# Functional Properties of MMSE

Yihong Wu
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA
Email: yihongwu@princeton.edu

Sergio Verdú
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA
Email: verdu@princeton.edu

***Abstract*—We show that the minimum mean-square error (MMSE) of estimating the input based on the channel output is a concave functional of the input-output joint distribution, and its various regularity properties are explored. In particular, the MMSE in Gaussian channels is shown to be weakly continuous in the input distribution and Lipschitz continuous with respect to the quadratic Wasserstein distance for peak-limited inputs. Regularity properties of mutual information are also obtained and some connections with rate-distortion theory are also drawn.**

## I. Introduction

Monotonicity, convexity and infinite differentiability of the minimum mean square error (MMSE) in Gaussian noise as a function of SNR have been shown in [1]. In contrast, this paper deals with the functional aspects of MMSE, i.e., as a function of input-output joint distribution, and in particular, as a function of the input distribution when the channel is fixed. We devote special attention to additive Gaussian channels.

The MMSE is a functional of the input-output joint distribution $P_{XY}$, or equivalently of of the pair $(P_X, P_{Y|X})$: Define

$$\begin{aligned} m(P_{XY}) &= m(P_X, P_{Y|X}) && (1)\\ &= \mathsf{mmse}(X|Y) && (2)\\ &= \mathbb{E}[(X - \mathbb{E}[X|Y])^2]. && (3) \end{aligned}$$

These notations will be used interchangeably. When $Y$ is related to $X$ through an additive-noise channel with gain $\sqrt{\mathsf{snr}}$, i.e., $Y = \sqrt{\mathsf{snr}}X + N$ where $N$ is independent of $X$, we denote

$$\begin{aligned} \mathsf{mmse}(X, N, \mathsf{snr}) &= \mathsf{mmse}(X|\sqrt{\mathsf{snr}}X + N), && (4)\\ \mathsf{mmse}(X, \mathsf{snr}) &= \mathsf{mmse}(X, N_{\mathsf{G}}, \mathsf{snr}), && (5) \end{aligned}$$

where $N_{\mathsf{G}}$ is standard Gaussian distributed.

In Section II we study various concavity properties of the MMSE functional defined in (3) – (5). Unlike the mutual information $I(P_X, P_{Y|X})$, which is *concave* in $P_X$, *convex* in $P_{Y|X}$ but neither convex nor concave in $P_{XY}$, the MMSE functional $m(P_{XY})$ is *concave* in the joint distribution $P_{XY}$, hence concave individually in $P_X$ when $P_{Y|X}$ is fixed and in $P_{Y|X}$ when $P_X$ is fixed. However, $m(P_X, P_{Y|X})$ is neither concave nor convex in the pair $(P_X, P_{Y|X})$.

In Section III we discuss the data processing inequality associated with MMSE, which implies $\mathsf{mmse}(X, N, \mathsf{snr})$ is decreasing in $\mathsf{snr}$ for $N$ with a stable distribution, *e.g.*, Gaussian.

In Section IV we present relevant results on the extremization of the MMSE functional with Gaussian inputs and/or Gaussian noise. In terms of MMSE, while the least favorable input for additive Gaussian channels is Gaussian, the worst channel for Gaussian inputs is *not* additive-Gaussian (but the *reverse* channel is). Moreover, it coincides with the optimal forward channel achieving the Gaussian rate-distortion function. Nonetheless, the worst additive-noise channel is still Gaussian.

Various regularity properties of MMSE are explored in Section V.

- We show that $\mathsf{mmse}(X, N, \mathsf{snr})$ is weakly lower semi-continuous (l.s.c.) in $P_X$ but not continuous in general.
- When $N$ has a continuous and bounded density, $P_X \mapsto \mathsf{mmse}(X, N, \mathsf{snr})$ is weakly continuous.
- When $N$ is Gaussian and $X$ is peak-limited, $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ is Lipschitz continuous with respect to the quadratic Wasserstein distance [2].

Via the I-MMSE relationship[1] [3]

$$I(X, \mathsf{snr}) = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma, \qquad (6)$$

where $I(X, \mathsf{snr}) = I(X; \sqrt{\mathsf{snr}}X + N_{\mathsf{G}})$, regularities of MMSE are inherited by the mutual information when the input power is bounded. This enables us to gauge the gap between the Gaussian channel capacity and the mutual information achieved by a given input by computing its Wasserstein distance to the Gaussian distribution.

Due to space limitations, several technical proofs are referred to [4].

## II. Concavity

**Theorem 1.** *$m(P_{XY})$ is a concave functional in $P_{XY}$.*

*Proof:* Fix arbitrary $P_{X_1Y_1}$ and $P_{X_2Y_2}$. Define a random variable $B$ on $\{1, 2\}$ with $\mathbb{P}\{B = 1\} = p$. Then $(X_B, Y_B)$ has joint distribution $\alpha P_{X_1Y_1} + (1 - \alpha)P_{X_2Y_2}$. Therefore

$$\begin{aligned} & m(\alpha P_{X_1Y_1} + (1 - \alpha)P_{X_2Y_2}) \\ &= \mathsf{mmse}(X_B|Y_B) && (7)\\ &\geq \mathsf{mmse}(X_B|Y_B, B) && (8)\\ &= \alpha\ \mathsf{mmse}(X_1|Y_1) + (1 - \alpha)\ \mathsf{mmse}(X_2|Y_2) && (9) \end{aligned}$$

[1] Throughout the paper natural logarithms are adopted and information units are nats.

■

**Corollary 1.** *$m(P_X, P_{Y|X})$ is individually concave in each of its arguments when the other one is fixed.*

**Remark 1.** MMSE is not concave in the pair $(P_X, P_{Y|X})$. We illustrate this point by the following example: for $i = 1, 2$, let $Y_i = X_i + N_i$, where $X_1$ and $N_1$ are independent and equiprobable Bernoulli, $X_2$ and $N_2$ are independent and equiprobable on $\{8, 10\}$ and $\{4, 6\}$ respectively. Let $Y = X + N$, where the distribution of $X$ (resp. $N$) is the equal mixture of those of $X_1$ and $X_2$ (resp. $N_1$ and $N_2$). Then

$$\mathsf{mmse}(X|Y) = \frac{1}{4}\left[\mathsf{mmse}(X_1|Y_1) + \mathsf{mmse}(X_2|Y_2)\right] \tag{10}$$
$$< \frac{1}{2}\left[\mathsf{mmse}(X_1|Y_1) + \mathsf{mmse}(X_2|Y_2)\right], \tag{11}$$

because $\mathsf{mmse}(X_1|Y_1) = \frac{1}{8}$ and $\mathsf{mmse}(X_2|Y_2) = \frac{1}{2}$.

**Remark 2** (Non-strict concavity)**.** In general MMSE is *not* strictly concave. It can be shown that $m(\alpha P_{XY} + (1 - \alpha)Q_{XY}) = \alpha m(P_{XY}) + (1 - \alpha)m(Q_{XY})$ holds for all $0 < \alpha < 1$ if and only if

$$\mathbb{E}_P[X|Y = y] = \mathbb{E}_Q[X|Y = y] \tag{12}$$

holds for $P_Y$-a.e. and $Q_Y$-a.e. $y$. Therefore, the non-strict concavity can be established by constructing two distributions which give rise to the same optimal estimator.

1) $P_{XY} \mapsto m(P_{XY})$ is not strictly concave: consider $Y = X + N$ where $X$ and $N$ are i.i.d. By symmetry, $\mathbb{E}[X|Y = y] = \mathbb{E}[N|Y = y]$. Then since $\mathbb{E}[X|Y = y] + \mathbb{E}[N|Y = y] = y$, the optimal estimator is given by $\mathbb{E}[X|Y = y] = y/2$, regardless of the distribution of $X$.
2) There exists $P_X$ such that the mapping $P_{Y|X} \mapsto m(P_X, P_{Y|X})$ is not strictly concave: consider $X$ and $N$ that are i.i.d. and standard Gaussian. Let $P_{XY}$ be the joint distribution of $(X, \sqrt{\mathsf{snr}}X + N)$ and $Q_{XY}$ be that of $(X, \frac{\mathsf{snr}+1}{\sqrt{\mathsf{snr}}}X)$. Then the optimal estimator of $X$ under $P_{XY}$ and $Q_{XY}$ are both $\hat{X}(y) = \frac{\mathsf{snr}}{\sqrt{\mathsf{snr}+1}}y$.
3) There exists $P_{Y|X}$ such that the mapping $P_X \mapsto m(P_X, P_{Y|X})$ is not strictly concave: consider an additive binary noise channel model $Y = X + 2\pi N$, where $N$ is independent of $X$ and equiprobable Bernoulli. Consider two densities of $X$:

$$f_{X_1}(x) = \varphi(x), \tag{13}$$
$$f_{X_2}(x) = \varphi(x)(1 + \sin x), \tag{14}$$

where $\varphi$ denotes the standard normal density. It can be shown that the optimal estimators for (13) and (14) are the same:

$$\hat{X}(y) = y - \frac{2\pi\varphi(y - 2\pi)}{\varphi(y) + \varphi(y - 2\pi)}, \tag{15}$$

hence the MMSE functional for this channel is the same for any mixture of (13) and (14).

Despite the non-strict concavity for general channels, the MMSE in the special case of additive Gaussian channels is indeed a strictly concave functional of the input distribution, as shown next. The proof exploits the relationship between the optimal estimator in Gaussian channels and the Weierstrass transform [5] of the input distribution.

**Theorem 2.** *For fixed $\mathsf{snr} > 0$, $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ is strictly concave.*

In view of (6) and the continuity of $\mathsf{snr} \mapsto I(X, \mathsf{snr})$, we have the following result:

**Corollary 2.** *For fixed $\mathsf{snr} > 0$, $P_X \mapsto I(X, \mathsf{snr})$ is strictly concave.*

## III. Data Processing Inequality

In [6] it is pointed out that MMSE satisfies a data processing inequality similar to the mutual information. We state it below together with a necessary and sufficient condition for equality to hold.

**Theorem 3.** *If[2] $X - Y - Z$, then*

$$\mathsf{mmse}(X|Y) \le \mathsf{mmse}(X|Z) \tag{16}$$

*with equality if and only if $\mathbb{E}[X|Y] = \mathbb{E}[X|Z]$ a.e.*

The monotonicity of $\mathsf{mmse}(X, \mathsf{snr})$ in $\mathsf{snr}$ is a consequence of Theorem 3, because for any $X$ and $\mathsf{snr}_1 \ge \mathsf{snr}_2 > 0$, $X - \left(X + \frac{1}{\sqrt{\mathsf{snr}_1}}N_{\mathsf{G}}\right) - \left(X + \frac{1}{\sqrt{\mathsf{snr}_2}}N_{\mathsf{G}}\right)$ forms a Markov chain. Using the same reasoning we can conclude that, for any $N$ with a stable law[3], $\mathsf{mmse}(X, N, \mathsf{snr})$ is monotonically decreasing in $\mathsf{snr}$ for all $X$.

It should be noted that unlike the data processing inequality for mutual information, equality in (16) does *not* imply that $Z$ is a sufficient statistic of $Y$ for $X$. Consider the following example of a *multiplicative channel*: Let $U, V, Z$ be independent square integrable random variables with zero mean. Let $Y = ZU$ and $X = YV$. Then $X - Y - Z$ and

$$\mathbb{E}[X|Y] = \mathbb{E}[YV|Y] = Y\mathbb{E}[V|ZU] = Y\mathbb{E}V = 0. \tag{17}$$

Similarly $\mathbb{E}[X|Z] = 0$. By Theorem 3, $\mathsf{mmse}(X|Z) = \mathsf{mmse}(X|Y) = \mathsf{var}X$. However, $X - Z - Y$ does not hold.

## IV. Extremizations

Unlike mutual information, the MMSE functional does not have a saddle-point behavior. Nonetheless, in view of Corollary 1, for fixed input (channel resp.) it is meaningful to investigate the worst channel (input resp.).

**Theorem 4** ([1, Proposition 12])**.** *Let $N \sim \mathcal{N}(0, \sigma_N^2)$ independent of $X$,*

$$\max_{P_X : \mathsf{var}X \le \sigma_X^2} \mathsf{mmse}(X|X + N) = \frac{\sigma_N^2 \sigma_X^2}{\sigma_N^2 + \sigma_X^2}, \tag{18}$$

[2] $X - Y - Z$ means that $X$ and $Z$ are conditionally independent given $Y$, i.e., $(X, Y, Z)$ is a Markov chain.

[3] A distribution is called *stable* if for $X_1, X_2$ independent identically distributed according to $P$, for any constants $a, b$, the random variable $aX_1 + bX_2$ has the same distribution as $cX + d$ for some constants $c$ and $d$ [7, Chapter 1].

*where the maximum is achieved if and only if* $X \sim \mathcal{N}(a, \sigma_X^2)$ *for some* $a \in \mathbb{R}$.

**Theorem 5.**

$$\max_{P_{Y|X}:\mathbb{E}[(Y-X)^2]\le D} \mathsf{mmse}(X|Y) = \min\{\mathsf{var}X, D\}, \tag{19}$$

*holds for any* $X$*, where it is understood that* $\mathsf{var}X = \infty$ *if* $\mathbb{E}\left[X^2\right] = \infty$.

*Proof of Theorem 5:* (Converse)

$$\begin{aligned} \mathsf{mmse}(X|Y) &= \mathsf{mmse}(Y-X|Y) && (20)\\ &\le \min\{\mathsf{var}X, \mathsf{var}(Y-X)\} && (21)\\ &\le \min\{\mathsf{var}X, D\}. && (22) \end{aligned}$$

(Achievability) If $D \ge \mathsf{var}X$, $\mathsf{mmse}(X|Y) = \mathsf{var}X$ is achieved by any $Y$ independent of $X$ with $\mathbb{E}Y^2 = D - \sigma_X^2$.

If $D < \mathsf{var}X$, we choose the channel according as follows: let $Z = \sqrt{\mathsf{snr}}X + N_{\mathsf{G}}$ with $N_{\mathsf{G}}$ independent of $X$ and $\mathsf{snr}$ chosen such that $\mathsf{mmse}(X, \mathsf{snr}) = D$. Such an $\mathsf{snr}$ always exists because $\mathsf{mmse}(X, \mathsf{snr})$ is a decreasing function in $\mathsf{snr}$ which vanishes as $\mathsf{snr} \to \infty$. Moreover it can be shown that $\mathsf{mmse}(X, \mathsf{snr}) \to \mathsf{var}X$ as $\mathsf{snr} \to 0$, even if $\mathsf{var}X = \infty$. Let $Y = \mathbb{E}[X|Z]$. Then $\mathbb{E}[(X-Y)^2] = \mathsf{mmse}(X|Y) = \mathsf{mmse}(X, \mathsf{snr}) = D$. ■

Through the I-MMSE relationship (6), Theorem 4 provides an alternative explanation for optimality of Gaussian inputs in Gaussian channels, because the integrand in (6) is maximized pointwise. Consequently, to achieve a mutual information near the capacity, it is equivalent to find $X$ whose MMSE profile approximates the Gaussian MMSE in the $L_1$ norm. This observation will be utilized in Section VI to study how discrete inputs approach the Gaussian channel capacity.

It is interesting to analyze the worst channel for Gaussian input $X \sim \mathcal{N}(0, \sigma_X^2)$. When $D < \sigma_X^2$, the maximal MMSE is achieved by

$$Y = \left(1 - \frac{D}{\sigma_X^2}\right) X + \sqrt{D - \frac{D^2}{\sigma_X^2}} N_{\mathsf{G}}, \tag{23}$$

which coincides with the minimizer of

$$\begin{aligned} R_X(D) &= \max_{P_{Y|X}:\mathbb{E}(Y-X)^2\le D} I(X;Y) && (24)\\ &= \frac{1}{2}\log^+\left(\frac{\sigma_X^2}{D}\right) && (25) \end{aligned}$$

hence the *reverse channel* $Y \to X$ is a Gaussian channel with noise variance $D$. Nonetheless, the worst additive-noise channel is still Gaussian, i.e.,

$$\max_{P_N:\mathsf{var}N\le D} \mathsf{mmse}(X|X+N) = \frac{\sigma_X^2 D}{\sigma_X^2 + D}, \tag{26}$$

because when $N$ is independent $X$, by (20), the problem reduces to the situation in Theorem 4 and the same expression applies.

## V. Regularity

In general the functional $m(P_{XY})$ is not weakly semi-continuous. To see this, consider $(X_n, Y_n) = (X, X/n)$, which converges in distribution to $(X, Y) = (X, 0)$. Therefore $\mathsf{mmse}(X|Y) = \mathsf{var}X$. However, $\mathsf{mmse}(X_n|Y_n) = 0$ for each $n$. Thus, whenever $\mathsf{var}X > 0$, $m(P_{XY})$ is not upper semi-continuous (u.s.c.):

$$\mathsf{mmse}(X|Y) > \limsup_{n\to\infty} \mathsf{mmse}(X_n|Y_n). \tag{27}$$

On the other hand, consider $Y_n = Y = 0$ and

$$X_n = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{n} \\ n & \text{w.p. } \frac{1}{n} \end{cases} \tag{28}$$

Then $X_n \xrightarrow{\mathrm{D}} X = 0$. Since $\mathsf{mmse}(X|Y) = \mathsf{var}X = 0$ and $\mathsf{mmse}(X_n|Y_n) = \mathsf{var}X_n = n - 1$, it holds that $m(P_{XY})$ is not l.s.c.:

$$\mathsf{mmse}(X|Y) < \liminf_{n\to\infty} \mathsf{mmse}(X_n|Y_n). \tag{29}$$

Nevertheless, under the assumptions of bounded input or additive-noise channel, MMSE is indeed a weakly u.s.c. functional.

**Theorem 6.** *Let* $E \in \mathcal{B}_{\mathbb{R}^2}$ *be such that* $\{x : (x,y) \in E\}$ *is bounded. Denote the collection of all Borel probability measures on* $E$ *by* $\mathcal{M}(E)$*. Then* $P_{XY} \mapsto m(P_{XY})$ *is weakly u.s.c. on* $\mathcal{M}(E)$.

*Proof:* Variational representation proves an effective tool in proving semi-continuity and convexity of information measures (for example relative entropy [8], Fisher information [9], etc). Here we follow the same approach by using the following variational characterization of MMSE:

$$\begin{aligned} & m(P_{XY}) \\ = \;& \inf\left\{\mathbb{E}[(X - f(Y))^2] : f \in \mathcal{B}(\mathbb{R}), \mathbb{E}[f^2(Y)] < \infty\right\} && (30)\\ = \;& \inf\left\{\mathbb{E}[(X - f(Y))^2] : f \in C^b(\mathbb{R})\right\} && (31) \end{aligned}$$

where $\mathcal{B}(\mathbb{R})$ and $C^b(\mathbb{R})$ denote the collection of all real-valued Borel and continuous bounded functions on $\mathbb{R}$ respectively, and (31) is due to the denseness of $C^b$ in $L^2$.

For a fixed estimator $f \in C^b(\mathbb{R})$,

$$\mathbb{E}[(X - f(Y))^2] = \iint (x - f(y))^2 P_{XY}(\mathrm{d}x, \mathrm{d}y) \tag{32}$$

is *weakly continuous* in $P_{XY}$. This is because $(x, y) \mapsto (x - f(y))^2 \in C^b(\mathbb{R}^2)$ since $E$ is bounded in $x$. Therefore by (31), $m(P_{XY})$ is weakly u.s.c. because it is the pointwise infimum of weakly continuous functions. In view of the counterexample in (29), we see that the boundedness assumption on $E$ is not superfluous. ■

**Theorem 7.** *For fixed* $\mathsf{snr} > 0$ *and* $N \in L^2(\Omega)$, $P_X \mapsto \mathsf{mmse}(X, N, \mathsf{snr})$ *is weakly u.s.c. If in addition the density of* $N$ *is continuous and bounded. Then* $P_X \mapsto \mathsf{mmse}(X, N, \mathsf{snr})$ *is weakly continuous.*

*Proof:* Note that

$$\mathsf{snr}\cdot\mathsf{mmse}(X,N,\mathsf{snr}) = \mathsf{mmse}(\sqrt{\mathsf{snr}}X|\sqrt{\mathsf{snr}}X+N) \quad (33)$$
$$= \mathsf{mmse}(N|\sqrt{\mathsf{snr}}X+N) \quad (34)$$
$$= \mathsf{mmse}\left(N,X,\mathsf{snr}^{-1}\right). \quad (35)$$

Therefore it is equivalent to prove the upper semi-continuity of $P_X \mapsto \mathsf{mmse}(N,X,\mathsf{snr}^{-1})$. Without loss of generality we shall assume that $\mathsf{snr}=1$.

Let $P_{X_k}$ be a sequence of distributions converging to $P_{X_0}$ weakly. By the Skorohod's representation [10, Theorem 25.6], there exist random variables with distribution $\{P_{X_k}\}$ and $P_X$ respectively, such that $X_k \xrightarrow{\text{a.s.}} X_0$. Let $N$ be a random variable defined on the same probability space and independent of $X$ and $\{X_k\}$.

Denote $Y_k = X_k + N$ and $g_k(y) = \mathbb{E}[N|Y_k=y]$. By the denseness of $C^b$ in $L^2$, for all $\epsilon>0$, there exists $\hat{g}\in C^b$ such that $\left\|g(Y_0)-\hat{g}(Y_0)\right\|_2 < \epsilon$. Then

$$\limsup_{k\to\infty}\sqrt{\mathsf{mmse}(N,X_k,1)}$$
$$\le \limsup_{k\to\infty}\left\|N-\hat{g}(Y_k)\right\|_2 \quad (36)$$
$$\le \|N-g(Y_0)\|_2 + \left\|g(Y_0)-\hat{g}(Y_0)\right\|_2 \quad (37)$$
$$+ \limsup_{k\to\infty}\left\|\hat{g}(Y_0)-\hat{g}(Y_k)\right\|_2 \quad (38)$$
$$\le \sqrt{\mathsf{mmse}(N,X_0,1)}+\epsilon, \quad (39)$$

where the last inequality follows from the dominated convergence theorem. By the arbitrariness of $\epsilon$, the proof for upper semi-continuity is complete.

The proof of weak continuity when $f_N$ is continuous and bounded is more technical [4]. Here we only give a proof under the additional assumption that $f_N$ decays rapidly enough: $f_N(z) = \mathrm{O}\left(|z|^{-2}\right)$ as $|z|\to\infty$. Denote $V_k = v_k(Y_k)$, where

$$v_k(y) = \mathsf{var}(N|Y_k=y) = \mathbb{E}[N^2|Y_k=y] - (\mathbb{E}[N|Y_k=y])^2, \quad (40)$$

and

$$\mathbb{E}[N|Y_k=y] = \frac{\mathbb{E}\left[(y-X_k)f_N(y-X_k)\right]}{\mathbb{E}\left[f_N(y-X_k)\right]}, \quad (41)$$

$$\mathbb{E}[N^2|Y_k=y] = \frac{\mathbb{E}\left[(y-X_k)^2 f_N(y-X_k)\right]}{\mathbb{E}\left[f_N(y-X_k)\right]}. \quad (42)$$

By assumption, $f_N$, $x\mapsto xf_N(x)$ and $x\mapsto x^2 f_N(x)$ are all continuous and bounded functions. Therefore $\{v_k\}$ is a sequence of nonnegative measurable functions converging pointwise to $v_0$. Also, the density $f_{Y_k}(y) = \mathbb{E}\left[f_N(y-X_k)\right]$ converges pointwise to $f_Y(y) = \mathbb{E}\left[f_N(y-X)\right]$. Applying Fatou's lemma yields the lower semi-continuity. ■

**Remark 3.** Theorem 7 cannot be extended to $\mathsf{snr}=0$, because $\mathsf{mmse}(X,N,0) = \mathsf{var}X$, which is weakly l.s.c. in $P_X$ but not continuous, as the example in (28) illustrates. For $\mathsf{snr}>0$, $P_X \mapsto \mathsf{mmse}(X,N,\mathsf{snr})$ need not be weakly continuous if the sufficient conditions in Theorem 7 are not satisfied. For example, suppose that $X$ and $N$ are both equiprobable Bernoulli. Let $X_k = q_k X$, where $q_k$ is a sequence of irrational numbers converging to 1. Then $X_k \to X$ in distribution, and $\mathsf{mmse}(X_k,N,1)=0$ for all $k$, but $\mathsf{mmse}(X,N,1)=\frac{1}{8}$. This also show that under the condition of Theorem 6, $m(P_{XY})$ need not to be weakly continuous in $P_{XY}$.

**Corollary 3.** *For fixed* $\mathsf{snr}>0$*,* $P_X\mapsto\mathsf{mmse}(X,\mathsf{snr})$ *is weakly continuous.*

In view of the representation of MMSE by the Fisher information of the channel output with additive Gaussian noise [3, (58)]:

$$\mathsf{snr}\cdot\mathsf{mmse}(X,\mathsf{snr}) = 1 - J(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}), \quad (43)$$

Corollary 3 implies the weak continuity of $J(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})$ in $P_X$. While Fisher information is only l.s.c. [9, p. 79], here the continuity is due to convolution with the Gaussian density.

Seeking a finer characterization of the modulus of continuity of $P_X\mapsto\mathsf{mmse}(X,\mathsf{snr})$, we introduce the *quadratic Wasserstein distance* [2, Theorem 6.8].

**Definition 1.** The *quadratic Wasserstein space* on $\mathbb{R}^n$ is defined as the collection of all Borel probability measures with finite second moments, denoted by $\mathcal{P}_2(\mathbb{R}^n)$. The *quadratic Wasserstein distance* is a metric on $\mathcal{P}_2(\mathbb{R}^n)$, defined for $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^n)$ as

$$W_2(\mu,\nu) = \inf\left\{\|X-Y\|_2 : X\sim\mu, Y\sim\nu\right\}, \quad (44)$$

where the infimum is over all joint distributions of $(X,Y)$.

The $W_2$ distance metrizes the topology of weak convergence plus convergence of second-order moments. Because in general convergence in distribution does not yield convergence of moments, this topology is strictly finer than the weak-* topology. Since convergence in $W_2$ implies convergence of variance, in view of Corollary 3, for all $\mathsf{snr}\ge 0$, $P_X\mapsto\mathsf{mmse}(X,\mathsf{snr})$ is continuous on the metric space $(\mathcal{P}_2(\mathbb{R}),W_2)$. Capitalizing on the Lipschitz continuity of the optimal estimator for bounded inputs in Gaussian channel, we obtain the Lipschitz continuity of $P_X\mapsto\mathsf{mmse}(X,\mathsf{snr})$ in this regime:

**Theorem 8.** *For all* $\mathsf{snr}\ge 0$, $\mathsf{mmse}(\cdot,\mathsf{snr})$ *is* $W_2$*-continuous. Moreover, if* $\mathsf{var}X,\mathsf{var}Z\le P$ *and* $\|X\|_\infty,\|Z\|_\infty\le K$*, then*

$$|\mathsf{mmse}(X,\mathsf{snr})-\mathsf{mmse}(Z,\mathsf{snr})|$$
$$\le 2(1+K^2\mathsf{snr})\min\left\{\sqrt{P},\frac{1}{\sqrt{\mathsf{snr}}}\right\}W_2(P_X,P_Z). \quad (45)$$

Corollary 3 guarantees that the MMSE of a random variable can be calculated using the MMSE of its successively finer discretizations, which paves the way for numerical calculating MMSE for singular inputs (*e.g.*, Cantor distribution) in [11]. However, one caveat is that to calculate the value of MMSE within a given accuracy, the quantization level needs to grow as $\mathsf{snr}$ grows (roughly as $\log\mathsf{snr}$ in view of Theorem 8) such that quantization error is much smaller than the noise.

## VI. Applications to Mutual Information

In view of the lower semi-continuity of relative entropy [8], $I(X, \mathsf{snr})$ is weakly l.s.c. in $P_X$ but not continuous in general, as the following example illustrates: Let $P_{X_k} = (1 - k^{-1})\,\mathcal{N}(0,1) + k^{-1}\,\mathcal{N}(0, \exp(k^2))$, which converges weakly to $P_X = \mathcal{N}(0,1)$. By the concavity of $I(\cdot, \mathsf{snr})$, $I(X_k, \mathsf{snr}) \to \infty$ but $I(X, \mathsf{snr}) = \frac{1}{2}\log 2$.

Nevertheless, if the input power is bounded (but not necessarily convergent), mutual information is indeed weakly continuous in the input distribution. Applying Corollary 3 and the dominated convergence theorem to (6), we obtain:

**Theorem 9.** *If* $X_k \xrightarrow{\mathrm{D}} X$ *and* $\sup \mathsf{var} X_k < \infty$*, then* $I(X_k, \mathsf{snr}) \to I(X, \mathsf{snr})$ *for any* $\mathsf{snr} \geq 0$.

By the $W_2$-continuity of MMSE (in Theorem 8), $I(\cdot, \mathsf{snr})$ is also $W_2$-continuous. In fact $W_2$-continuity also holds for $P_X \mapsto I(X; \sqrt{\mathsf{snr}}X + N)$ whenever $N$ has finite *non-Gaussianness*

$$D(N) \triangleq D(P_N \,||\, \mathcal{N}(\mathbb{E}[N], \mathsf{var}N)). \tag{46}$$

This can be seen by writing

$$\begin{aligned} &I(X; \sqrt{\mathsf{snr}}X + N) \\ =&\frac{1}{2}\log\left(1 + \frac{\mathsf{snr}\,\mathsf{var}X}{\mathsf{var}N}\right) - D(\sqrt{\mathsf{snr}}X + N) + D(N), \end{aligned} \tag{47}$$

Since variance converges under $W_2$ convergence, upper semi-continuity follows from the lower semi-continuity of relative entropy.

As a consequence of Theorem 9, we can restrict inputs to a weakly dense subset (*e.g.*, discrete distributions) in the maximization

$$C(\mathsf{snr}) = \max_{\mathbb{E}[X^2] \leq 1} I(X, \mathsf{snr}) = \frac{1}{2}\log(1 + \mathsf{snr}). \tag{48}$$

It is interesting to analyze how the gap between $C(\mathsf{snr})$ and the maximal mutual information achieved by unit-variance inputs taking $m$ values, denoted by $C_m(\mathsf{snr})$, closes as $m$ grows. The $W_2$-Lipschitz continuity of mutual information allows us to obtain an upper bound on the convergence rate. It is known that the optimal $W_2$ distance between a given distribution and discrete distributions taking $m$ values coincides with the square root of the quantization error of the optimal $m$-point quantizer [12], which scales according to $\frac{1}{m}$ [13]. Choosing $X$ to be the output of the optimal quantizer and applying some truncation argument, we conclude that $C(\mathsf{snr}) - C_m(\mathsf{snr}) = O\left(\frac{1}{m}\right)$. In fact, the gap vanishes at least exponentially fast [14].

To conclude this section, we give an example where the non-Gaussianness of a sequence of absolutely continuous distributions does not vanish in the central limit theorem. Consider the following example [15, 17.4]: let $\{Z_k\}$ be a sequence of independent random variables, with

$$\mathbb{P}\{Z_k = 1\} = \mathbb{P}\{Z_k = -1\} = \frac{1}{2}(1 - k^{-2}), \tag{49}$$

$$\mathbb{P}\{Z_k = k\} = \mathbb{P}\{Z_k = -k\} = \frac{1}{2}k^{-2}. \tag{50}$$

Define $X_n = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k$. While $\mathsf{var}\, X_n \to 2$, direct computation of characteristic functions reveals that $X_n \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$. Now let $Y_n = X_n + N_{\mathsf{G}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,2)$. Since $\{\mathsf{var} Y_n\}$ is bounded, by Theorem 9, $I(X_n; X_n + N_{\mathsf{G}}) \to \frac{1}{2}\log 2$. In view of (47), we have $D(Y_n) \to \frac{1}{2}\log\frac{3}{2}$.

## VII. Conclusions

In this paper we explored various concavity and regularity properties of the MMSE functional and its connections to Shannon theory. Through the I-MMSE integral formula (6), new results on mutual information are uncovered. In particular, the Lipschitz continuity of MMSE in Theorem 8 measures how fast a discrete input reaches the Gaussian channel capacity as the the constellation cardinality grows, without evaluating the mutual information numerically.

## References


[1] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," submitted to *IEEE Transactions on Information Theory*, 2010.

[2] C. Villani, *Optimal Transport: Old and New*. Springer Verlag, 2008.

[3] D. Guo, S. Shamai, and S. Verdú, "Mutual Information and Minimum Mean-Square Error in Gaussian Channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261 – 1283, April 2005.

[4] Y. Wu and S. Verdú, "MMSE Dimension," submitted to *IEEE Transactions on Information Theory*, 2009.

[5] Y. A. Brychkov and A. P. Prudnikov, *Integral Transforms of Generalized Functions*. CRC Press, 1989.

[6] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1246–1250, 1998.

[7] J. P. Nolan, *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhäuser, 2010, in progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[8] P. Dupuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley-Interscience, 1997.

[9] P. Huber, *Robust Statistics*. Wiley-Interscience, 1981.

[10] P. Billingsley, *Probability and Measure*, 3rd ed. New York: John Wiley & Sons, 1995.

[11] Y. Wu and S. Verdú, "MMSE Dimension," to appear in *IEEE Int. Symp. Information Theory (ISIT)*, Austin, TX, 2010.

[12] J. A. Cuesta-Albertos, "Shape of a distribution through the $L_2$-Wasserstein distance," in *Proceedings of Distributions with Given Marginals and Statistical Modelling*. Kluwer Academic Publishers, 2002, pp. 51 – 62.

[13] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[14] Y. Wu and S. Verdú, "Capacity of Gaussian channels under finite input alphabet Constraint," Draft, 2010.

[15] J. Stoyanov, *Counterexamples in Probability*, 2nd ed. Wiley, 1997.