

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 the pair $(P_X, P_{Y|X})$: Define

$$m(P_{XY}) = m(P_X, P_{Y|X}) \quad (1)$$

$$= \text{mmse}(X|Y) \quad (2)$$

$$= \mathbb{E}[(X - \mathbb{E}[X|Y])^2]. \quad (3)$$

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

$$\text{mmse}(X, N, \text{snr}) = \text{mmse}(X|\sqrt{\text{snr}}X + N), \quad (4)$$

$$\text{mmse}(X, \text{snr}) = \text{mmse}(X, N_G, \text{snr}), \quad (5)$$

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

Via the I-MMSE relationship¹ [3]

$$I(X, \text{snr}) = \int_0^{\text{snr}} \text{mmse}(X, \gamma) d\gamma, \quad (6)$$

where $I(X, \text{snr}) = I(X; \sqrt{\text{snr}}X + N_G)$, 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

$$m(\alpha P_{X_1Y_1} + (1 - \alpha)P_{X_2Y_2}) = \text{mmse}(X_B|Y_B) \quad (7)$$

$$\geq \text{mmse}(X_B|Y_B, B) \quad (8)$$

$$= \alpha \text{mmse}(X_1|Y_1) + (1 - \alpha) \text{mmse}(X_2|Y_2) \quad (9)$$

¹Throughout 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

$$\text{mmse}(X|Y) = \frac{1}{4} [\text{mmse}(X_1|Y_1) + \text{mmse}(X_2|Y_2)] \quad (10)$$

$$< \frac{1}{2} [\text{mmse}(X_1|Y_1) + \text{mmse}(X_2|Y_2)], \quad (11)$$

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

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

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

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 $(X, \sqrt{\text{snr}}X + N)$ and Q_{XY} be that of $(X, \frac{\text{snr}+1}{\sqrt{\text{snr}}}X)$. Then the optimal estimator of X under P_{XY} and Q_{XY} are both $\hat{X}(y) = \frac{\text{snr}}{\sqrt{\text{snr}+1}}y$.
- 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 :

$$f_{X_1}(x) = \varphi(x), \quad (13)$$

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

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

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

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 $\text{snr} > 0$, $P_X \mapsto \text{mmse}(X, \text{snr})$ is strictly concave.

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

Corollary 2. For fixed $\text{snr} > 0$, $P_X \mapsto I(X, \text{snr})$ 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. If² $X - Y - Z$, then

$$\text{mmse}(X|Y) \leq \text{mmse}(X|Z) \quad (16)$$

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

The monotonicity of $\text{mmse}(X, \text{snr})$ in snr is a consequence of Theorem 3, because for any X and $\text{snr}_1 \geq \text{snr}_2 > 0$, $X - \left(X + \frac{1}{\sqrt{\text{snr}_1}}N_G\right) - \left(X + \frac{1}{\sqrt{\text{snr}_2}}N_G\right)$ forms a Markov chain. Using the same reasoning we can conclude that, for any N with a stable law³, $\text{mmse}(X, N, \text{snr})$ is monotonically decreasing in 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

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

Similarly $\mathbb{E}[X|Z] = 0$. By Theorem 3, $\text{mmse}(X|Z) = \text{mmse}(X|Y) = \text{var}X$. 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 ,

$$\max_{P_X: \text{var}X \leq \sigma_X^2} \text{mmse}(X|X + N) = \frac{\sigma_N^2 \sigma_X^2}{\sigma_N^2 + \sigma_X^2}, \quad (18)$$

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

³A 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.

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

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

Proof of Theorem 5: (Converse)

$$\text{mmse}(X|Y) = \text{mmse}(Y - X|Y) \quad (20)$$

$$\leq \min\{\text{var}X, \text{var}(Y - X)\} \quad (21)$$

$$\leq \min\{\text{var}X, D\}. \quad (22)$$

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

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

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

$$Y = \left(1 - \frac{D}{\sigma_X^2}\right) X + \sqrt{D - \frac{D^2}{\sigma_X^2}} N_G, \quad (23)$$

which coincides with the minimizer of

$$R_X(D) = \max_{P_{Y|X}: \mathbb{E}(Y-X)^2 \leq D} I(X; Y) \quad (24)$$

$$= \frac{1}{2} \log^+ \left(\frac{\sigma_X^2}{D} \right) \quad (25)$$

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

$$\max_{P_N: \text{var}N \leq D} \text{mmse}(X|X + N) = \frac{\sigma_X^2 D}{\sigma_X^2 + D}, \quad (26)$$

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 $\text{mmse}(X|Y) = \text{var}X$. However, $\text{mmse}(X_n|Y_n) = 0$ for each n . Thus, whenever $\text{var}X > 0$, $m(P_{XY})$ is not upper semi-continuous (u.s.c.):

$$\text{mmse}(X|Y) > \limsup_{n \rightarrow \infty} \text{mmse}(X_n|Y_n). \quad (27)$$

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

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

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

$$\text{mmse}(X|Y) < \liminf_{n \rightarrow \infty} \text{mmse}(X_n|Y_n). \quad (29)$$

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:

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

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 $f \in C^b(\mathbb{R})$,

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

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

Proof: Note that

$$\text{snr} \cdot \text{mmse}(X, N, \text{snr}) = \text{mmse}(\sqrt{\text{snr}}X | \sqrt{\text{snr}}X + N) \quad (33)$$

$$= \text{mmse}(N | \sqrt{\text{snr}}X + N) \quad (34)$$

$$= \text{mmse}(N, X, \text{snr}^{-1}). \quad (35)$$

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

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 $\|g(Y_0) - \hat{g}(Y_0)\|_2 < \epsilon$. Then

$$\begin{aligned} & \limsup_{k \rightarrow \infty} \sqrt{\text{mmse}(N, X_k, 1)} \\ & \leq \limsup_{k \rightarrow \infty} \|N - \hat{g}(Y_k)\|_2 \end{aligned} \quad (36)$$

$$\leq \|N - g(Y_0)\|_2 + \|g(Y_0) - \hat{g}(Y_0)\|_2 \quad (37)$$

$$+ \limsup_{k \rightarrow \infty} \|\hat{g}(Y_0) - \hat{g}(Y_k)\|_2 \quad (38)$$

$$\leq \sqrt{\text{mmse}(N, X_0, 1)} + \epsilon, \quad (39)$$

where the last inequality follows from the dominated convergence theorem. By the arbitrariness of ϵ , 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: $f_N(z) = O(|z|^{-2})$ as $|z| \rightarrow \infty$. Denote $V_k = v_k(Y_k)$, where

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

and

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

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

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 $f_{Y_k}(y) = \mathbb{E}[f_N(y - X_k)]$ converges pointwise to $f_Y(y) = \mathbb{E}[f_N(y - X)]$. Applying Fatou's lemma yields the lower semi-continuity. ■

Remark 3. Theorem 7 cannot be extended to $\text{snr} = 0$, because $\text{mmse}(X, N, 0) = \text{var}X$, which is weakly l.s.c. in P_X but not continuous, as the example in (28) illustrates. For $\text{snr} > 0$, $P_X \mapsto \text{mmse}(X, N, \text{snr})$ 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 $X_k \rightarrow X$ in distribution, and $\text{mmse}(X_k, N, 1) = 0$ for all k , but $\text{mmse}(X, N, 1) = \frac{1}{8}$. 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 $\text{snr} > 0$, $P_X \mapsto \text{mmse}(X, \text{snr})$ is weakly continuous.

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

$$\text{snr} \cdot \text{mmse}(X, \text{snr}) = 1 - J(\sqrt{\text{snr}}X + N_G), \quad (43)$$

Corollary 3 implies the weak continuity of $J(\sqrt{\text{snr}}X + N_G)$ 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 $P_X \mapsto \text{mmse}(X, \text{snr})$, 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

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

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 $\text{snr} \geq 0$, $P_X \mapsto \text{mmse}(X, \text{snr})$ 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 $P_X \mapsto \text{mmse}(X, \text{snr})$ in this regime:

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

$$\begin{aligned} & |\text{mmse}(X, \text{snr}) - \text{mmse}(Z, \text{snr})| \\ & \leq 2(1 + K^2 \text{snr}) \min \left\{ \sqrt{P}, \frac{1}{\sqrt{\text{snr}}} \right\} W_2(P_X, P_Z). \end{aligned} \quad (45)$$

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 snr grows (roughly as $\log \text{snr}$ 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], $I(X, \text{snr})$ 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 $I(\cdot, \text{snr})$, $I(X_k, \text{snr}) \rightarrow \infty$ but $I(X, \text{snr}) = \frac{1}{2} \log 2$.

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 $X_k \xrightarrow{D} X$ and $\sup \text{var} X_k < \infty$, then $I(X_k, \text{snr}) \rightarrow I(X, \text{snr})$ for any $\text{snr} \geq 0$.*

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

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

This can be seen by writing

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

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

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

It is interesting to analyze how the gap between $C(\text{snr})$ and the maximal mutual information achieved by unit-variance inputs taking m values, denoted by $C_m(\text{snr})$, 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 $C(\text{snr}) - C_m(\text{snr}) = O\left(\frac{1}{m}\right)$. 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

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

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

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

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. 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.