and $\gamma(\cdot), \phi(\cdot) \in C[0,T]$, we follow [25] and define a truncated version of g as follows:

$$g_{(n)}(t, \gamma_0^t, \phi_0^t) = g(t, \gamma_0^t, \phi_0^t) \mathbf{1}_{\int_0^t g^2(s,\gamma_0^t,\phi_0^s)ds<n}.$$

Now, define a truncated version of Y as follows:

$$Y_{(n)}(t) = \rho \int_0^t g_{(n)}(s, W_0^s, Y_0^s) ds + B(t), \quad t \in [0,T],$$

which, as elaborated on Page 265 in [25], can be rewritten as

$$Y_{(n)}(t) = \rho \int_0^t g_{(n)}(s, W_0^s, Y_{(n),0}^s) ds + B(t), \quad t \in [0,T].$$

It is well known that (see, e.g., Theorem 6.2.1 of [22]) that

$$I(W_0^T; Y_{(n),0}^T) = \frac{1}{2} \int_0^T \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^s)] - \mathbb{E}[\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^s]]ds,$$

and

$$I(W_0^T; Y_0^T) = \frac{1}{2} \int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)] - \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^s]]ds.$$

Moreover, it follows from Theorem 4.2 (here, note that extra yet minor care has to be taken since $g_{(n)}(s, W_0^s, Y_{(n),0}^s)$ is only a piecewise differentiable function in ρ; cf. Condition (c)) that

$$\begin{aligned}\frac{d}{d\rho} I(W_0^T; Y_{(n),0}^T) &= \rho \int_0^T \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^s)] - \mathbb{E}[\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]]ds \\ &+ \rho^2 \int_0^T \mathbb{E}\left[(g_{(n)}(s, W_0^s, Y_{(n),0}^s) - \mathbb{E}[g_{(n)}(s, , W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]) \frac{d}{d\rho} g_{(n)}(s, W_0^s, Y_{(n),0}^s)\right] ds.\end{aligned} \tag{18}$$

Step 3. In this step, we will prove that

$$\begin{aligned}\lim_{n\to\infty} \frac{d}{d\rho} I(W_0^T; Y_{(n),0}^T) &= \rho \int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)] - \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^T]]ds \\ &+ \rho^2 \int_0^T \mathbb{E}\left[(g(s, W_0^s, Y_0^s) - \mathbb{E}[g(s, W_0^s, Y_0^s)|Y_0^T]) \frac{d}{d\rho} g(s, W_0^s, Y_0^s)\right] ds.\end{aligned} \tag{19}$$

Step 3.1. In this step, we observe that, with Condition (d), an application of the dominated convergence theorem will yield

$$\lim_{n\to\infty} \int_0^T \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^s)]ds = \int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds.$$

Step 3.2. In this step, we will prove that

$$\lim_{n\to\infty}\int_0^T \mathbb{E}[\mathbb{E}^2[g_{(n)}(s,W_0^s,Y_{(n),0}^s)|Y_{(n),0}^T]]ds=\int_0^T \mathbb{E}[\mathbb{E}^2[g(s,W_0^s,Y_0^s)|Y_0^T]]ds. \tag{20}$$

First of all, we note that

$$\begin{aligned}\mathbb{E}[\mathbb{E}^2[g_{(n)}(s,W_0^s,Y_{(n),0}^s)|Y_{(n),0}^T]] &= \mathbb{E}\left[\left(\int g_{(n)}(s,w_0^s,Y_{(n),0}^s)\mu_{W|Y_{(n)}}(dw|Y_{(n),0}^T)\right)^2\right]\\ &=\mathbb{E}\left[\left(\int g_{(n)}(s,w_0^s,Y_{(n),0}^s)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(Y_{(n),0}^T|w_0^T)\mu_W(dw)/\frac{d\mu_{Y_{(n)}}}{d\mu_B}(Y_{(n),0}^T)\right)^2\right]\\ &=\mathbb{E}\left[\left(\int g_{(n)}(s,w_0^s,B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right)^2\times\left(\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right)^{-1}\right].\end{aligned}$$

We now proceed with the following steps:

Step 3.2.1. In this step, we prove that in probability

$$\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\to\frac{d\mu_Y}{d\mu_B}(B_0^T).$$

First of all,

$$\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)=\int\exp\left(\rho\int_0^T g_{(n)}(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g_{(n)}^2(s,w_0^s,B_0^s)ds\right)\mu_W(dw).$$

It then follows from the Itô isometry that

$$\exp\left(\rho\int_0^T g_{(n)}(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g_{(n)}^2(s,w_0^s,B_0^s)ds\right)$$

converges to

$$\exp\left(\rho\int_0^T g(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g^2(s,w_0^s,B_0^s)ds\right)$$

in probability. And moreover, it can be easily checked that

$$\mathbb{E}\left[\int\exp\left(\rho\int_0^T g_{(n)}(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g_{(n)}^2(s,w_0^s,B_0^s)ds\right)\mu_W(dw)\right]=\mathbb{E}\left[\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right]=1$$

and

$$\mathbb{E}\left[\int\exp\left(\rho\int_0^T g(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g^2(s,w_0^s,B_0^s)ds\right)\mu_W(dw)\right]=\mathbb{E}\left[\frac{d\mu_Y}{d\mu_B}(B_0^T)\right]=1.$$

It then follows from Theorem 5.5.2 of [11] that

$$\lim_{n\to\infty}\mathbb{E}\left[\int\left|\left(\exp\left(\rho\int_0^T g_{(n)}(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g_{(n)}^2(s,w_0^s,B_0^s)ds\right)\right.\right.\right.$$

$$-\exp\left(\rho\int_0^T g(s,w_0^s,B_0^s)dB(s)-\frac{\rho^2}{2}\int_0^t g^2(s,w_0^s,B_0^s)ds\right)\right)\Big|\mu_W(dw)\Big]=0,$$

which further implies that

$$\int \exp\left(\rho\int_0^T g_{(n)}(s,w_0^s,B_0^s)dB_s-\frac{\rho^2}{2}\int_0^t g_{(n)}^2(s,w_0^s,B_0^s)ds\right)\mu_W(dw)$$

converges to

$$\int \exp\left(\rho\int_0^T g(s,w_0^s,B_0^s)dB_s-\frac{\rho^2}{2}\int_0^t g^2(s,w_0^s,B_0^s)ds\right)\mu_W(dw)$$

in probability.

Step 3.2.2. In this step, we will prove that in probability

$$\int g_{(n)}(s,w_0^s,B_0^s)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\to\int g(s,w_0^s,B_0^s)\frac{d\mu_{Y|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw).$$

First of all, it is easy to check that in probability

$$g_{(n)}(s,w_0^s,B_0^s)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\to g(s,w_0^s,B_0^s)\frac{d\mu_{Y|W}}{d\mu_B}(B_0^T|w_0^T).$$

And moreover, we have

$$\mathbb{E}\left[\int|g_{(n)}(s,w_0^s,B_0^s)|\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right]=\mathbb{E}[|g_{(n)}(s,W(s),Y_{(n),0}^s)|]$$

converges to

$$\mathbb{E}[|g(s,W(s),Y_0^s)|]=\mathbb{E}\left[\int|g(s,w_0^s,B_0^s)|\frac{d\mu_{Y|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right].$$

So, similarly as in Step 3.1.1, we deduce that

$$\int g_{(n)}(s,w_0^s,B_0^s)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w)\mu_W(dw)\to\int g(s,w_0^s,B_0^s)\frac{d\mu_{Y|W}}{d\mu_B}(B_0^T|w)\mu_W(dw).$$

in probability.

Step 3.2.3. Note that Steps 3.2.1 and 3.2.2 collectively yield that

$$\left(\int g_{(n)}(s,w_0^s,B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right)^2\times\left(\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right)^{-1}$$

converges to

$$\left(\int g(s,w_0^s,B_0^T)\frac{d\mu_{Y|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right)^2\times\left(\frac{d\mu_{Y}}{d\mu_B}(B_0^T)\right)^{-1}$$

in probability. Now, applying Jensen's inequality, we have

$$\left(\int g_{(n)}(s, w_0^s, B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right)^2 \times \left(\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right)^{-1}$$
$$= \left(\int g_{(n)}(s, w_0^s, B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)/\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right)^2 \times \frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)$$
$$\leq \left(\int g_{(n)}^2(s, w_0^s, B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)/\frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)\right) \times \frac{d\mu_{Y_{(n)}}}{d\mu_B}(B_0^T)$$
$$= \int g_{(n)}^2(s, w_0^s, B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw).$$

Note that

$$\mathbb{E}\left[\int g_{(n)}^2(s, w_0^s, B_0^T)\frac{d\mu_{Y_{(n)}|W}}{d\mu_B}(B_0^T|w_0^T)\mu_W(dw)\right] = \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^s)] \to \mathbb{E}[g^2(s, W_0^s, Y_0^s)] < \infty,$$

where the finiteness is due to Condition (d). Finally, the desired (20) follows from the generalized dominated convergence theorem (see, e.g., Theorem 19 on Page 89 of [36]).

Step 3.3. In this step, we establish the following two convergences:

$$\lim_{n\to\infty}\int_0^T \mathbb{E}\left[g_{(n)}(s, W_0^s, Y_{(n),0}^s)\frac{d}{d\rho}g_{(n)}(s, W_0^s, Y_{(n),0}^s)\right]ds = \int_0^T \mathbb{E}\left[g(s, W_0^s, Y_0^s)\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right]ds \tag{21}$$

and

$$\lim_{n\to\infty}\int_0^T \mathbb{E}\left[\mathbb{E}[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]\frac{d}{d\rho}g_{(n)}(s, W_0^s, Y_{(n),0}^s)\right]ds = \int_0^T \mathbb{E}\left[\mathbb{E}[g(s, W_0^s, Y_0^s)|Y_0^T]\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right]ds. \tag{22}$$

Step 3.3.1. In this step, we will prove (21). Writing $g_{(n)}(s, W_0^s, Y_{(n),0}^s), g(s, W_0^s, Y_0^s)$ as $g_{(n)}(s), g(s)$ for notational simplicity, we have

$$\int_0^T \mathbb{E}\left[g_{(n)}(s)\frac{d}{d\rho}g_{(n)}(s)\right]ds - \int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\right]ds$$
$$= \int_0^T \mathbb{E}\left[g_{(n)}(s)\frac{d}{d\rho}g_{(n)}(s)\right]ds - \int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g_{(n)}(s)\right]ds + \int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g_{(n)}(s)\right]ds - \int_0^T \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\right]ds$$
$$= \int_0^T \mathbb{E}\left[(g_{(n)}(s) - g(s))\frac{d}{d\rho}g_{(n)}(s)\right]ds - \int_0^T \mathbb{E}\left[g(s)\left(\frac{d}{d\rho}g_{(n)}(s) - \frac{d}{d\rho}g(s)\right)\right]ds.$$

The desired convergences then follow from the fact that as n tends to infinity,

$$\left(\int_0^T \mathbb{E}\left[(g_{(n)}(s) - g(s))\frac{d}{d\rho}g_{(n)}(s)\right]ds\right)^2 \leq \int_0^T \mathbb{E}[(g_{(n)}(s)-g(s))^2]ds\int_0^T \mathbb{E}\left[\left(\frac{d}{d\rho}g_{(n)}(s)\right)^2\right]ds \to 0$$

and

$$\left(\int_0^T \mathbb{E}\left[g(s)\left(\frac{d}{d\rho}g_{(n)}(s) - \frac{d}{d\rho}g(s)\right)\right]ds\right)^2 \leq \int_0^T \mathbb{E}[g^2(s)]ds\int_0^T \mathbb{E}\left[\left(\frac{d}{d\rho}g_{(n)}(s) - \frac{d}{d\rho}g(s)\right)^2\right]ds \to 0.$$

Step 3.3.2. In this step, we will prove (22). To see this, note that

$$\int_0^T \mathbb{E}\left[\mathbb{E}[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]\frac{d}{d\rho}g_{(n)}(s, W_0^s, Y_{(n),0}^s)\right] ds$$
$$= \int_0^T \mathbb{E}\left[\mathbb{E}[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]\mathbb{E}\left[\frac{d}{d\rho}g_{(n)}(s, W_0^s, Y_{(n),0}^s)\middle| Y_{(n),0}^T\right]\right] ds,$$

whose convergence to

$$\int_0^T \mathbb{E}\left[\mathbb{E}[g(s, W_0^s, Y_0^s)|Y_0^T]\mathbb{E}\left[\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\middle| Y_0^T\right]\right] ds$$

can be established using a similar argument as in Step 3.2.

Step 3.4. Note that Steps 3.1, 3.2 and 3.3 collectively yield (19).

Step 4. In this step, we will prove

$$\lim_{n\to\infty} I(W_0^T; Y_{(n),0}^T) = I(W_0^T; Y_0^T). \tag{23}$$

Obviously, it suffices to prove that

$$\lim_{n\to\infty} \int_0^T \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^s)]ds = \int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds, \tag{24}$$

and

$$\lim_{n\to\infty} \int_0^T \mathbb{E}[\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^s]]ds = \int_0^T \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^s]]ds. \tag{25}$$

Note that (24) has been established in Step 3.1, and the proof of (25) can be established using a parallel argument as in Step 3.2.

Step 5. In this step, we will establish the continuity of the following terms with respect to ρ:

$$\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds, \quad \int_0^T \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^T]]ds,$$

and

$$\int_0^T \mathbb{E}\left[g(s, W_0^s, Y_0^s)\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right] ds, \quad \int_0^T \mathbb{E}\left[\mathbb{E}[g(s, W_0^s, Y_0^s)|Y_0^T]\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right] ds.$$

Note that the continuity of $\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds$ immediately follows from the dominated convergence theorem together with Condition (d) and the fact that $g(s, W_0^s, Y_0^s)$ is continuous in ρ. And moreover, a parallel argument can be used to establish the continuity of

$$\int_0^T \mathbb{E}\left[g(s, W_0^s, Y_0^s)\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right] ds.$$

To establish the continuity of $\int_0^T \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^T]]ds$, it suffices to prove that for any sequence $\{\rho_n\}$ convergent to ρ,

$$\lim_{n\to\infty} \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^{(\rho_n),s})|Y_0^{(\rho_n),T}]] = \int_0^T \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^{(\rho),s})|Y_0^{(\rho),T}]]ds,$$

which can be shown in a parallel argument as in Step 3.2, where the following similar convergence is proven:

$$\lim_{n\to\infty}\int_0^T \mathbb{E}[\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]]ds = \int_0^T \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^s)|Y_0^T]]ds.$$

Furthermore, similarly as in Step 3.3.2, the continuity of

$$\int_0^T \mathbb{E}\left[\mathbb{E}[g(s, W_0^s, Y_0^s)|Y_0^T]\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right] ds$$

can be established as well.

Step 6. It then follows from (18) that, for any $\tau > 0$,

$$\begin{aligned}
I(W_0^T; Y_{(n),0}^{(\tau),T}) &= \int_0^\tau \frac{d}{d\rho} I(W_0^T; Y_{(n),0}^{(\rho),T})d\rho \\
&= \rho\int_0^\tau \int_0^T \mathbb{E}[g_{(n)}^2(s, W_0^s, Y_{(n),0}^{(\rho),s})] - \mathbb{E}[\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^{(\rho),s})|Y_{(n),0}^{(\rho),T}]]dsd\rho \\
&+ \rho^2\int_0^\tau \int_0^T \mathbb{E}\left[(g_{(n)}(s, W_0^s, Y_{(n),0}^{(\rho),s}) - \mathbb{E}[g_{(n)}(s, , W_0^s, Y_{(n),0}^{(\rho),s})|Y_{(n),0}^{(\rho),T}])\frac{d}{d\rho}g_{(n)}(s, W_0^s, Y_{(n),0}^{(\rho),s})\right] dsd\rho,
\end{aligned}$$

where we have used the superscripts (ρ) and (τ) to specify the underlying parameters. It then follows from the dominated convergence theorem that

$$\begin{aligned}
I(W_0^T; Y_0^{(\tau),T}) &= \int_0^\tau \rho\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^{(\rho),s})] - \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^{(\rho),s})|Y_0^{(\rho),T}]]dsd\rho \\
+ \int_0^\tau \rho^2\int_0^T \mathbb{E}&\left[(g(s, W_0^s, Y_0^{(\rho),s}) - \mathbb{E}[g(s, , W_0^s, Y_0^{(\rho),s})|Y_{(n),0}^{(\rho),T}])\frac{d}{d\rho}g(s, W_0^s, Y_0^{(\rho),s})\right] dsd\rho. \quad (26)
\end{aligned}$$

Note that Step (5) has established the continuity of the following terms in ρ,

$$\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^{(\rho),s})] - \mathbb{E}[\mathbb{E}^2[g(s, W_0^s, Y_0^{(\rho),s})|Y_0^T]]ds$$

and

$$\int_0^T \mathbb{E}\left[(g(s, W_0^s, Y_0^{(\rho),s}) - \mathbb{E}[g(s, , W_0^s, Y_0^{(\rho),s})|Y_{(n),0}^T])\frac{d}{d\rho}g(s, W_0^s, Y_0^{(\rho),s})\right] ds.$$

So, the desired formula (15) then follows from taking the derivative of (26) with respect to τ, and the proof of the theorem is then complete.

Remark 4.8. Theorem 4.1 is indeed included by Theorem 4.3 as a special case. More precisely, the power constraint (12) trivially implies Conditions (b), (c) and (d). Note that Theorem 4.3 still holds true if Condition (f) is replaced by the following somewhat cumbersome condition: for any n,

$$\frac{d}{d\rho}g_{(n)}(s, W(s), Y_{(n),0}^s) = \left(\frac{d}{d\rho}g(s, W(s), Y_0^s)\right)\mathbf{1}_{\int_0^s g^2(t, W(t), Y_0^t)dt < n}, \quad \text{a.s.},$$

which is also implied by (12). So, Theorem 4.3 recovers Theorem 4.1 with a direct and rigorous proof [1].

Remark 4.9. To show (20), as opposed to our approach in Step 3.2, a possible and seemingly more natural first step is to establish the convergence of $\mathbb{E}^2[g_{(n)}(s, W_0^s, Y_{(n),0}^s)|Y_{(n),0}^T]$ (either in probability or distribution) as n tends to infinity, which, however, has eluded our multiple attempts. Note that for the above-mentioned convergence, the martingale convergence theorem may not be applied, since it is not clear if the σ-algebra generated by $Y_{(n),0}^T$ gets larger at n increases. Similar hurdles were encountered in our attempts to prove (22) and (25), and parallel arguments as in Step 3.2 have to be used instead. Here, we remark that, in general, the problem of establishing the convergence of a sequence of conditional expectations can be rather subtle and challenging; see some positive results in [13] and [6] where some fairly strong assumptions are imposed.

5 Possible Future Directions

The significant impact of the original I-MMSE relation (2) on non-feedback/memoryless channels presages many possible applications of the extended I-MMSE relations (10), (11), (16), (17) to situations where the feedback/memory are present; moreover, we envision that our new approach can provide new perspectives to examine a number of aspects in information theory. In this section, we will discuss some promising future directions one can further pursue based on this work. In a nutshell, the possible further directions can be summarized as follows:

1. further extend the I-MMSE relation to colored Gaussian feedback channels, general feedback channels, and its limiting version in terms of mutual information rate;
2. explore the properties of the extended MMSE;
3. explore the applications of the extended I-MMSE relation to Gaussian feedback channels, multi-user Gaussian channels, Gaussian channels with input/output memory;
4. explore the applications of our new approach to other information-theoretic quantities, higher order derivatives, entropy power inequalities, and so on.

5.1 Further Extensions of the I-MMSE Relation

Colored Gaussian feedback channels. The discrete-time I-MMSE relation (2) carries over verbatim to linear vector Gaussian channels [14], and its extensions to more general settings include derivatives with respect to arbitrary parameterizations [31], higher order derivatives [33], and so on. Extensions of the continuous-time I-MMSE relation (13) have

[1] For sticklers demanding mathematical rigor and perfection: It is known that there are multiple "missing steps" in the proof of Theorem 4.1 in [14]: For instance, the differentiability of $I(X_0^T; Y_0^T)$ with respect to snr does not seem to be trivial and thereby demands careful justifications, which are however absent in [14]; also, from (259) to (270) in the proof of Lemma 5 (a key lemma for the proof of Theorem 4.1), the authors assumed that for a sequence of random variable X_n convergent to 0 almost surely, $\lim_{n\to\infty} \mathbb{E}[X_n] = 0$, which is not true in general.

been studied as well; representative work include fractional Brownian motion noise [9] and an abstract Wiener space [48, 45]. On the other hand, all the above-mentioned extensions have been confined to the scenarios where the feedback are absent.

In view of our results on extensions of the I-MMSE relation, one of the possible future directions is to further extend the I-MMSE relation to colored Gaussian feedback channels in both discrete time and continuous time.

While the proposed direction is well within reach in discrete time, the same problem appears to be far more challenging in continuos time due to the inherent intractability of continuous-time Gaussian processes. A natural goal in this direction is to find the broadest class of continuous-time Gaussian processes for which the extended I-MMSE relation holds. One special class of Gaussian processes that appear to be tractable are those featuring canonical representations [21] (in terms of the standard Brownian motions) without discrete spectrum terms (see (6.8.2) of [22]), and thereby Girsanov's theorem [25], a key technical ingredient used in our proofs of Theorems 4.2 and 4.3, can be carried over to such processes. Since fractional Brownian motions are a special class of separable Gaussian processes, one would arrive at results which include the ones in [9] as special cases.

General feedback channels. The exploration of fundamental relationships between information and estimation measures has not been confined to Gaussian channels only. As a matter of fact, a considerable amount of work, largely inspired by the I-MMSE relation for Gaussian channels, have been devoted to investigating non-Gaussian channels for parallel relations. In this direction, representative work include additive channels [15], arbitrary channels [32], Poisson channels [16, 2, 41], binomial and negative binomial channels [40, 41]. This thread of efforts have culminated in a recent paper [23], where a unified general formula relating information and estimation measures was derived for Lévy channels, which encompass Gaussian channels and a number of other non-Gaussian channels as special cases.

One of the possible directions is to further generalize the result in [23] to Levy channels with feedback/memory, in either discrete or continuous time. Alternatively, one can also consider deriving the extended I-MMSE relation for channel featuring noise with jumps (obviously, noise of this type naturally exists in a variety of real-life situations). For this direction, it might be wiser to first consider additive Levy processes (which are different from Levy channels in [23] in spite of the same name), which have been extensively studied in mathematical theory and practical applications. Note that such extension, if successful, would generalize the one in [10], which only deals with pure jump processes. A key ingredient for success would be an "explicit" Girsanov-type theorem for Leyy processes.

Limiting version. For most non-degenerate channels with feedback/memory, the capacity is computed via maximizing the (directed) mutual information rate, rather than the mutual information. This fact necessitates the consideration of the limiting version of the extended I-MMSE relation in discrete time as n tends to infinity.

There are hurdles for the journey along this direction: First of all, not all input processes will guarantee the limit of the mutual information rate is well-defined. Another issue is the differentiability/smoothness/analyticity of the mutual information rate, which may fail for certain channels [18, 19]. So, it makes senses to focus one's attention on identifying channels with explicit and reasonable assumptions on the input process for the existence of the mutual information rate and its derivative.

Probably a feasible first step is to examine Gaussian channels with Markovian input processes: at least for discrete-time Gaussian channels with ARMA noise, the capacity will be achieved by Markovian input processes [24]. Moreover, for certain Gaussian channels with a finite input alphabet, the analyticity/smoothness/asymptotics of the mutual information rate has been established [19].

5.2 Properties of the Extended MMSE

Properties of the discrete-time MMSE associated with Gaussian non-feedback channels, such as monotonicity, continuity, smoothness, analyticity, concavity and asymptotics, have been extensively studied [17, 46]. These properties have been utilized in a wide range of applications; in particular, the following two properties [17] of the MMSE are of great interest and of direct use in deriving the capacity regions of some multi-user Gaussian channels, such as Gaussian wiretap channels [3] and Gaussian broadcast channels [4]:

- Gaussian inputs are the hardest to estimate, which means that any non-Gaussian input yields strictly smaller MMSE than a Gaussian input of the same variance;
- The single-crossing property, which, roughly speaking, says that a Gaussian MMSE curve (with respect to the snr) only intersects with a non-Gaussian MMSE curve at most once.

Naturally one may consider exploring whether or to what extent these properties hold for the extended MMSE in both discrete and continuous time. It is clear that for the extended MMSE, whether these two properties will hold depends on the adopted encoding schemes, which points out a natural future direction: to explore for what encoding schemes these two properties hold for the extended MMSE. In this direction, one reasonable candidate would be Gaussian channels with linear feedback encoding schemes; see, e.g., [37, 22].

5.3 Applications to Gaussian Feedback Channels

Despite extensive efforts spent on colored Gaussian feedback channels, the capacity of such channels has largely remained unknown, except for some special cases [24]. The extended I-MMSE relations may be helpful to deepen our understanding of colored Gaussian feedback channels: First, notice that an application of the Cauchy-Schwarz inequality yields that the correctional term of an extended MMSE can be upper bounded by the MMSE term, up to a multiplicative constant. Since the MMSE term "corresponds" to Gaussian channels without feedback, it is plausible to at least derive some bound [12] (which may depend on the signal-to-noise ratio) between the ratio of the feedback capacity and non-feedback capacity. Second, written as the sum of an MMSE term and a correctional term, an extended MMSE can be of great help, in both discrete and continuous time, to describe the asymptotical behavior [8] of the feedback capacity for the regime when snr is small or large.

While deriving the capacity of a general colored Gaussian feedback channel seems to be far-fetched, one may consider making use of the extended MMSE relations to derive the feedback capacity for some special colored Gaussian feedback channels. It is well known (see, e.g., Ihara [22]) that for colored Gaussian feedback channels, linear feedback schemes are sufficient to achieve the capacity. This fact can be a major boost of the chance of deriving the exact capacity using the extended I-MMSE: under a linear feedback encoding scheme, the inputs and the outputs are de facto jointly Gaussian, which means both the MMSE and the correctional terms can be explicitly computed. Note that the above-mentioned idea is particularly promising for the case when the Gaussian noise is Markovian, which implies that the correctional term is a scaled version of the MMSE term, and further the desired property that the extended MMSE is maximizeable by a Gaussian distribution.

5.4 Applications to Multi-User Gaussian Channels

Discrete-time. The original I-MMSE relation has been applied to discrete-time multi-user non-feedback Gaussian channels including Gaussian broadcast channels, wiretap channels and interference channels and so on. Naturally, one tempting direction is to explore the possible applications of the extended I-MMSE relation to discrete-time multi-user Gaussian channels when the feedback is present. For this purpose, one of the imminent problems is to identify those multi-user Gaussian channels for which linear feedback coding schemes achieve the capacity regions. Alternatively, one can also look into whether a "multi-user" version of the extended I-MMSE relation exists, which may involve conditional mutual information with multiple message sets. As might be expected, such a multi-user extended I-MMSE relation can provide more insights between the interactions among the users.

Continuous-time. Recently, the infinite bandwidth capacity regions of a continuous-time white Gaussian multiple access channel with/without feedback, a continuous-time white Gaussian interference channel without feedback and a continuous-time white Gaussian broadcast channel without feedback have been derived in [26]. The continuous-time I-MMSE relation has been applied to derive the capacity region of continuous-time white Gaussian broadcast channels. It is very natural to further extend the above-mentioned results and derive the capacity region for more general Gaussian multi-user channels with feedback, such formulas might be of great help for the derivation of the capacity region of continuous-time white Gaussian broadcast channels with feedback, or even more general continuous-time multi-user channels.

5.5 Applications to Gaussian Memory Channels

It is conceivable that the extended I-MMSE relations (11) and (17) may be helpful for us to further understand Gaussian memory channels, which are suitable for modeling some storage systems, such as flash memories [1]. To be more precise, we believe that such extended relations will be helpful in terms of estimating/computing the capacity (region) of (multi-user) Gaussian channels with input/output memory.

5.6 Applications of Our New Approach

Other than the extended I-MMSE relations, one may also consider whether/how the proposed new approach for deriving the extended I-MMSE relation can be applied elsewhere. Below is a list of several scenarios where it can be instrumental.

Other information-theoretic quantities. Other than recovering and extending the original I-MMSE relation, the proposed approach in this paper may be further applied to study other information-theoretic quantities as well, which has been evidenced by the simple and direct proof (see Section 2.2) for the classical de Brunij's identity [39, 5]. It is our opinion that investigations on whether our approach can be applied elsewhere, particularly to the situations where the derivatives of certain information-theoretic quantity are needed, is highly likely to bear fruit. Here, we remark that the derivative of relative entropy has been examined for channels involving mismatched estimation without feedback; see [42, 44].

Higher order derivatives. The second order derivative of the mutual information and entropy power function have also been computed in [17, 33], which, among many other applications, have played a key role in understanding the concavity of the mutual information and deriving entropy power inequalities for Gaussian channels [5, 7, 33, 34]. We expect that such results can be extended to Gaussian feedback channels. Rough computations suggest that the framework of our approach can also be applied to compute higher order derivatives explicitly. Other than understanding concavity, such explicit expressions can also help to characterize the asymptotic behavior of the mutual information and entropy power function associated with Gaussian feedback channels. In this direction, some Talyor-series-expansion-like formulae seem to be within reach, which, undoubtedly, will yield a finer characterization of the behavior of the mutual information and entropy power function of Gaussian feedback channels.

Entropy power inequalities. The ideas and techniques in the proof of the original I-MMSE relation has been used to give new and simpler proofs of a number of entropy power inequalities [43] associated with Gaussian non-feedback channels. It is certainly worthwhile to look into whether these inequalities can be extended to Gaussian feedback channels using our new approach. And, obviously, the same questions can be asked in the continuous-time setting, which, however, appears to be much more challenging.

Acknowledgement. We would like to thank Dongning Guo, Young-Han Kim, Tsachy Weissman for insightful suggestions and comments, and for pointing out relevant references.

Appendices

A Key Lemmas

The following two well-known lemmas are the main tools that will be used to justify the interchanges between a differentiation and an integration in this paper; for their proofs, see [11, Theorem A.5.1, Theorem A.5.2].

Lemma A.1. *Let $f(x,\theta)$ be a continuously differentiable function with respect to θ and X be a random variable. Let $\varepsilon > 0$ and suppose that*
(i) $u(\theta) = \mathbb{E}[f(X,\theta)] < \infty$ for all $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, and
(ii) $v(\theta) = \mathbb{E}[\frac{\partial}{\partial\theta} f(X,\theta)]$ is continuous at $\theta = \theta_0$, and
(iii) $\mathbb{E}\left(\int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon} \left|\frac{\partial}{\partial\theta} f(X,\theta)\right| d\theta\right) < \infty$,
then we have $u'(\theta_0) = v(\theta_0)$, i.e.,

$$\frac{d}{d\theta}\mathbb{E}[f(X,\theta)]\bigg|_{\theta=\theta_0} = \mathbb{E}\left[\frac{\partial}{\partial\theta} f(X,\theta)\right]\bigg|_{\theta=\theta_0}.$$

The following lemma is a direct consequence of the above one.

Lemma A.2. *Let $f(x, \theta)$ be a continuously differentiable function with respect to θ and X be a random variable. Let $\varepsilon > 0$ and suppose that*
(i) $u(\theta) = \mathbb{E}[f(X, \theta)] < \infty$ for $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, and
(ii) $\mathbb{E}\left[\sup_{\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)} \left|\frac{\partial}{\partial \theta} f(X, \theta)\right|\right] < \infty$,
then we have $u'(\theta_0) = v(\theta_0)$, i.e.,

$$\frac{d}{d\theta}\mathbb{E}[f(X, \theta)]\bigg|_{\theta=\theta_0} = \mathbb{E}\left[\frac{\partial}{\partial \theta} f(X, \theta)\right]\bigg|_{\theta=\theta_0}.$$

B Justifications for the interchanges in Section 2.1

1. We first prove that for any $\rho_0 \in \mathbb{R}$,

$$\frac{d}{d\rho}\int_{\mathbb{R}} f_{Y|X}(Y|x) f_X(x) dx\bigg|_{\rho=\rho_0} = \int_{\mathbb{R}} \frac{d}{d\rho} f_{Y|X}(Y|x) f_X(x) dx\bigg|_{\rho=\rho_0},$$

or equivalently, we prove that for any $\rho_0 \in \mathbb{R}$ and for any $x', z' \in \mathbb{R}$,

$$\frac{d}{d\rho}\int_{\mathbb{R}} f_{Y|X}(\rho x' + z'|x) f_X(x) dx\bigg|_{\rho=\rho_0} = \int_{\mathbb{R}} \frac{d}{d\rho} f_{Y|X}(\rho x' + z'|x) f_X(x) dx\bigg|_{\rho=\rho_0}. \quad (27)$$

In what follows, fix $x', z' \in \mathbb{R}$ and $\varepsilon > 0$. Straightforward computations yield that for all $\rho \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)$

$$\int_{\mathbb{R}} f_{Y|X}(\rho x' + z'|x) f_X(x) dx \leq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f_X(x) dx \leq \frac{1}{\sqrt{2\pi}},$$

and moreover,

$$\begin{aligned} \frac{\partial}{\partial \rho} f_{Y|X}(\rho x' + z'|x) &= \frac{\partial}{\partial \rho}\left[e^{-(\rho x' - \rho x + z')^2/2}\right] \\ &= -e^{-(\rho x' - \rho x + z')^2/2}(\rho x' - \rho x + z')(x' - x), \end{aligned}$$

which, together with the assumption that $E[X^2] < \infty$, immediately implies that

$$\int_{\mathbb{R}} \sup_{\rho \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)} \left|\frac{\partial}{\partial \rho} f_{Y|X}(\rho x' + z'|x) f_X(x)\right| dx < \infty.$$

The interchange as in (27) then immediately follows from an invocation of Lemma A.2.

2. We next prove that for any $\rho_0 \in \mathbb{R}$,

$$\frac{d}{d\rho}\mathbb{E}[\log f_Y(Y)]\bigg|_{\rho=\rho_0} = \mathbb{E}\left[\frac{d}{d\rho}\log f_Y(Y)\right]\bigg|_{\rho=\rho_0}. \quad (28)$$

Note that, by the assumption that $E[X^2] < \infty$, we have

$$E[Y^2] = \rho^2 E[X^2] + E[N^2] < \infty,$$

which immediately implies the finiteness of $\mathbb{E}[\log f_Y(Y)]$ for all $\rho \in (\rho - \varepsilon, \rho + \varepsilon)$. As in the proof of Section 2.1, we have

$$\mathbb{E}\left[\frac{d}{d\rho}\log f_Y(Y)\right] = \rho\mathbb{E}[(X - \mathbb{E}[X|Y])^2] = \rho(\mathbb{E}[X^2] - \mathbb{E}[\mathbb{E}^2[X|Y]]),$$

which means to prove the continuity of $\mathbb{E}\left[\frac{d}{d\rho}\log f_Y(Y)\right]$ at $\rho = \rho_0$, it suffices to prove that of $\mathbb{E}[\mathbb{E}^2[X|Y]]$ at $\rho = \rho_0$.

As a matter of fact, we will prove the aforementioned continuity at any ρ. We first show that

$$\mathbb{E}[X|Y] = \frac{1}{f_Y(Y)}\int_{\mathbb{R}} \frac{x}{\sqrt{2\pi}} e^{-(Y-\rho x)^2/2} f_X(x)dx$$

is continuous in ρ. To see this, note that for any ρ, we have

$$\frac{x}{\sqrt{2\pi}} e^{-(Y-\rho x)^2/2} f_X(x) \leq \frac{|x|}{\sqrt{2\pi}} f_X(x),$$

of which the right hand side is integrable. It then follows from the fact that $\frac{x}{\sqrt{2\pi}} e^{-(Y-\rho x)^2/2} f_X(x)$ is continuous at any ρ and the dominated convergence theorem that

$$\int_{\mathbb{R}} \frac{x}{\sqrt{2\pi}} e^{-(Y-\rho x)^2/2} f_X(x)dx$$

is continuous in ρ. A similar argument can be applied to show that $f_Y(Y)$ is also continuous in ρ, which immediately implies the continuity of $\mathbb{E}[X|Y]$ in ρ.

We are now ready to show that $\mathbb{E}[\mathbb{E}^2[X|Y]]$ is continuous in ρ. To see this, note that it follows from $\mathbb{E}[X^2] < \infty$ that $\{\mathbb{E}[X^2|Y], \rho \geq 0\}$ forms a family of uniformly integrable random variables. This, together with the fact that $\mathbb{E}^2[X|Y] \leq \mathbb{E}[X^2|Y]$, implies that $\{\mathbb{E}^2[X|Y], \rho \geq 0\}$ also forms a collection of uniformly integrable random variables. The continuity of $\mathbb{E}[\mathbb{E}^2[X|Y]]$ then follows from that of $\mathbb{E}[X|Y]$ and the uniform integrability of $\{\mathbb{E}^2[X|Y], \rho \geq 0\}$.

Moreover, it can be readily verified that

$$\begin{aligned}
\mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \left|\frac{d}{d\rho}\log f_Y(Y)\right| d\rho\right] &= \mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \left|\int_{\mathbb{R}} (Y-\rho x)(X-x) f_{X|Y}(x|Y)dx\right| d\rho\right] \\
&\leq \mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \int_{\mathbb{R}} |(Y-\rho x)(X-x)|\, f_{X|Y}(x|Y)dxd\rho\right] \\
&\leq \mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \int_{\mathbb{R}} (|YX| + |Yx| + \rho|xX| + \rho|x^2|) f_{X|Y}(x|Y)dxd\rho\right] \\
&= \mathbb{E}[\mathbb{E}[|YX|] + |Y|\mathbb{E}[|X||Y] + \rho|X|\mathbb{E}[|X||Y] + \rho\mathbb{E}[X^2|Y]] \\
&= 2\mathbb{E}[|YX|] + \rho\mathbb{E}[\mathbb{E}^2[|X||Y]] + \rho\mathbb{E}[X^2] \\
&\leq \rho\mathbb{E}[X^2] + \frac{1}{2}\mathbb{E}[X^2] + \frac{1}{2}\mathbb{E}[N^2] + 2\rho\mathbb{E}[X^2],
\end{aligned}$$

which is finite due to the assumption that $\mathbb{E}[X^2] < \infty$ and the fact that $\mathbb{E}[N^2] < \infty$. So, by Lemma A.1, we can switch the integration and differentiation as in (28), and therefore

$$\frac{d}{d\rho} I(X;Y) = -\mathbb{E}\left[\frac{d}{d\rho}\log p_Y(Y)\right] = \mathbb{E}[(X - E[X|Y])^2].$$

C Justifications for the interchanges in Section 2.2

We need to verify that for any $t_0 > 0$,

$$\frac{d}{d\rho}\int_{\mathbb{R}} f_{Y|X}(Y|x)f_X(x)dx\bigg|_{t=t_0} = \int_{\mathbb{R}} \frac{d}{d\rho} f_{Y|X}(Y|x)f_X(x)dx\bigg|_{t=t_0},$$

or equivalently, we prove that for any $t_0 > 0$ and for any $x', z' \in \mathbb{R}$,

$$\frac{d}{dt}\int_{\mathbb{R}} f_{Y|X}(x'+\sqrt{t}z'|x)f_X(x)dx\bigg|_{t=t_0} = \int_{\mathbb{R}} \frac{d}{dt} f_{Y|X}(x'+\sqrt{t}z'|x)f_X(x)dx\bigg|_{t=t_0},$$

which follows from a parallel argument as in the proof of (27). We also need to verify that for any $t_0 > 0$,

$$\frac{d}{d\rho}\mathbb{E}[\log f_Y(Y)]\bigg|_{t=t_0} = \mathbb{E}\left[\frac{d}{d\rho}\log f_Y(Y)\right]\bigg|_{t=t_0},$$

which follows from a parallel argument as in the proof of (28).

D Justifications for the interchanges in the Proof of Theorem 3.1

In this section, we fix $\varepsilon > 0$ and we sometimes write $g_i(W_1^i, Y_1^{i-1})$ as g_i for notational simplicity.

1. We first prove that for any $\rho_0 \in \mathbb{R}_+$, with probability 1,

$$\frac{d}{d\rho}\int_{\mathbb{R}^n} f(Y_1^n|w_1^n)f(w_1^n)dw_1^n\bigg|_{\rho=\rho_0} = \int_{\mathbb{R}^n} \frac{d}{d\rho} f(Y_1^n|w_1^n)f(w_1^n)dw_1^n\bigg|_{\rho=\rho_0}. \tag{29}$$

It follows from straightforward computations that for all $\rho \in (\rho_0 - \varepsilon, \rho_0 + \varepsilon)$

$$\int_{\mathbb{R}} f(Y_1^n|w_1^n)f(w_1^n)dw_1^n \le \frac{1}{(\sqrt{2\pi})^n}\int_{\mathbb{R}} f(w_1^n)dw_1^n \le \frac{1}{(\sqrt{2\pi})^n}.$$

Moreover, we have

$$\frac{d}{d\rho}f(Y_1^n|w_1^n) = \sum_{i=1}^n (Y_i - \rho g_i(w_1^i, Y_1^{i-1}))\left(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1}) + \rho\frac{d}{d\rho}(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1}))\right) f(Y_1^n|w_1^n).$$

It then follows from (8) that, with probability 1,

$$\int_{\mathbb{R}^n} \sup_{\rho\in[\rho_0-\varepsilon,\rho_0+\varepsilon]}\left|\frac{d}{d\rho}f(Y_1^n|w_1^n)f(w_1^n)\right| dw_1^n < \infty.$$

The interchange as in (29) then immediately follows from an invocation of Lemma A.2.

2. We next prove that for any $\rho_0 \in \mathbb{R}$,

$$\frac{d}{d\rho}\mathbb{E}[\log f(Y_1^n)]\bigg|_{\rho=\rho_0} = \mathbb{E}\left[\frac{d}{d\rho}\log f(Y_1^n)\right]\bigg|_{\rho=\rho_0}. \tag{30}$$

Note that, by (8), we have for all $\rho \in [\rho_0 - \varepsilon, \rho_0 + \varepsilon]$ and for all i,

$$E[Y_i^2] = \rho^2 E[\sup_{\rho \in [\rho_0 - \varepsilon, \rho_0 + \varepsilon]} g_i^2(W_1^i, Y_1^{i-1})] + E[Z_i^2] < \infty,$$

which implies that $H(Y_i)$ is upper bounded. On the other hand, it follows from

$$H(Y_i) \geq H(Y_i | Y_1^{i-1}, W_i) = H(Z_i)$$

that $H(Y_i)$ is lower bounded, and so we have obtained the finiteness of $\mathbb{E}[\log f(Y_1^n)]$. As in the proof of Theorem 3.1, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{d}{d\rho}\log f(Y_1^n)\right] &= \rho \sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}[g_i|Y_1^n])^2\right] + \rho^2 \sum_{i=1}^n \mathbb{E}\left[(g_i - \mathbb{E}[g_i|Y_1^n])\frac{d}{d\rho}g_i\right] \\
&= \rho \sum_{i=1}^n (\mathbb{E}[g_i^2] - \mathbb{E}[\mathbb{E}^2[g_i|Y_1^n]]) + \rho^2 \sum_{i=1}^n (\mathbb{E}[g_i \frac{d}{d\rho} g_i] - \mathbb{E}[\mathbb{E}[g_i|Y_1^n]\frac{d}{d\rho}g_i]).
\end{aligned}$$

So, to prove the continuity of $\mathbb{E}\left[\frac{d}{d\rho}\log f(Y_1^n)\right]$ at $\rho = \rho_0$, it suffices to prove that of

$$\mathbb{E}[g_i^2(W_1^i, Y_1^{i-1})],\ \mathbb{E}[\mathbb{E}^2[g_i(W_1^i, Y_1^{i-1})|Y_1^n]],\ \mathbb{E}[g_i(W_1^i, Y_1^{i-1})\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})],\ \mathbb{E}[\mathbb{E}[g_i(W_1^i, Y_1^{i-1})|Y_1^n]\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})]$$

at $\rho = \rho_0$. With Condition (8) and the fact that for all feasible i, $g_i(W_1^i, Y_1^{i-1})$ is continuous in ρ, the continuity of $\mathbb{E}[g_i^2(W_1^i, Y_1^{i-1})]$ immediately follows from the dominated convergence theorem. Similarly, it can be also verified that

$$\mathbb{E}\left[\sup_{\rho \in [\rho_0 - \varepsilon, \rho_0 + \varepsilon]} \left| g_i(W_1^i, Y_1^{i-1})\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})\right|\right] < \infty,$$

which implies the continuity of $\mathbb{E}[g_i(W_1^i, Y_1^{i-1})\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})]$. Moreover, a similar argument as in Section B can be used to establish the continuity of $\mathbb{E}[\mathbb{E}^2[g_i(W_1^i, Y_1^{i-1})|Y_1^n]]$ and $\mathbb{E}[\mathbb{E}[g_i(W_1^i, Y_1^{i-1})|Y_1^n]\frac{d}{d\rho}g_i(W_1^i, Y_1^{i-1})]$ in ρ. We then obtain the continuity of $\mathbb{E}\left[\frac{d}{d\rho}\log f(Y_1^n)\right]$, as desired.

Moreover, we verify that

$$\begin{aligned}
\mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon}\left|\frac{d}{d\rho}\log f(Y_1^n)\right| d\rho\right] &= \mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon}\left|\int_{\mathbb{R}^n}\sum_{i=1}^n (Y_i - \rho g_i(w_1^i, Y_1^{i-1})\Big(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1})\right.\right. \\
&\qquad \left.\left. + \rho\frac{d}{d\rho}(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1}))\Big) f(w_1^n|Y_1^n)dw_1^n\right| d\rho\right] \\
&\leq \mathbb{E}\left[\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon}\int_{\mathbb{R}^n}\sum_{i=1}^n \left|(Y_i - \rho g_i(w_1^i, Y_1^{i-1})\Big(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1})\right.\right. \\
&\qquad \left.\left. + \rho\frac{d}{d\rho}(g_i(W_1^i, Y_1^{i-1}) - g_i(w_1^i, Y_1^{i-1}))\Big)\right| f(w_1^n|Y_1^n)dw_1^n d\rho\right] \\
&< \infty,
\end{aligned}$$

where the finiteness then follows from (8). So, by Lemma A.1, the integration and differentiation in (30) can be interchanged.

E Justifications for the interchanges in the Proof of Theorem 4.2

In this section, let $\varepsilon > 0$ and we sometimes write $g(s, W_0^s, Y_0^s)$ as $g(s)$ for notational simplicity.

1. We first prove that for any $\rho_0 \in \mathbb{R}_+$,

$$\frac{d}{d\rho}\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds\bigg|_{\rho=\rho_0} = 2\int_0^T \mathbb{E}\left[g(s, W_0^s, Y_0^s)\frac{d}{d\rho}g(s, W_0^s, Y_0^s)\right]ds\bigg|_{\rho=\rho_0}.$$

It immediately follows from Condition (d) that for any $\rho \in [\rho_0 - \varepsilon, \rho_0 + \varepsilon]$,

$$\int_0^T \mathbb{E}[g^2(s, W_0^s, Y_0^s)]ds < \infty,$$

and moreover,

$$\int_0^T \mathbb{E}\left[\sup_{\rho\in[\rho_0-\varepsilon,\rho_0+\varepsilon]} 2g(s, W(s), Y_0^s)\frac{d}{d\rho}g(s, W(s), Y_0^s)\right]$$

$$\leq \int_0^T \mathbb{E}\left[\sup_{\rho\in[\rho_0-\varepsilon,\rho_0+\varepsilon]} g^2(s, W(s), Y_0^s)\right]ds + \int_0^T \mathbb{E}\left[\sup_{\rho\in[\rho_0-\varepsilon,\rho_0+\varepsilon]}\left(\frac{d}{d\rho}g(s, W(s), Y_0^s)\right)^2\right]ds < \infty.$$

The desired interchange then immediately follows from Lemma A.2.

2. We next prove that for any $\rho_0 \in \mathbb{R}_+$, we have, with probability 1,

$$\frac{d}{d\rho}\int \frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T|w)\mu_W(dw)\bigg|_{\rho=\rho_0} = \int \frac{d}{d\rho}\frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T|w)\mu_W(dw)\bigg|_{\rho=\rho_0}.$$

First of all, it follows from Theorem 7.1 of [25] that $\mu_Y \sim \mu_B$, and

$$\frac{d\mu_Y}{d\mu_B}(Y_0^T) = \int \frac{d\mu_{Y|W}}{d\mu_B}(Y_0^T|w)\mu_W(dw)$$

is finite almost surely, which can be further written as

$$\frac{d\mu_Y}{d\mu_B}(Y_0^T) = \int \exp\left\{\rho^2\int_0^T \tilde{g}(s)g(s)ds + \rho\int_0^T \tilde{g}(s)dB_s - \frac{\rho^2}{2}\int_0^T \tilde{g}^2(s)ds\right\}\mu_W(dw), \quad (31)$$

where $g(s, w_0^s, Y_0^s)$ is written as $\tilde{g}(s)$ for notational simplicity. Emphasizing the dependence on ρ, we write

$$b(\rho) = \exp\left\{\rho^2\int_0^T \tilde{g}(s)g(s)ds + \rho\int_0^T \tilde{g}(s)dB_s - \frac{\rho^2}{2}\int_0^T \tilde{g}^2(s)ds\right\},$$

and write

$$a(\rho) = \int \frac{d}{d\rho}b(\rho)\mu_W(dw).$$

We now establish the continuity of $a(\rho)$ with respect to ρ. To see this, note that it follows from a routine estimation that

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\left|\frac{a(\rho+\varepsilon)-a(\rho)}{\varepsilon}-\int \frac{d^2}{d\rho^2}b(\rho)\mu_W(dw)\right|^2\right]=0,$$

where we have used the boundedness of $g(s)$. This further implies that for any ρ, we have, with probability 1,

$$a(\rho)-a(0)=\int_0^\rho\int \frac{d^2}{d\gamma^2}b(\gamma)\mu_W(dw)d\gamma.$$

The continuity of $a(\rho)$ then immediately follows (or, more precisely, $a(\rho)$ has a continuous modificaiton). Moreover, it is straightforward to verify that

$$\mathbb{E}\left[\left|\frac{d}{d\rho}b(\rho)\right|\right]<\infty.$$

Finally, with all the technical conditions checked, the desired interchange then immediately follows from Lemma A.1.

3. Finally, we will prove that for any $\rho_0\in\mathbb{R}_+$,

$$\frac{d}{d\rho}\mathbb{E}\left[\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right]\bigg|_{\rho=\rho_0}=\mathbb{E}\left[\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right]\bigg|_{\rho=\rho_0}.$$

First of all, we will show that for all $\rho\in[\rho_0-\varepsilon,\rho_0+\varepsilon]$, $\mathbb{E}\left[\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right]$ is finite. To see this, first note that it follows from Theorem 7.1 of [25] that

$$\frac{d\mu_Y}{d\mu_B}(Y_0^T)=\frac{1}{\mathbb{E}[e^{-\int_0^T XdY+1/2\int_0^T X^2ds}|Y_0^T]}.$$

By Jensen's inequality, we have

$$\mathbb{E}\left[-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds|Y_0^T\right]\le\log\mathbb{E}[e^{-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds}|Y_0^T],$$

and, by the easy fact that $\log x\le x$ for any $x>0$,

$$\log\mathbb{E}[e^{-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds}|Y_0^T]\le\mathbb{E}[e^{-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds}|Y_0^T],$$

The desired finiteness then follows from

$$\left|\log\mathbb{E}[e^{-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds}|Y_0^T]\right|\le\left|\mathbb{E}[-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds|Y_0^T]\right|+\mathbb{E}[e^{-\int_0^T XdY_s+\frac{1}{2}\int_0^T X^2ds}|Y_0^T].$$

Next, as in the proof of Theorem 4.2, we have

$$\mathbb{E}\left[\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right]=\rho\int_0^T\mathbb{E}[\mathbb{E}^2[g(s)|Y_0^T]]ds+\rho^2\int_0^T\mathbb{E}\left[\mathbb{E}[g(s)|Y_0^T]\frac{d}{d\rho}g(s)\right]ds.$$

Note that

$$\int_0^T \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}[\mathbb{E}^2[g(s)|Y_0^T]]ds \leq \int_0^T \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}[\mathbb{E}[g^2(s)|Y_0^T]]ds = \int_0^T \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}[g^2(s)]ds < \infty,$$

and furthermore,

$$\int_0^T \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}\left[\mathbb{E}[g(s)|Y_0^T]\frac{d}{d\rho}g(s)\right] ds \leq \frac{1}{2}\left(\int_0^T \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}\left[\mathbb{E}^2[g(s)|Y_0^T]\right] + \sup_{\rho\in[\rho_0-\varepsilon,\rho+\varepsilon]} \mathbb{E}\left[\left(\frac{d}{d\rho}g(s)\right)^2\right] ds\right) < \infty.$$

It then immediately follows that $\mathbb{E}\left[\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right]$ is continuous with respect to ρ. Moreover, note that

$$\begin{aligned}\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \mathbb{E}\left[\left|\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right|\right] d\rho &= \int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \mathbb{E}\left[\left|\frac{d}{d\rho}\left(\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right)/\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right|\right] d\rho \\ &\leq \int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \mathbb{E}\left[\left|\mathbb{E}\left[\int_0^T g(s)dY(s)\middle| Y_0^T\right]\right| + \rho\left|\mathbb{E}\left[\int_0^T \frac{d}{d\rho}g(s)dY(s)\middle| Y_0^T\right]\right|\right. \\ &\left. + \rho\int_0^T \left|\mathbb{E}[g(s)|Y_0^T]g(s) - \mathbb{E}[g^2(s)|Y_0^T]\right| ds + \rho^2\int_0^T \left|\frac{d}{d\rho}g(s)\mathbb{E}[g(s)|Y_0^T] - \mathbb{E}\left[g(s)\frac{d}{d\rho}g(s)\middle| Y_0^T\right]\right| ds\right] d\rho.\end{aligned}$$

It then follows from Condition (d) that

$$\int_{\rho_0-\varepsilon}^{\rho_0+\varepsilon} \mathbb{E}\left[\left|\frac{d}{d\rho}\log\frac{d\mu_Y}{d\mu_B}(Y_0^T)\right|\right] d\rho < \infty.$$

Finally, with all the technical conditions checked, the desired interchange follows from Lemma A.1.

References

[1] M. Asadi, Xiujie Huang, Aleksandar Kavcic and Narayana (Prasad) Santhanam. Optimal detector for multilevel NAND flash memory channels with intercell interference. *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 5, pp. 825-835, 2014.

[2] R. Atar and T. Weissman. Mutual information, relative entropy, and estimation in the Poisson channel. *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1302-1318, 2012.

[3] R. Bustin, R. Liu, H. V. Poor and S. Shamai. A MMSE Approach to the secrecy capacity of the MIMO Gaussian wiretap channel. *EURASIP Journal on Wireless Communications and Networking*, special issue on wireless physical security, 2009.

[4] R. Bustin and S. Shamai. The I-MMSE approach on the weak Gaussian Z-inteference channel and the type I Gaussian broadcast-Z-inteference channel. *IEEE 26-th Convention of Electrical and Electronic Engineers in Israel*, pp. 000291-000295, Nov., 2010.

[5] M. Costa. A new entropy power inequality. *IEEE Trans. Info. Theory*, vol. 31, no. 11, pp. 751-760, 1985.

[6] I. Crimaldi and L. Pratelli. Convergence results for conditional expectations. *Bernoulli,* vol. 11, no. 4, pp. 737-745, 2005.

[7] A. Dembo. Simple proof on the concavity of the entropy power with respect to added Gaussian noise. *IEEE Trans. Info. Theory*, vol. 35, no. 4, pp. 887-888, 1989.

[8] A. Dembo. On Gaussian feedback capacity. *IEEE Trans. Info. Theory*, vol. 35, no. 5, pp. 1072-1076, 1989.

[9] T. Duncan. Mutual information for stochastic signals and fractional Brownian motion. *IEEE Trans. Info. Theory*, vol. 54, no. 10, pp. 4432-4438, 2008.

[10] T. Duncan. Mutual information for stochastic signals and Levy processes. *IEEE Trans. Info. Theory*, vol. 56, no. 1, pp. 18-24, 2010.

[11] R. Durrett. *Probability: theory and examples*, 4th ed., Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010.

[12] P. Ebert. The capacity of the Gaussian channel with feedback. *Bell Syst. Tech. J.*, vol. 49, pp. 1705-1712, 1970.

[13] E. Goggin. Convergence in distribution of conditional expectations. *The Annals of Probability*, vol. 22, no. 2, pp. 1097-1114, 1994.

[14] D. Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Info. Theory*, vol. 51, no. 4, pp. 1261-1282, 2005.

[15] D. Guo, S. Shamai, and S. Verdu. *Additive non-Gaussian noise channels: mutual information and conditional mean estimation. IEEE ISIT*, pp. 719-723, 2005.

[16] D. Guo, S. Shamai and S. Verdu. Mutual information and conditional mean estimation in Poisson channels. *IEEE Trans. Info. Theory*, vol. 54, no. 5, pp. 1837-1849, 2008.

[17] D. Guo, Y. Wu, S. Shamai and S. Verdu. Estimation in Gaussian Noise: Properties of the Minimum Mean-Square Error. *IEEE Trans. Info. Theory*, vol. 57, no. 4, pp. 2371-2385, 2011.

[18] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5251-5266, 2006.

[19] G. Han and B. Marcus. Analyticity of Entropy Rate of Hidden Markov Chains with Continuous Alphabet. vol. 61, no. 6, pp. 3013 - 3028, 2015.

[20] G. Han and J. Song. Extensions of the I-MMSE relation. *IEEE ISIT*, pp. 2202-2206, 2014.

[21] T. Hida and M. Hitsuda. *Gaussian Processes*, Amer. Math. Soc, Providence, Rhode Island, 1993.

[22] S. Ihara. *Information Theory for Continuous Systems*, World Scientific Publishing Co., Inc., River Edge, NJ, 1993.

[23] J. Jiao, K. Venkat and T. Weissman. Relations between information and estimation in scalar Levy channels. *IEEE ISIT*, pp. 2212-2216, 2014.

[24] Young-Han Kim. Feedback capacity of stationary Gaussian channels. *IEEE Trans. Info. Theory*, vol. 56, no. 1, pp. 57-85, 2010.

[25] R. Liptser and A. Shiryaev. *Statistics of random processes (I): General theory*, 2nd edition, Springer-Verlag, Berlin, 2001.

[26] X. Liu and G. Han. Recent results in continuous-time network information theory. *IEEE ISIT*, pp. 2421-2425, 2014.

[27] A. Lozano, A. Tulino, and S. Verdu. Optimum power allocation for parallel Gaussian channels with arbitrary input distributions. *IEEE Trans. Info. Theory*, vol. 52, no. 7, pp. 3033-3051, 2006.

[28] X. Mao. *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.

[29] X. Mao. *Stochastic Differential Equations and Applications*, Horwood Pub., 2008.

[30] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, 1995.

[31] D. Palomar and S. Verdu. Gradient of mutual information in linear vector Gaussian channels. *IEEE Trans. Info. Theory*, vol. 52, no. 1, pp. 141-154, 2006.

[32] D. Palomar and S. Verdu. Representation of mutual information via input estimates. *IEEE Trans. Info. Theory*, vol. 53, no. 2, pp. 453-470, 2007.

[33] M. Payaro and D. Palomar. Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels. *IEEE Trans. Info. Theory*, vol. 55, no. 8, pp. 3613-3628, 2009.

[34] M. Payaro, M. Gregori and D. Palomar. Yet another entropy power inequality with an application. *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, pp. 1-5, 2011.

[35] M. Peleg, A. Sanderovich and S. Shamai. On extrinsic information of good binary codes operating on Gaussian channels. *Europ. Trans. Telecommun.*, vol. 17, no. 6, 2006.

[36] H. Royden. *Real analysis*, 4th edition, Prentice Hall, Boston, 2010.

[37] J. Schalkwijk and T. Kailath. Coding scheme for additive noise channels with feedback I: No bandwidth constraint. *IEEE Trans. Info. Theory*, vol. 12, no. 2, pp. 172-182, 1966.

[38] S. Shamai. From constrained signaling to network interference via an information-estimation perspective. Shannon Lecture, *IEEE ISIT*, 2011.

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

[40] C. Taborda, D. Guo and F. Pérez-Cruz. Information-estimation relationships over binomial and negative binomial models. *IEEE Trans. Info. Theory*, vol. 60, pp. 2630-2646, 2014.

[41] C. Taborda, F. Pérez-Cruz and D. Guo. New information-estimation results for poisson, binomial and negative binomial models. *IEEE ISIT*, pp. 2207-2211, 2014.

[42] S. Verdu. Mismatched estimation and relative entropy. *IEEE Trans. Info. Theory*, vol. 56, no. 8, pp. 3712-3720, 2010.

[43] S. Verdu and D. Guo. A simple proof of the entropy power inequality. *IEEE Trans. Info. Theory*, vol. 52, no. 5, pp. 2165-2166, 2006.

[44] T. Weissman. The relationship between causal and noncausal mismatched estimation in continuous-time AWGN channels. *IEEE Trans. Info. Theory*, vol. 56, no. 9, pp. 4256-4273, 2010.

[45] E. Mayer-Wolf and M. Zakai. Some relations between mutual information and estimation error in Wiener space. *Ann. Appl. Probab.*, vol. 17, pp. 1102-1116, 2007.

[46] Y. Wu and S. Verdu. Functional properties of minimum mean-square error and mutual information. *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1289-1301, 2012.

[47] Y. Wu, S. Shamai and S. Verdu. Degrees of freedom of the interference Channel: a general formula. *IEEE ISIT*, 2011.

[48] M. Zakai. On mutual information, likelihood ratios, and estimation error for the additive Gaussian channel. *IEEE Trans. Info. Theory*, vol. 51, no. 9, pp. 3017-3024, 2005.