Functional Properties of Minimum Mean-square Error and Mutual Information

Yihong Wu and Sergio Verdú

***Abstract*—In addition to exploring its various regularity properties, we show that the minimum mean-square error (MMSE) is a concave functional of the input-output joint distribution. In the case of additive Gaussian noise, the MMSE 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. Several applications to information theory and the central limit theorem are discussed.**

***Index Terms*—Bayesian statistics, minimum mean-square error (MMSE), mutual information, Gaussian noise, non-Gaussian noise, central limit theorem.**

I. Introduction

Monotonicity, convexity and infinite differentiability of the minimum mean square error (MMSE) in Gaussian noise as a function of the signal to noise ratio (SNR) have been shown in [2]. In contrast, this paper deals with the functional aspects of MMSE, i.e., as a function of the input-output joint distribution P_{XY}, and in particular, as a function of the input distribution P_X when $P_{Y|X}$ is fixed. We devote special attention to additive Gaussian noise.

The MMSE is a functional of the input-output joint distribution P_{XY} defined on $(\mathbb{R}^2, \mathcal{B})$, or equivalently of the pair $(P_X, P_{Y|X})$: Define

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

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

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

where N_{G} is standard Gaussian distributed. Similarly, we denote the mutual information by

$$I(X, N, \mathsf{snr}) = I(X; \sqrt{\mathsf{snr}}X + N), \tag{6}$$
$$I(X, \mathsf{snr}) = I(X, N_{\mathsf{G}}, \mathsf{snr}), \tag{7}$$

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}, we show that 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. $m(P_X, P_{Y|X})$ is neither concave nor convex in the pair $(P_X, P_{Y|X})$.

Various regularity properties of MMSE are explored in Section III. In particular, 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 has a finite moment of a certain order, $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ is Lipschitz continuous with respect to the Wasserstein distance [3].

For more general distortion functions (not necessarily mean-square), concavity and lower semicontinuity properties of the optimal distortion as a function of the input distribution or channel statistics have recently been studied in [4] in the context of stochastic control.

Via the I-MMSE relationship[1] [5]

$$I(X, \mathsf{snr}) = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma, \tag{8}$$

regularities of MMSE are inherited by the mutual information when the input power is bounded. Several applications of these continuity properties are given in Section IV:

- Using the weak continuity of MMSE, we prove that the differential version of the I-MMSE relationship
$$\frac{\mathrm{d}I(X, \mathsf{snr})}{\mathrm{d}\mathsf{snr}} = \frac{1}{2}\mathsf{mmse}(X, \mathsf{snr}) \tag{9}$$
holds for all $\mathsf{snr} > 0$ as long as the mutual information is finite, thus dropping the finite-variance condition imposed on X in [5, Theorem 1].
- The Lipschitz continuity of MMSE 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.
- We give upper bounds on the convergence rate (in terms of relative entropy) in the central limit theorem for densities obtained as the convolution of Gaussian with another distribution under various moment assumptions.
- We give an example of the central limit theorem in the sense of weak convergence where the non-Gaussianness is finite but does not vanish.

The results of this paper were presented in part at the IEEE International Symposium on Information Theory, Austin, TX, June 13-18, 2010 [1].

Y. Wu and S. Verdú are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: yihongwu@princeton.edu; verdu@princeton.edu).

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

We also give a new bound on the decrease of MMSE due to additional observations in terms of the mutual information.

In Section V we discuss the data processing inequality associated with MMSE, which implies $\mathsf{mmse}(X, N, \mathsf{snr})$ is decreasing in snr for those N with a stable distribution,[2] e.g., Gaussian.

In Section VI we present relevant results on the extremization of the MMSE functional with Gaussian inputs and/or Gaussian noise. While the least favorable input with additive Gaussian noise is Gaussian, the worst random transformation faced by Gaussian inputs is an attenuation followed by additive Gaussian noise, which coincides with the optimal forward random transformation achieving the Gaussian rate-distortion function, i.e. the backward random transformation is an additive Gaussian noise. Nonetheless, the worst additive-noise channel is still Gaussian. We also discuss MMSE-maximizing input distribution under amplitude constraint and its discrete nature.

II. Concavity

Theorem 1. *For any P_{XY}, Q_{XY} and $0 \le \alpha \le 1$,*

$$\begin{aligned} &m(\alpha P_{XY} + (1-\alpha)Q_{XY}) \\ &= \alpha\, m(P_{XY}) + (1-\alpha) m(Q_{XY}) + \alpha(1-\alpha)\times \\ &\int \lambda(\mathrm{d}y) \left(\mathbb{E}_P[X|Y=y] - \mathbb{E}_Q[X|Y=y]\right)^2 \frac{\mathrm{d}P_Y}{\mathrm{d}\lambda}(y) \frac{\mathrm{d}Q_Y}{\mathrm{d}\lambda}(y) \end{aligned} \tag{10}$$

where $\lambda = \alpha P_Y + (1-\alpha)Q_Y$, P_Y and Q_Y denote the marginals of Y under P_{XY} and Q_{XY} respectively. Consequently, $m(P_{XY})$ is a concave functional in P_{XY}.

Proof: Appendix A. ∎

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 on $\{0, 1\}$, X_2 and N_2 are independent and equiprobable on $\{8, 10\}$ and $\{4, 6\}$ respectively. Then $\mathsf{mmse}(X_1|Y_1) = \frac{1}{8}$ and $\mathsf{mmse}(X_2|Y_2) = \frac{1}{2}$. 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

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

$$< \frac{1}{2}\left[\mathsf{mmse}(X_1|Y_1) + \mathsf{mmse}(X_2|Y_2)\right]. \tag{12}$$

Corollary 2 (Non-strict concavity)**.** *In general MMSE is* not *strictly concave.*

Proof: According to (10), it can be shown that

$$m(\alpha P_{XY} + (1-\alpha)Q_{XY}) = \alpha m(P_{XY}) + (1-\alpha) m(Q_{XY}) \tag{13}$$

holds for all $0 < \alpha < 1$ if and only if

$$\mathbb{E}_P[X|Y=y] = \mathbb{E}_Q[X|Y=y] \tag{14}$$

holds for P_Y-a.e. and Q_Y-a.e. y. Therefore, instances of non-strict concavity can be established by constructing pairs of distributions which give rise to the same optimal estimator. Consider the following examples:

1) Let $Y = X + N$ 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. Therefore, in this case, the mapping $P_{XY} \mapsto m(P_{XY})$ is affine between those joint distributions.
2) Let X and N be independent and standard Gaussian. Denote the joint distribution of $(X, \sqrt{\mathsf{snr}}X + N)$ and $(X, \frac{\mathsf{snr}+1}{\sqrt{\mathsf{snr}}}X)$ by P_{XY} and Q_{XY} respectively. Then the optimal estimators of X under P_{XY} and Q_{XY} are both $\hat{X}(y) = \frac{\sqrt{\mathsf{snr}}}{\sqrt{\mathsf{snr}}+1}y$. Therefore, $P_{Y|X} \mapsto m(P_X, P_{Y|X})$ is not strictly concave.
3) Let $Y = X + 2\pi N$, where N is independent of X and equiprobable Bernoulli. Consider two densities of X:

$$f_{X_1}(x) = \varphi(x), \tag{15}$$

$$f_{X_2}(x) = \varphi(x)(1 + \sin x), \tag{16}$$

where $\varphi(x) \triangleq \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ denotes the standard normal density. It can be shown that the optimal estimators for (15) and (16) are the same:

$$\hat{X}(y) = y - \frac{2\pi\varphi(y - 2\pi)}{\varphi(y) + \varphi(y - 2\pi)}, \tag{17}$$

hence the MMSE functional for this channel is the same for any mixture of (15) and (16).

∎

Despite the non-strict concavity in P_X for general $P_{Y|X}$, in the special case of additive Gaussian noise, MMSE is indeed a strictly concave functional of P_X, as shown next. The proof exploits the relationship between the optimal estimator in Gaussian channels and the Weierstrass transform [7] of the input distribution.

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

Proof: Appendix B. ∎

III. Regularity of MMSE

A. Continuity and semi-continuity

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,

$$\mathsf{mmse}(X|Y) > \limsup_{n\to\infty} \mathsf{mmse}(X_n|Y_n) \tag{18}$$

[2] A distribution P is called *stable* if for X_1, X_2 independent identically distributed according to P, for any $a, b \in \mathbb{R}$, the random variable $aX_1 + bX_2$ has the same distribution as $cX + d$ for some $c, d \in \mathbb{R}$ [6, p. 6].

and therefore $m(P_{XY})$ is not lower semi-continuous (l.s.c.) in P_{XY}. On the other hand, consider $Y_n = Y = 0$ and

$$X_n = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{n} \\ n & \text{w.p. } \frac{1}{n} \end{cases} \tag{19}$$

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 u.s.c.:

$$\mathsf{mmse}(X|Y) < \liminf_{n\to\infty} \mathsf{mmse}(X_n|Y_n). \tag{20}$$

Nevertheless, assuming either bounded input or additive noise, MMSE is indeed a weakly u.s.c. functional.

Theorem 3. *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 representations prove effective tools 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[3]:

$$m(P_{XY}) = \inf_{f\in\mathcal{B}(\mathbb{R})} \left\{\mathbb{E}[(X - f(Y))^2] : \mathbb{E}[f^2(Y)] < \infty\right\} \tag{21}$$

$$= \inf_{f\in\mathcal{C}_0(\mathbb{R})} \mathbb{E}[(X - f(Y))^2] \tag{22}$$

where $\mathcal{B}(\mathbb{R})$ and $\mathcal{C}_0(\mathbb{R})$ denote the collection of all real-valued Borel and continuous bounded functions on $\mathbb{R}$ respectively, and (22) is due to the denseness of $\mathcal{C}_0$ in L^2.

For a fixed estimator $f \in \mathcal{C}_0(\mathbb{R})$,

$$\mathbb{E}[(X - f(Y))^2] = \iint (x - f(y))^2 P_{XY}(\mathrm{d}x, \mathrm{d}y) \tag{23}$$

is *weakly continuous* in P_{XY}. This is because $(x, y) \mapsto (x - f(y))^2 \in \mathcal{C}_0(\mathbb{R}^2)$ since E is bounded in x. Therefore by (22), $m(P_{XY})$ is weakly u.s.c. because it is the pointwise infimum of weakly continuous functions. In view of the counterexample in (20), we see that the boundedness assumption on E is not superfluous. ∎

Remark 2. The variational representation of MMSE in (22) provides an alternative proof of its concavity as follows: since for any $f \in \mathcal{B}(\mathbb{R})$, $\mathbb{E}[(X - f(Y))^2]$ is *affine* in P_{XY}, $m(P_{XY})$ is *concave* because it is the pointwise infimum of affine functions. This proof does not rely on the boundedness of E. Hence we could set $E = \mathbb{R}^2$.

Theorem 4. *Let $\mathbb{E}[N^2] < \infty$. Then for any $\mathsf{snr} > 0$, $P_X \mapsto \mathsf{mmse}(X, N, \mathsf{snr})$ is weakly u.s.c. In addition, if N has a continuous and bounded density, then $P_X \mapsto \mathsf{mmse}(X, N, \mathsf{snr})$ is weakly continuous.*

Proof: Appendix C. ∎

[3]The Borel measurability of estimators in (22) is not superfluous. For example [10], it is possible to construct a random variable Y and a non-measurable function $\hat{f}$ such that $X = \hat{f}(Y)$ is a random variable independent of Y. Then $\mathsf{mmse}(X|Y) = \mathsf{var}X$ while $\mathbb{E}[(X - \hat{f}(Y))^2] = 0$.

Remark 3. Theorem 4 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 (19) 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 4 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 3, $m(P_{XY})$ need not be weakly continuous in P_{XY}.

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

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 numerically 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 with snr such that the quantization error is much smaller than the noise.

In view of the representation of the MMSE by the Fisher information of the channel output with additive Gaussian noise [5, (58)] (known as Brown's identity in the statistics literature [12, (1.3.4)]):

$$\mathsf{snr} \cdot \mathsf{mmse}(X, \mathsf{snr}) = 1 - J(\sqrt{\mathsf{snr}}X + N_{\mathsf{G}}), \tag{24}$$

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.

B. Lipschitz continuity

Seeking a finer characterization of the modulus of continuity of $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$, we introduce the *Wasserstein distance* [3].

Definition 1. For $1 \le p \le \infty$, the *Wasserstein space of order p* on $\mathbb{R}$ is defined as the collection of all Borel probability measures with finite p^{th}-order moments, denoted by $\mathcal{P}_p(\mathbb{R})$. The *Wasserstein distance of order p* (W_p distance) is a metric on $\mathcal{P}_p(\mathbb{R})$, defined for $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ as

$$W_p(\mu, \nu) = \inf \left\{ \|X - Y\|_p : X \sim \mu, Y \sim \nu \right\}, \tag{25}$$

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

On the real line the W_p distance coincides with the L_p distance between the quantile functions of two distributions [13], [14]:

$$W_p(P_X, P_Y) = \left\| F_X^{-1} - F_Y^{-1} \right\|_p, \tag{26}$$

where F_X denotes the cumulative distribution function of X. By Hölder's inequality,

$$W_p(P_X, P_Y) \le W_q(P_X, P_Y), \quad p \le q. \tag{27}$$

The W_p distance metrizes the topology of weak convergence plus convergence of p^{th}-order moments. Because in general,

convergence in distribution does not yield convergence of moments, this topology is strictly finer than the weak-* topology. Since W_2 convergence implies convergence of variance, in view of Corollary 3, for all $\mathsf{snr} \geq 0$, $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ is continuous on the metric space $(\mathcal{P}_2(\mathbb{R}), W_2)$. Capitalizing on the smoothness of the optimal estimator in Gaussian channel, the Lipschitz continuity of $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ can be established as follows:

Theorem 5. *For any $\mathsf{snr} \geq 0$ and any $1 \leq p, q \leq \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\begin{aligned} &\frac{|\mathsf{mmse}(Z, \mathsf{snr}) - \mathsf{mmse}(X, \mathsf{snr})|}{\sqrt{\mathsf{var} X} + \sqrt{\mathsf{var} Z}} \\ \leq\ & \left[\sqrt{2}(\|X\|_{4q}^2 + \|Z\|_{4q}^2) + 2\,\mathsf{snr}\,\mathsf{var} X + 1\right] W_{2p}(X, Z). \end{aligned} \tag{28}$$

Consequently, $P_X \mapsto \mathsf{mmse}(X, \mathsf{snr})$ is W_r-Lipschitz continuous on any compact set in $\mathcal{P}_{4r/(r-2)}(\mathbb{R})$ for any $2 \leq r \leq 6$.

Proof: Appendix D. ∎

Finally we present two results which frequently enable us to approximate MMSE of a given input using its truncated version, where the approximation error is *uniform* in the random transformation.

Lemma 1. *For any P_{XYZ},*

$$\left|\sqrt{\mathsf{mmse}(X|Y)} - \sqrt{\mathsf{mmse}(Z|Y)}\right| \leq \|X - Z\|_2. \tag{29}$$

Proof:

$$\begin{aligned} \sqrt{\mathsf{mmse}(X|Y)} &\leq \|X - \mathbb{E}[Z|Y]\|_2 && (30)\\ &\leq \|X - Z\|_2 + \|Z - \mathbb{E}[Z|Y]\|_2 && (31)\\ &= \|X - Z\|_2 + \sqrt{\mathsf{mmse}(Z|Y)}. && (32) \end{aligned}$$

Interchanging the roles of X and Z, (29) follows. ∎

Lemma 2. *Let X be distributed according to P. For $A > 0$, denote by P_A the distribution of X conditioned on the event $\{|X| \leq A\}$, i.e., $P_A(E) = \frac{P(E \cap [-A, A])}{P([-A, A])}$ for any measurable set E. Then For any $P_{Y|X}$,*

$$\begin{aligned} & m(P, P_{Y|X}) - 4\,\mathbb{E}\left[X^2 \mathbf{1}_{\{|X|>A\}}\right] \\ \leq\ & m(P_A, P_{Y|X}) && (33)\\ \leq\ & \frac{m(P, P_{Y|X})}{\mathbb{P}\{|X| \leq A\}} && (34) \end{aligned}$$

Proof of Lemma 2: Let X_A be distributed according to P_A and Y_A be the output of $P_{Y|X}$ when the input is X_A. Then the joint distribution of (X_A, Y_A) is equal to the distribution of (X, Y) conditioned on the event $\{|X| \leq A\}$, i.e.,

$$\mathbb{P}\{X_A \in E, Y_A \in B\} = \frac{\mathbb{P}\{X \in E \cap [-A, A], Y \in B\}}{\mathbb{P}\{X \cap [-A, A]\}}. \tag{35}$$

Since

$$\hat{X}_A(y) \triangleq \mathbb{E}[X_A | Y_A = y] \tag{36}$$

is a suboptimal estimator of X, we have

$$\begin{aligned} & \mathsf{mmse}(X|Y) \\ \leq\ & \mathbb{E}\left[(X - \hat{X}_A(Y))^2\right] && (37)\\ =\ & \mathbb{E}\left[(X - \hat{X}_A(Y))^2 \big| |X| \leq A\right] \mathbb{P}\{|X| \leq A\} \\ & + \mathbb{E}\left[(X - \hat{X}_A(Y))^2 \mathbf{1}_{\{|X|>A\}}\right] && (38)\\ \leq\ & \mathbb{E}\left[(X_A - \hat{X}_A(Y))^2\right] \mathbb{P}\{|X| \leq A\} \\ & + 4\,\mathbb{E}\left[|X|^2 \mathbf{1}_{\{|X|>A\}}\right] && (39)\\ =\ & \mathsf{mmse}(X_A|Y_A)\mathbb{P}\{|X| \leq A\} + 4\,\mathbb{E}\left[|X|^2 \mathbf{1}_{\{|X|>A\}}\right], && (40) \end{aligned}$$

which implies (33). To show (34), note that

$$\begin{aligned} & \mathsf{mmse}(X|Y) \\ =\ & \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right] && (41)\\ \geq\ & \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 || X| \leq A\right] \mathbb{P}\{|X| \leq A\} && (42)\\ \geq\ & \mathsf{mmse}(X_A|Y_A)\mathbb{P}\{|X| \leq A\}. && (43) \end{aligned}$$

∎

IV. Applications to mutual information

A. Finite mutual information and the I-MMSE relationship

Capitalizing on the weak continuity of MMSE proved in Theorem 4, we prove that the I-MMSE relationship holds as long as the mutual information is finite, thus removing the finite-variance condition imposed on the input in [5, Theorem 1].

Theorem 6. *For any random variable X, the following are equivalent:*

a) $I(X, \mathsf{snr}) < \infty$ for any $\mathsf{snr} > 0$;
b) $I(X, \mathsf{snr}) < \infty$ for all $\mathsf{snr} > 0$;
c) $H(\lfloor X \rfloor) < \infty$.

Furthermore, if $I(X, \mathsf{snr}) < \infty$, then (9) holds for all $\mathsf{snr} > 0$.

It should be remarked that $I(X, \mathsf{snr}) < \infty$ is much weaker than $\mathbb{E}[X^2] < \infty$, as the following sufficient condition shows, which also applies to non-Gaussian noise:

Lemma 3. *Let N have a density with $h(N) > -\infty$. Let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be an increasing continuous function that satisfies the following conditions:*

1)

$$\int_{\mathbb{R}_+} \mathrm{e}^{-\psi(x)} \mathrm{d}x < \infty. \tag{44}$$

2) For any $0 \leq \lambda \leq 1$, there exists $a_\lambda, b_\lambda, c_\lambda \geq 0$, such that

$$\psi(\lambda x + (1-\lambda) y) \leq a_\lambda \psi(x) + b_\lambda \psi(y) + c_\lambda, \quad \forall x, y \geq 0. \tag{45}$$

If

$$\mathbb{E}[\psi(|X|)] < \infty \tag{46}$$
$$\mathbb{E}[\psi(|N|)] < \infty \tag{47}$$

Then

1) $I(X, N, \mathsf{snr}) < \infty$ for all $\mathsf{snr} \geq 0$.

2) $\mathsf{snr} \mapsto I(X, N, \mathsf{snr})$ *is continuous on* $\mathbb{R}_+$.

Proof: See Appendix E. ∎

Examples of function ψ that satisfies (45) include:

1) ψ is convex.
2) ψ is increasing and subadditive (e.g., ψ concave with $\psi(0) = 0$).

Consequently, any convex, increasing and non-constant ψ satisfies both (44) and (45). Particularizing to Gaussian noise, choosing ψ to quadratic implies that $\mathsf{var}X < \infty$ is sufficient for $I(X, \mathsf{snr}) < \infty$. In fact, choosing ψ to be nested logarithmic, we obtain a family of sufficient conditions much weaker than square integrability: define

$$l_k(x) \triangleq \underbrace{\log \circ \cdots \circ \log}_{k \text{ times}}(t_k + |x|), \tag{48}$$

where t_k denotes the $(k-1)^{\text{th}}$ tetration (iterated exponential) of e. Then $\mathbb{E}[l_k(X)] < \infty$ for some $k \in \mathbb{N}$ implies $I(X, \mathsf{snr}) < \infty$ for all $\mathsf{snr} > 0$.

Proof of Theorem 6: The equivalence between a) and c) has been shown in [15, Theorem 1]. To prove that a) $\Leftrightarrow$ b), assume that $I(X, \mathsf{snr}') < \infty$ for some $\mathsf{snr}' > 0$. By monotonicity, it is sufficient to consider $\mathsf{snr} > \mathsf{snr}'$. Since $X - (\sqrt{\mathsf{snr}}X + N_{\mathsf{G}}) - (\sqrt{\mathsf{snr}'}X + N_{\mathsf{G}})$ forms a Markov chain, we have

$$\begin{aligned} & I(X, \mathsf{snr}) - I(X, \mathsf{snr}') \\ &= I(X; \sqrt{\mathsf{snr}}X + N_{\mathsf{G}} | \sqrt{\mathsf{snr}'}X + N_{\mathsf{G}}) && (49) \\ &\le \frac{1}{2}\mathbb{E}\left[\log\left(1 + \mathsf{var}(X|\sqrt{\mathsf{snr}'}X + N_{\mathsf{G}})\right)\right] && (50) \\ &\le \frac{1}{2}\log(1 + \mathsf{mmse}(X, \mathsf{snr}')) && (51) \\ &\le \frac{1}{2}\log\left(1 + \frac{1}{\mathsf{snr}'}\right), && (52) \end{aligned}$$

where (50) follows from the concavity of the logarithm. Therefore $I(X, \mathsf{snr}) < \infty$ for all $\mathsf{snr} > 0$.

Next we show that for any $0 < \mathsf{snr}' < \mathsf{snr} < \infty$,

$$I(X, \mathsf{snr}) - I(X, \mathsf{snr}') = \frac{1}{2}\int_{\mathsf{snr}'}^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma. \tag{53}$$

For $m \in \mathbb{N}$, let X_m be a random variable distributed according to the distribution of X conditioned on the event $\{|X| \le m\}$, i.e., $\mathbb{P}\{X_m \in A\} = \frac{\mathbb{P}\{X \in A \cap [-m,m]\}}{\mathbb{P}\{|X| \le m\}}$ for any Borel subset A. This is well-defined because the denominator is positive for sufficiently large m. Since X_m is bounded, we have

$$I(X_m, \mathsf{snr}) - I(X_m, \mathsf{snr}') = \frac{1}{2}\int_{\mathsf{snr}'}^{\mathsf{snr}} \mathsf{mmse}(X_m, \gamma)\mathrm{d}\gamma. \tag{54}$$

Since $X_m \to X$ in distribution, by the weak continuity of MMSE proved in Corollary 3, $\lim_{m\to\infty} \mathsf{mmse}(X_m, \gamma) = \mathsf{mmse}(X, \gamma)$. Since $\mathsf{mmse}(X, \gamma) \le \frac{1}{\gamma}$, which in integrable on the interval $(\mathsf{snr}', \mathsf{snr})$, applying dominated convergence theorem to (54) yields

$$\lim_{m\to\infty} I(X_m, \mathsf{snr}) - I(X_m, \mathsf{snr}') = \frac{1}{2}\int_{\mathsf{snr}'}^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma. \tag{55}$$

To establish (53), it remains to show

$$\lim_{m\to\infty} I(X_m, \mathsf{snr}) = I(X, \mathsf{snr}). \tag{56}$$

By the lower semicontinuity of relative entropy, it is sufficient to show

$$\limsup_{m\to\infty} I(X_m, \mathsf{snr}) \le I(X, \mathsf{snr}). \tag{57}$$

Note that

$$\begin{aligned} I(X, \mathsf{snr}) &= I\left(X, 1_{\{|X|\le m\}}; \sqrt{\mathsf{snr}}X + N_{\mathsf{G}}\right) && (58) \\ &\ge I\left(X; \sqrt{\mathsf{snr}}\,X + N_{\mathsf{G}} \,\middle|\, 1_{\{|X|\le m\}}\right) && (59) \\ &\ge \mathbb{P}(|X| \le m) I\left(X; \sqrt{\mathsf{snr}}\,X + N_{\mathsf{G}} \,\middle|\, |X| \le m\right) && (60) \\ &= \mathbb{P}(|X| \le m) I(X_m, \mathsf{snr}), && (61) \end{aligned}$$

which implies the desired (57).

In view of the continuity of $\mathsf{mmse}(X, \cdot)$ on $(0, \infty)$ established in [2, Proposition 7], differentiating (53) yields (9) for *every* $\mathsf{snr} > 0$. ∎

Remark 4. It can be shown that the integral I-MMSE relationship (8) holds for all $\mathsf{snr} > 0$. First consider the case of $I(X, \mathsf{snr}) < \infty$. Sending $\mathsf{snr}' \to 0$ in (53) and applying the monotone convergence theorem, we have

$$I(X, \mathsf{snr}) - \lim_{\mathsf{snr}'\downarrow 0} I(X, \mathsf{snr}') = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma. \tag{62}$$

Therefore establishing (8) for any X amounts to proving the continuity of $I(X, \cdot)$ at $\mathsf{snr} = 0+$, which has been established in [16] for arbitrary X with finite mutual information.

In case of $I(X, \mathsf{snr}) = \infty$, the integral I-MMSE relationship (8) still holds in the sense that both sides are infinity. To see this, let X_A be distributed according to the conditional distribution P_A defined in Lemma 2. Then

$$\begin{aligned} & \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma \\ &= \frac{1}{2}\liminf_{A\to\infty} \mathbb{P}\{|X| \le A\} \int_0^{\mathsf{snr}} \mathsf{mmse}(X, \gamma)\mathrm{d}\gamma && (63) \\ &\ge \frac{1}{2}\liminf_{A\to\infty} \int_0^{\mathsf{snr}} \mathsf{mmse}(X_A, \gamma)\mathrm{d}\gamma && (64) \\ &\ge \liminf_{A\to\infty} I(X_A, \mathsf{snr}) && (65) \\ &\ge I(X, \mathsf{snr}) = \infty, && (66) \end{aligned}$$

where

- (64): by (34);
- (65): by (8), since X_A is bounded;
- (66): by the lower semicontinuity of $P_X \mapsto I(X, \mathsf{snr})$.

B. Regularity of mutual information

Note that unlike the finite-dimensional setting (e.g. [17]), the concavity of mutual information does not imply continuity. In view of the weak 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

$$P_{X_k} = (1 - k^{-1})\mathcal{N}(0, 1) + k^{-1}\mathcal{N}(0, \exp(\beta k^{\alpha})) \tag{67}$$

with $\alpha \geq 1$ and $\beta > 0$, which converges weakly to $P_X = \mathcal{N}(0,1)$ regardless of the choice of (α, β). Using the concavity of $I(\cdot, \mathsf{snr})$ and the dominated convergence theorem, it can be shown that for any $\mathsf{snr} > 0$,

$$I(X_k, \mathsf{snr}) \to \begin{cases} \frac{1}{2}(\beta + \log(1+\mathsf{snr})) & \alpha = 1 \\ \infty & \alpha > 1 \end{cases} \tag{68}$$

but $I(X, \mathsf{snr}) = \frac{1}{2}\log(1+\mathsf{snr})$.

Nevertheless, mutual information is indeed weakly continuous in the input distribution if the input variance is bounded and the additive noise is Gaussian. Applying Corollary 3 and the dominated convergence theorem to (8), we obtain:

Theorem 7. *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 5), $I(\cdot, \mathsf{snr})$ is also W_2-continuous. Moreover, Lipschitz continuity holds when the input is peak-limited, as the following result shows, which is obtained by integrating both sides of (28) and (8):

Corollary 4. *Under the conditions in Theorem 5, For any $\mathsf{snr} \geq 0$, $1 \leq p, q \leq \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\begin{aligned} &\frac{I(Z, \mathsf{snr}) - I(X, \mathsf{snr})}{\sqrt{\mathsf{var} X} + \sqrt{\mathsf{var} Z}} \\ &\leq \frac{\mathsf{snr}}{2}\left[\sqrt{2}(\|X\|_{4q}^2 + \|Z\|_{4q}^2) + \mathsf{snr}\,\mathsf{var} X + 1\right] W_{2p}(X, Z). \end{aligned} \tag{69}$$

Consequently, $P_X \mapsto I(X, \mathsf{snr})$ is W_r-Lipschitz continuous on any compact set in $\mathcal{P}_{4r/(r-2)}(\mathbb{R})$ for any $2 \leq r \leq 6$.

In fact W_2-continuity also carries over to non-Gaussian noise:

Theorem 8. *Let $W_2(P_{X_k}, P_X) \to 0$, i.e., $X_k \xrightarrow{\mathrm{D}} X$ and second-order moments also converge. Then $I(X_k, N, \mathsf{snr}) \to I(X, N, \mathsf{snr})$ whenever N has finite* non-Gaussianness

$$\mathcal{D}(N) \triangleq D(P_N \,||\, \mathcal{N}(\mathbb{E}[N], \mathsf{var} N)). \tag{70}$$

Proof: Note that

$$\begin{aligned} &I(X, N, \mathsf{snr}) \\ =&\frac{1}{2}\log\left(1 + \frac{\mathsf{snr}\,\mathsf{var} X}{\mathsf{var} N}\right) - \mathcal{D}(\sqrt{\mathsf{snr}} X + N) + \mathcal{D}(N), \end{aligned} \tag{71}$$

Since variance converges under W_2 convergence, the upper semi-continuity $P_X \mapsto I(X, N, \mathsf{snr})$ follows from the lower semi-continuity of relative entropy. As we mentioned before, the lower semicontinuity of $P_X \mapsto I(X, N, \mathsf{snr})$ is inherited by that of relative entropy. ∎

As a consequence of Theorem 7, we can restrict inputs to a weakly dense subset (e.g., discrete distributions) in the maximization

$$C(\mathsf{snr}) = \max_{\mathbb{E}[X^2] \leq 1} I(X, \mathsf{snr}) = \frac{1}{2}\log(1+\mathsf{snr}) \tag{72}$$

with max replaced by sup. 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 [18], which scales according to $\frac{1}{m}$ [19]. Choosing X to be the output of the optimal quantizer and applying a truncation argument, we conclude that $C(\mathsf{snr}) - C_m(\mathsf{snr}) = O\left(\frac{1}{m}\right)$. In fact, the gap vanishes exponentially fast [20].

C. Applications to the central limit theorem

We proceed to prove upper bounds for the convergence rate of non-Gaussianness in central limit theorem with i.i.d. random variables whose distribution is the convolution between a Gaussian distribution and an arbitrary distribution satisfying certain moment assumptions.

Theorem 9. *Let $\{X_i\}$ and $\{W_i\}$ be independent i.i.d. sequences of random variables where $W_i \sim \mathcal{N}(0, \sigma^2)$. Let $Z_i = X_i + W_i$ and*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i. \tag{73}$$

Then

- *If $\mathbb{E}\left[X_1^6\right] < \infty$, then*

$$\mathcal{D}(S_n) = O(n^{-\frac{1}{3}}). \tag{74}$$

- *If the moment generating function of X_1 is finite, then*

$$\mathcal{D}(S_n) = O(n^{-\frac{1}{2}}). \tag{75}$$

Proof: Let $T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$. Then $S_n = T_n + \sigma N_{\mathsf{G}}$ with independent $N_{\mathsf{G}} \sim \mathcal{N}(0,1)$. Let $\bar{X}_1$ denote a normal random variable with the same mean and variance as X_1. Let $1 \leq p, q \leq \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. According to (71), we have

$$\begin{aligned} \mathcal{D}(S_n) &= \mathcal{D}\left(\frac{T_n}{\sigma} + N_{\mathsf{G}}\right) && (76) \\ &= I(\bar{X}_1, \sigma^{-2}) - I(T_n, \sigma^{-2}) && (77) \\ &\leq L_n\, W_{2p}(P_{\bar{X}_1}, P_{T_n}), && (78) \end{aligned}$$

with

$$L_n \triangleq \sqrt{P}\sigma^{-2}\left[\sqrt{2}(\left\|\bar{X}_1\right\|_{4q}^2 + \|T_n\|_{4q}^2) + P\sigma^{-2} + 1\right]. \tag{79}$$

where (78) follows from Corollary 4.

The Wasserstein distance between P_{T_n} and $P_{\bar{X}_1}$ satisfies the following:

- For $r > 2$, if $\mathbb{E}\left[|X_1|^r\right] < \infty$, then [21], [22, (1.5)]

$$W_r(P_{\bar{X}_1}, P_{T_n}) = O(n^{\frac{1}{r} - \frac{1}{2}}). \tag{80}$$

- If the moment generating function of X_1 is finite, then [22, (1.6)] for any $r \geq 1$,

$$W_r(P_{\bar{X}_1}, P_{T_n}) = O(n^{-\frac{1}{2}}). \tag{81}$$

Note that $\|T_n\|_r = \left\|\bar{X}_1\right\|_r + o(1)$ if $\mathbb{E}\left[|X|^r\right] < \infty$. Applying (80) with $r = 6$ and (81) to (78) respectively, we obtain the desired results in (74) and (75). ∎

Remark 5. Non-asymptotic upper bounds on $\mathcal{D}(S_n)$ can be obtained by combining (78) and the non-asymptotic results on the Wasserstein distance in [22, Theorem 2.1 and (1.6)]. Theorem 9 can also be generalized to non-i.i.d. sequences by applying [22, Theorem 4.1].

Remark 6. The asymptotics of $\mathcal{D}(S_n)$ is studied in [23], where it is shown that $\mathcal{D}(S_n) = o(1)$ if and only if $\mathcal{D}(S_{n_0}) < \infty$ for some $n_0 \in \mathbb{N}$. Previous work in [24], [25] showed that if Z_1 has finite Poincaré constant [25, Definition 1.2], the non-Gaussianness $\mathcal{D}(S_n)$ vanishes as the optimal rate $O(n^{-1})$. However, finiteness of the Poincare constant implies Z_1 has moments of all orders [26], while (74) only assumes the existence of the sixth moment. In the special case of compactly-supported X_1, [27, Theorem 1.8] implies that Z_1 has a finite Poincaré constant. Therefore in this case $\mathcal{D}(S_n) = O(n^{-1})$, which is stronger than (75).

After the submission of the present paper, the exact asymptotics of $\mathcal{D}(S_n)$ has been established in [28, Theorem 1.1] using Edgeworth-type expansions: Let S_n be defined in (73) with $\mathbb{E}\left[Z_1^4\right] < \infty$. Then

$$\mathcal{D}(S_n) = \frac{\left(\mathbb{E}\left[Z_1^3\right]\right)^2}{12n} + o\left(\frac{1}{n \log n}\right), \tag{82}$$

which improves Theorem 9 significantly.

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 [29, 17.4]: let $\{Z_k\}$ be a sequence of independent random variables, with

$$\mathbb{P}\{Z_k = 1\} = \mathbb{P}\{Z_k = -1\} = \frac{1}{2}(1 - k^{-2}), \tag{83}$$

$$\mathbb{P}\{Z_k = k\} = \mathbb{P}\{Z_k = -k\} = \frac{1}{2}k^{-2}. \tag{84}$$

Define $S_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n Z_k$. While $\mathsf{var}\, S_n \to 2$, direct computation of characteristic functions reveals that $S_n \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$. Now let $Y_n = S_n + N_{\mathsf{G}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,2)$. Since $\{\mathsf{var} Y_n\}$ is bounded, by Theorem 7, $I(S_n; S_n + N_{\mathsf{G}}) \to \frac{1}{2}\log 2$. In view of (71), we have $\mathcal{D}(Y_n) \to \frac{1}{2}\log\frac{3}{2}$.

D. Bounds on MMSE via mutual information

To conclude this section, we present a result related to Tao's inequality [30], [31], which shows that the contribution of Z in estimating X never exceeds half of the mutual information between X and Z.

Theorem 10. *Let X take values in the unit ball of an Euclidean space. For any $P_{YZ|X}$,*

$$\mathsf{mmse}(X|Y) - \mathsf{mmse}(X|Y,Z) \le \frac{1}{2} I(X; Z|Y). \tag{85}$$

In particular,

$$\mathsf{var}(X) - \mathsf{mmse}(X|Z) \le \frac{1}{2} I(X; Z). \tag{86}$$

Proof: Let $B = \{x : \|x\|_2 \le 1\}$. Then

$$\begin{aligned}
&\mathsf{mmse}(X|Y) - \mathsf{mmse}(X|Y,Z) \\
&= \mathbb{E}\left[\left\|\mathbb{E}\left[X|Y\right] - \mathbb{E}\left[X|Y,Z\right]\right\|_2^2\right] && (87)\\
&= \int P_{YZ}(\mathrm{d}y, \mathrm{d}z) \left\| \int_B x (P_{X|Y}(\mathrm{d}x|y) - P_{X|Y,Z}(\mathrm{d}x|y,z)) \right\|_2^2 && (88)\\
&\le \int P_{YZ}(\mathrm{d}y, \mathrm{d}z) \left(\int_B \|x\|_2 \left| P_{X|Y}(\mathrm{d}x|y) - P_{X|Y,Z}(\mathrm{d}x|y,z) \right| \right)^2 && (89)\\
&\le \int P_{YZ}(\mathrm{d}y, \mathrm{d}z) \left\| P_{X|Y=y} - P_{X|Y=y,Z=z} \right\|_1^2 && (90)\\
&\le 2 \int P_{YZ}(\mathrm{d}y, \mathrm{d}z) D(P_{X|Y=y} \,||\, P_{X|Y=y,Z=z}) && (91)\\
&= 2 I(X; Z|Y). && (92)
\end{aligned}$$

where (90) makes uses of the fact that x belongs to the unit ball, (91) follows from the Kullback-Csiszár-Kemperman-Pinsker inequality (e.g. [32]) and (89) follows from:

$$\left\| \int_B x (\mathrm{d}P - \mathrm{d}Q) \right\|_2 \le \int_B \|x\|_2 \, \mathrm{d}|P - Q|. \tag{93}$$

To verify (93), assume $P \ne Q$. Let μ denote the normalized version of $|P - Q|$, i.e., $\mu = \frac{1}{|P-Q|(B)}|P - Q|$. Denote by $|x|$ the absolutely value of x taken componentwise. Then

$$\begin{aligned}
\left\| \int x (\mathrm{d}P - \mathrm{d}Q) \right\|_2 &\le \left\| \int |x| \mathrm{d}|P - Q| \right\|_2 && (94)\\
&= |P - Q|(B) \left\| \int |x| \mathrm{d}\mu \right\|_2 && (95)\\
&\le |P - Q|(B) \int \| |x| \|_2 \, \mathrm{d}\mu && (96)\\
&= \int \|x\|_2 \, \mathrm{d}|P - Q|, && (97)
\end{aligned}$$

where (96) follows from Jensen's inequality and the convexity of norms. ∎

V. Data processing inequality

In [33] it is pointed out that, like mutual information, MMSE satisfies a data processing inequality. We prove this property along with a necessary and sufficient condition for equality to hold.

Theorem 11. *If*[4] *$X - Y - Z$, then*

$$\mathsf{mmse}(X|Y) \le \mathsf{mmse}(X|Z) \tag{98}$$

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

Proof: By the orthogonality principle and the Markov property,

$$\mathsf{mmse}(X|Z) - \mathsf{mmse}(X|Y) = \mathbb{E}(\mathbb{E}[X|Y] - \mathbb{E}[X|Z])^2. \tag{99}$$

∎

Corollary 5. *For any X, $\mathsf{mmse}(X, \mathsf{snr})$ is decreasing in snr.*

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

Proof: The monotonicity is a consequence of Theorem 11: for any X and $\mathsf{snr}_1 \geq \mathsf{snr}_2 > 0$, $X - \left(X + \frac{1}{\sqrt{\mathsf{snr}_1}} N_{\mathsf{G}}\right) - \left(X + \frac{1}{\sqrt{\mathsf{snr}_2}} N_{\mathsf{G}}\right)$ forms a Markov chain. ∎

Using the same reasoning we can conclude that, for any N with a stable law, $\mathsf{mmse}(X, N, \mathsf{snr})$ is monotonically decreasing in snr for all X.

It should be noted that unlike the data processing inequality for mutual information, equality in (98) 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

$$\mathbb{E}[X|Y] = \mathbb{E}[YV|Y] = Y\,\mathbb{E}[V|ZU] = Y\,\mathbb{E}V = 0. \tag{100}$$

Similarly $\mathbb{E}[X|Z] = 0$. By Theorem 11, $\mathsf{mmse}(X|Z) = \mathsf{mmse}(X|Y) = \mathsf{var}X$. However, $X - Z - Y$ does not hold.

VI. Maximization of MMSE

In this section we consider the maximization of MMSE over a convex set of P_X for fixed $P_{Y|X}$ and vice versa.

A. The worst input distribution

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

$$\max_{P_X : \mathsf{var}X \leq \sigma_X^2} \mathsf{mmse}(X|X+N) = \frac{\sigma_N^2 \sigma_X^2}{\sigma_N^2 + \sigma_X^2}, \tag{101}$$

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

Through the I-MMSE relationship (8), Theorem 12 provides an alternative explanation for optimality of Gaussian inputs in Gaussian channels, because the integrand in (8) is maximized pointwise. Consequently, to approximate the capacity-achieving Gaussian distribution under some given constraints, it is equivalent to find X whose MMSE profile approximates the Gaussian MMSE in the L_1 norm. This observation has been exploited in Section IV to estimate how discrete inputs approach the Gaussian channel capacity.

Theorem 12 states that the least favorable input distribution under the variance constraint is Gaussian. However, under the amplitude constraint, the input distribution P_A^* that achieves

$$\max_{P_X : |X| \leq A} \mathsf{mmse}(X|X+N) \tag{102}$$

is *finitely supported* (see, for example, [34]), a consequence of the analyticity of the Gaussian density. This phenomenon is reminiscent of the fact that the capacity-achieving input distribution for the Gaussian channel under amplitude constraints is also finitely supported [35]. In general, there are no closed-form solutions for P_A^* (see [36] for numerical recipes). However, when $A \leq 1.05$, it has been shown that the worst input distribution is binary [37]:

$$P_A^* = \frac{1}{2}(\delta_A + \delta_{-A}). \tag{103}$$

Capitalizing on the I-MMSE relationship (8), this result has been used to establish that (103) also achieves

$$\max_{P_X : |X| \leq A} I(X; X+N) \tag{104}$$

when $A \leq 1.05$ [38].

The behavior of P_A^* when A is large has been investigated in [39], where it is shown that if X_A^* is distributed according to P_A^*, then $\frac{1}{A} X_A^* \xrightarrow{\mathrm{D}} P^*$ as $A \to \infty$, where the limiting distribution P^* has the following density:

$$p^*(x) = \cos^2 \frac{\pi x}{2} \mathbf{1}_{\{|x| \leq 1\}}. \tag{105}$$

Moreover,

$$\max_{P_X : |X| \leq A} \mathsf{mmse}(X|X+N) = 1 - \frac{\pi^2}{A^2} + o\left(\frac{1}{A^2}\right). \tag{106}$$

An intuitive explanation for (105) and (106) is the following: Let $X_A^* = AZ_A^*$. By (24),

$$\mathsf{mmse}(X_A^*|X_A^* + N) = A^2 \mathsf{mmse}(Z_A^*, A^2) \tag{107}$$
$$= 1 - J(AZ_A^* + N) \tag{108}$$
$$= 1 - A^{-2} J(Z_A^* + A^{-1}N). \tag{109}$$

Suppose that Z_A^* converges to Z^* in distribution and that $J(Z_A^* + A^{-1}N) = J(Z^*) + o(1)$. Then Z^* must minimize the Fisher information among all distributions supported on $[-1, 1]$. This unique minimizer is given by (105) [40], [39]. Similarly, if a variance constraint $\mathsf{var}X \leq \alpha A^2$ is added to (102) in addition to the amplitude constraint, it can be shown that $\frac{1}{A} X_A^*$ converges in distribution to the density supported on $[-1, 1]$ that minimizes the Fisher information with variance not exceeding α, which has been obtained in [41].

Compared to (105), it is interesting to observe that the capacity-achieving input distribution in (104) has a different limiting behavior: Let $\tilde{X}_A^*$ achieves (104). Then $\frac{1}{A} \tilde{X}_A^*$ converges to the uniform distribution on $[-1, 1]$, which maximizes the differential entropy instead of minimizing the Fisher information.

B. The worst random transformation

In this subsection we investigate the variance-constrained additive noise that maximizes the MMSE of the input given its noisy version. It is important to note that we do *not* constrain the noise to be independent of the input. If the criterion is the minimization of mutual information and the additive noise is not allowed to depend on the input, the worst-case noise is known for specific input distributions such as Gaussian and binary [42].

Theorem 13.

$$\max_{P_{Y|X} : \mathbb{E}[(Y-X)^2] \leq D} \mathsf{mmse}(X|Y) = \min\{\sigma_X^2, D\}, \tag{110}$$

holds for any X with finite mean.

Proof of Theorem 13: (Converse)

$$\mathsf{mmse}(X|Y) = \mathsf{mmse}(Y - X|Y) \tag{111}$$
$$\leq \min\{\sigma_X^2, \mathsf{var}(Y - X)\} \tag{112}$$
$$\leq \min\{\sigma_X^2, D\}. \tag{113}$$

(Achievability) If $D \geq \sigma_X^2$, $\mathsf{mmse}(X|Y) = \sigma_X^2$ is achieved by $Y = \mathbb{E}[X]$. If $D < \sigma_X^2$, let $Z = \sqrt{\mathsf{snr}}X + N_{\mathsf{G}}$ with N_{G} independent of X and snr chosen such that $\mathsf{mmse}(X, \mathsf{snr}) = D$. Such an snr always exists because $\mathsf{mmse}(X, \mathsf{snr})$ is a decreasing continuous function [2, Proposition 7] in snr which vanishes as $\mathsf{snr} \to \infty$. Moreover, it can be shown that $\mathsf{mmse}(X, \mathsf{snr}) \to \sigma_X^2$ as $\mathsf{snr} \to 0$, even if $\sigma_X^2 = \infty$. Then $Y = \mathbb{E}[X|Z]$ achieves the upper bound since $\mathbb{E}[(X-Y)^2] = \mathsf{mmse}(X|Y) = \mathsf{mmse}(X, \mathsf{snr}) = D$. ∎

It is interesting to analyze the worst channel in (110) for Gaussian input $X \sim \mathcal{N}(0, \sigma_X^2)$. When $D < \sigma_X^2$, the maximal MMSE is achieved by an attenuator followed by contamination by additive independent Gaussian noise

$$Y = \left(1 - \frac{D}{\sigma_X^2}\right) X + \sqrt{D - \frac{D^2}{\sigma_X^2}} N_{\mathsf{G}}, \tag{114}$$

Interestingly, (114) is the minimizer of

$$R_X(D) = \min_{P_{Y|X}: \mathbb{E}(Y-X)^2 \leq D} I(X;Y) \tag{115}$$

$$= \frac{1}{2} \log^+ \left(\frac{\sigma_X^2}{D}\right), \tag{116}$$

where $\log^+(x) \triangleq \max\{\log x, 0\}$. Hence the *backward* random transformation $P_{X|Y}$ consists of additive Gaussian noise with variance D. Nonetheless, the worst independent additive-noise channel is still Gaussian, i.e.,

$$\max_{P_N: \mathsf{var} N \leq D} \mathsf{mmse}(X|X+N) = \frac{\sigma_X^2 D}{\sigma_X^2 + D}, \tag{117}$$

because when N is independent X, by (111), the problem reduces to the situation in Theorem 12 and the same expression applies.

Appendix A
Proof of Theorem 1

Proof: Let $P_{X_1Y_1}$ and $P_{X_2Y_2}$ be two joint distributions. Let B be a random variable taking values on $\{1, 2\}$ with $\mathbb{P}\{B = 1\} = \alpha_1 = 1 - \alpha_2$. Let (X, Y) be distributed according to $P_{X_iY_i}$ conditioned on $B = i$ for $i = 1$ or 2. Then (X, Y) has joint distribution $P_{XY} = \alpha_1 P_{X_1Y_1} + \alpha_2 P_{X_2Y_2}$. Denote the densities of $P_{X_iY_i}$ and P_{Y_i} with respect to P_{XY} and P_Y by h_i and g_i respectively.

The optimal estimator for X_i based on Y_i is given by

$$\hat{X}_i(y) \triangleq \mathbb{E}[X_i|Y_i = y] = \frac{f_i(y)}{g_i(y)} \tag{118}$$

where

$$f_i(y) \triangleq \int x h_i(x,y) P_{X|Y}(\mathrm{d}x|y). \tag{119}$$

Then the optimal estimator for X based on Y is given by

$$\hat{X}(y) \triangleq \mathbb{E}[X|Y = y] = \int x P_{X|Y}(\mathrm{d}x|y) = \alpha_1 f_1(y) + \alpha_2 f_2(y) \tag{120}$$

since $\alpha_1 h_1 + \alpha_2 h_2 \equiv 1$.

By the orthogonality principle and $\alpha_1 g_1 + \alpha_2 g_2 \equiv 1$,

$$m(\alpha_1 P_{X_1Y_1} + \alpha_2 P_{X_2Y_2}) - \alpha_1 m(P_{X_1Y_1}) - \alpha_2 m(P_{X_2Y_2})$$

$$= \mathsf{mmse}(X|Y) - \mathsf{mmse}(X|Y, B) \tag{121}$$

$$= \mathbb{E}\left[(\mathbb{E}[X|Y,B] - \mathbb{E}[X|Y])^2\right] \tag{122}$$

$$= \alpha_1 \mathbb{E}\left[\left(\frac{f_1}{g_1} - \hat{X}\right)^2 (Y) \middle| B = 1\right] + \alpha_2 \mathbb{E}\left[\left(\frac{f_2}{g_2} - \hat{X}\right)^2 (Y) \middle| B = 2\right] \tag{123}$$

$$= \alpha_1 \alpha_2 \int \frac{(f_1 g_2 - f_2 g_1)^2}{g_1 g_2} P_Y(\mathrm{d}y) \tag{124}$$

$$= \alpha_1 \alpha_2 \mathbb{E}\left[g_1(Y) g_2(Y) (\hat{X}_1(Y) - \hat{X}_2(Y))^2\right]. \tag{125}$$

This proves the desired equality (10). ∎

Appendix B
Proof of Theorem 2

Proof: The concavity follows from Corollary 1. To prove strict concavity, it is sufficient to consider $\mathsf{snr} = 1$. Suppose that for X_i with distribution P_i $(i = 1, 2)$ and $0 < \alpha < 1$, we have

$$\mathsf{mmse}(\alpha P_1 + (1-\alpha) P_2, 1) = \alpha \mathsf{mmse}(P_1, 1) + (1-\alpha) \mathsf{mmse}(P_2, 1). \tag{126}$$

Denote $Y_i = X_i + N_{\mathsf{G}}$ and $\hat{X}_i(y) = \mathbb{E}[X_i|Y_i = y]$ for $i = 1, 2$. Then by (14),

$$\hat{X}_1(y) = \hat{X}_2(y) \tag{127}$$

holds for Lebesgue-a.e. $y \in \mathbb{R}$.

The density of Y_i is given by:

$$f_{Y_i}(y) = \mathbb{E}[\varphi(y - X_i)] = \frac{1}{\sqrt{2\pi}} \mathbb{E}\left[\exp\left(-\frac{1}{2}(y - X_i)^2\right)\right], \tag{128}$$

which is the Weierstrass transform [7] of P_i. Define the likelihood ratio as

$$l_i(y) = \frac{f_{Y_i}(y)}{\varphi(y)} = \mathbb{E}\left[\exp\left(-\frac{X_i^2}{2} + y X_i\right)\right]. \tag{129}$$

The following properties of the log-likelihood ratio are proved in [43] and [44, Property 3] respectively:

$$\frac{\mathrm{d}}{\mathrm{d}y} \log l_i(y) = \hat{X}_i(y), \tag{130}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2} \log l_i(y) = \mathsf{var}(X_i|Y_i = y). \tag{131}$$

The solution of (130) is

$$l_i(y) = l_i(0) \exp\left(\int_0^y \hat{X}_i(t) \mathrm{d}t\right), \tag{132}$$

hence in view of (127), we have for all $y \in \mathbb{R}$,

$$\frac{l_1(y)}{l_1(0)} = \frac{l_2(y)}{l_2(0)}. \tag{133}$$

Next we show that $P_1 = P_2$. For $i = 1, 2$, define a probability measure Q_i according to

$$\mathbb{E}_{Q_i}[f(X)] = \frac{\mathbb{E}[f(X_i) \exp(-X_i^2/2)]}{\mathbb{E}[\exp(-X_i^2/2)]} \tag{134}$$

for each positive Borel function f. From (129), we observe that $\frac{l_i(y)}{l_i(0)}$ is the Laplace transform of Q_i. By (133), we conclude that $Q_1 = Q_2$. Then

$$\begin{aligned}\mathbb{E}[f(X_1)] &= l_1(0)\mathbb{E}_{Q_1}[f(X)\exp(-X^2/2)] &(135)\\ &= l_2(0)\mathbb{E}_{Q_2}[f(X)\exp(-X^2/2)] &(136)\\ &= \mathbb{E}[f(X_2)], &(137)\end{aligned}$$

where (136) follows from

$$l_1(0) = \frac{1}{\mathbb{E}_{Q_1}[\exp(X^2/2)]} = \frac{1}{\mathbb{E}_{Q_2}[\exp(X^2/2)]} = l_2(0). \tag{138}$$

By (137) and the arbitrariness of f, we conclude that $P_1 = P_2$. In view of (8), the strict concavity of $P_X \mapsto I(X, \mathsf{snr})$ follows. ∎

Appendix C
Proof of Theorem 4

Proof of Theorem 4: Note that

$$\begin{aligned}\mathsf{snr}\cdot\mathsf{mmse}(X,N,\mathsf{snr}) &= \mathsf{mmse}(\sqrt{\mathsf{snr}}X|\sqrt{\mathsf{snr}}X+N) &(139)\\ &= \mathsf{mmse}(N|\sqrt{\mathsf{snr}}X+N) &(140)\\ &= \mathsf{mmse}\left(N,X,\mathsf{snr}^{-1}\right). &(141)\end{aligned}$$

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 [45, Theorem 25.6], there exist a sequence of random variables $\{X_k\}_{k\geq 0}$ with distributions $\{P_{X_k}\}$ 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 $\{X_k\}$.

Denote $Y_k = X_k + N$. Let $g_k(y) = \mathbb{E}[N|Y_k = y]$. By the denseness of $\mathcal{C}_0$ in L^2, for all $\epsilon > 0$, there exists $\hat{g} \in \mathcal{C}_0$ such that $\left\|g_0(Y_0) - \hat{g}(Y_0)\right\|_2 < \epsilon$. Then

$$\begin{aligned}&\limsup_{k\to\infty}\sqrt{\mathsf{mmse}(N,X_k,1)}\\ &\leq \limsup_{k\to\infty}\left\|N-\hat{g}(Y_k)\right\|_2 &(142)\\ &\leq \|N-g(Y_0)\|_2 + \left\|g(Y_0)-\hat{g}(Y_0)\right\|_2\\ &\quad + \limsup_{k\to\infty}\left\|\hat{g}(Y_0)-\hat{g}(Y_k)\right\|_2 &(143)\\ &\leq \sqrt{\mathsf{mmse}(N,X_0,1)}+\epsilon, &(144)\end{aligned}$$

where (142) is due to the suboptimality of $\hat{g}$ and (144) follows from the dominated convergence theorem. By the arbitrariness of ϵ, the proof for upper semi-continuity is complete, and it remains to show

$$\liminf_{k\to\infty}\mathsf{mmse}(N,X_k,1) \geq \mathsf{mmse}(N,X_0,1) \tag{145}$$

under the assumption that N has a continuous and bounded density f_N. For every positive integer m define the following continuous and compactly supported function

$$h_m(x) = \begin{cases} x & |x|\leq m\\ m(1+m-|x|) & m<|x|\leq m+1\\ 0 & |x|>m+1\end{cases}. \tag{146}$$

Denote the density of Y_k by $p_k(y) \triangleq \mathbb{E}\left[f_N(y-X_k)\right]$. For fixed $m > 0$, define

$$\begin{aligned}v_k(y) &\triangleq \mathsf{var}(h_m(N)|Y_k=y) &(147)\\ &= \mathbb{E}[h_m^2(N)|Y_k=y] - (\mathbb{E}[h_m(N)|Y_k=y])^2, &(148)\end{aligned}$$

where

$$\mathbb{E}[h_m(N)|Y_k=y] = \frac{\mathbb{E}\left[h_m(y-X_k)f_N(y-X_k)\right]}{\mathbb{E}\left[f_N(y-X_k)\right]}, \tag{149}$$

$$\mathbb{E}[h_m(N)^2|Y_k=y] = \frac{\mathbb{E}\left[h_m^2(y-X_k)f_N(y-X_k)\right]}{\mathbb{E}\left[f_N(y-X_k)\right]}. \tag{150}$$

Since f_N is continuous and bounded, $x \mapsto h_m(x)f_N(x)$ and $x \mapsto h_m^2(x)f_N(x)$ are both continuous and bounded functions. Therefore $\{v_k p_k\}$ is a sequence of nonnegative measurable functions converging pointwise to $v_0 p_0$. By Fatou's lemma,

$$\begin{aligned}\liminf_{k\to\infty}\mathsf{mmse}(h_m(N)|Y_k) &= \liminf_{k\to\infty}\int_{\mathbb{R}} v_k(y)p_k(y)\mathrm{d}y &(151)\\ &\geq \int_{\mathbb{R}} v_0(y)p_0(y)\mathrm{d}y &(152)\\ &= \mathsf{mmse}(h_m(N)|Y_0). &(153)\end{aligned}$$

Note that $\mathsf{var}(h_m(N)) \leq \mathsf{var}(N)$ for any m. By Lemma 1, for any k,

$$\begin{aligned}&|\mathsf{mmse}(N|Y_k) - \mathsf{mmse}(h_m(N)|Y_k)|\\ &\leq 2\sqrt{\mathsf{var}(N)}\,\|N-h_m(N)\|_2 &(154)\\ &\leq 2\sqrt{\mathsf{var}(N)\mathbb{E}\left[N^2\mathbf{1}_{\{|N|\geq m\}}\right]}. &(155)\end{aligned}$$

Plugging (155) into (153) yields

$$\begin{aligned}\liminf_{k\to\infty}\mathsf{mmse}(N|Y_k) \geq\ &\mathsf{mmse}(N|Y_0) -\\ &2\sqrt{\mathsf{var}(N)\mathbb{E}\left[N^2\mathbf{1}_{\{|N|\geq m\}}\right]}, &(156)\end{aligned}$$

which, upon sending $m \to \infty$, gives the desired (145). ∎

Appendix D
Proof of Theorem 5

First we prove an upper bound on the conditional variance, which improves the estimate in [46, Proposition 1.2]:

Lemma 4. *Let $Y = \sqrt{\mathsf{snr}}X + N_{\mathsf{G}}$, where $N_{\mathsf{G}} \sim \mathcal{N}(0,1)$ is independent of X. If $\mathsf{var}X < \infty$, then for any $y \in \mathbb{R}$,*

$$\begin{aligned}&\mathsf{var}(X|Y=y)\\ &\leq \frac{2}{\mathsf{snr}}\left(1+\log\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2\sqrt{2\pi}f_Y(y)}\right) &(157)\\ &\leq \frac{2}{\mathsf{snr}}\left(1+\log\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2}+\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2}\right), &(158)\end{aligned}$$

where $f_Y(y) = \mathbb{E}\left[\varphi(y-\sqrt{\mathsf{snr}}X)\right]$ is the probability density function of Y.

Proof: Without loss of generality we assume that $\mathbb{E}[X] = 0$. Fix $A > 0$. Then

$$\begin{aligned}
&\mathbb{E}\left[N_{\mathsf{G}}^2|Y=y\right] \\
&= \frac{\mathbb{E}\left[(y-\sqrt{\mathsf{snr}}X)^2\varphi(y-\sqrt{\mathsf{snr}}X)\right]}{f_Y(y)} && (159)\\
&\le A\,\mathbb{E}\left[(y-\sqrt{\mathsf{snr}}X)^2\right] + && (160)\\
&\quad \mathbb{E}\left[\frac{(y-\sqrt{\mathsf{snr}}X)^2\varphi(y-\sqrt{\mathsf{snr}}X)}{f_Y(y)}\mathbf{1}_{\{\varphi(y-\sqrt{\mathsf{snr}}X)>Af_Y(y)\}}\right] \\
&\le A(y^2+\mathsf{snr}\,\mathsf{var}X) + 2\log\frac{1}{\sqrt{2\pi}Af_Y(y)} && (161)
\end{aligned}$$

where (161) follows from that $\{\varphi(y-\sqrt{\mathsf{snr}}X) > \alpha\} = \{(y-\sqrt{\mathsf{snr}}X)^2 < 2\log\frac{1}{\sqrt{2\pi}\alpha}\}$. Minimizing the upper bound in (161) by choosing $A = \frac{2}{y^2+\mathsf{snr}\,\mathsf{var}X}$, we have

$$\begin{aligned}
&\mathbb{E}\left[N_{\mathsf{G}}^2|Y=y\right] \\
&\le 2\left(1+\log\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2\sqrt{2\pi}f_Y(y)}\right) && (162)\\
&\le 2\left(1+\log\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2}+\frac{y^2+\mathsf{snr}+\mathsf{var}X}{2}\right), && (163)
\end{aligned}$$

where (163) follows from Jensen's inequality:

$$\begin{aligned}
\sqrt{2\pi}f_Y(y) &= \mathbb{E}\left[\exp\left(-\frac{(y-\sqrt{\mathsf{snr}}X)^2}{2}\right)\right] && (164)\\
&\ge \exp\left(-\frac{y^2+\mathsf{snr}\,\mathsf{var}X}{2}\right). && (165)
\end{aligned}$$

The proof of (157) and (158) is completed by observing that

$$\mathsf{var}(X|Y=y) = \frac{1}{\mathsf{snr}}\mathsf{var}(N_{\mathsf{G}}|Y=y) \le \frac{1}{\mathsf{snr}}\mathbb{E}\left[N_{\mathsf{G}}^2|Y=y\right]. \quad (166)$$

■

Proof of Theorem 5: Let X and Z be jointly distributed according to their optimal coupling so that

$$\|X-Z\|_{2p} = W_{2p}(P_X,P_Z). \quad (167)$$

Let $\hat{X}(y) = \mathbb{E}\left[X|\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}=y\right]$. In view of (130) and (131), $\hat{X}$ is an increasing differentiable function with derivative

$$\begin{aligned}
0 \le \hat{X}'(y) &= \sqrt{\mathsf{snr}}\,\mathsf{var}(X|\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}=y) && (168)\\
&\le \frac{2}{\sqrt{\mathsf{snr}}}(y^2+\mathsf{snr}\,\mathsf{var}X), && (169)
\end{aligned}$$

where the upper bound follows from further weakening (158) by applying the inequality $\log(1+x) \le x$. Then

$$\begin{aligned}
&\sqrt{\mathsf{mmse}(Z,\mathsf{snr})} \\
&\le \left\|Z-\hat{X}(\sqrt{\mathsf{snr}}Z+N_{\mathsf{G}})\right\|_2 && (170)\\
&\le \left\|X-\hat{X}(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})\right\|_2 + \|X-Z\|_2 \\
&\quad + \left\|\hat{X}(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})-\hat{X}(\sqrt{\mathsf{snr}}Z+N_{\mathsf{G}})\right\|_2 && (171)\\
&= \sqrt{\mathsf{mmse}(X,\mathsf{snr})} + \|X-Z\|_2 + \\
&\quad \left\|\hat{X}(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})-\hat{X}(\sqrt{\mathsf{snr}}Z+N_{\mathsf{G}})\right\|_2 && (172)
\end{aligned}$$

where (170) follows from the suboptimality of $\hat{Z}$ for estimating X.

Next we upper bound the third term in (172): fix $w \in \mathbb{R}$ and let $f(x) = \hat{X}(\sqrt{\mathsf{snr}}x+w)$. Then for any $x, z \in \mathbb{R}$,

$$\begin{aligned}
&|\hat{X}(\sqrt{\mathsf{snr}}x+w)-\hat{X}(\sqrt{\mathsf{snr}}z+w)| \\
&= \left|\int_z^x f'(y)\mathrm{d}y\right| && (173)\\
&\le 2\left|\int_z^x (y^2+\mathsf{snr}\,\mathsf{var}X)\mathrm{d}y\right| && (174)\\
&\le \frac{2}{3}|x^3-z^3| + 2\,\mathsf{snr}\,\mathsf{var}X|x-z|, && (175)
\end{aligned}$$

where (174) follows from (169). Therefore

$$\begin{aligned}
&\left\|\hat{X}(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})-\hat{X}(\sqrt{\mathsf{snr}}Z+N_{\mathsf{G}})\right\|_2 \\
&\le \frac{2}{3}\left\|X^3-Z^3\right\|_2 + 2\,\mathsf{snr}\,\mathsf{var}X\,\|X-Z\|_2. && (176)
\end{aligned}$$

Since

$$\begin{aligned}
\left\|X^3-Z^3\right\|_2^2 &= \mathbb{E}\left[(X^3-Z^3)^2\right] && (177)\\
&\le \frac{9}{2}\mathbb{E}\left[(X-Z)^2(X^4+Z^4)\right] && (178)\\
&\le \frac{9}{2}\mathbb{E}\left[(X-Z)^{2p}\right]^{\frac{1}{p}}\mathbb{E}\left[(X^4+Z^4)^q\right]^{\frac{1}{q}} && (179)\\
&\le \frac{9}{2}\|X-Z\|_{2p}^2(\|X\|_{4q}^4+\|Z\|_{4q}^4). && (180)
\end{aligned}$$

where

- (178): by $(u^2+v^2+uv)^2 \le \frac{9}{4}(u^2+v^2)^2 \le \frac{9}{2}(u^4+v^4)$.
- (179): by Hölder's inequality with $1 \le p,q \le \infty$ and $\frac{1}{p}+\frac{1}{q}=1$.

Taking square root on both sides of (180) before applying it to (176), we have

$$\begin{aligned}
&\left\|\hat{X}(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})-\hat{X}(\sqrt{\mathsf{snr}}Z+N_{\mathsf{G}})\right\|_2 \\
&\le \sqrt{2}\,\|X-Z\|_{2p}(\|X\|_{4q}^2+\|Z\|_{4q}^2) + 2\,\mathsf{snr}\,\mathsf{var}X\,\|X-Z\|_2. && (181)
\end{aligned}$$

Substituting (167) and (181) into (172) and using the fact that $\mathsf{mmse}(X,\mathsf{snr}) \le \mathsf{var}X$, we obtain

$$\begin{aligned}
&\frac{\mathsf{mmse}(Z,\mathsf{snr})-\mathsf{mmse}(X,\mathsf{snr})}{\sqrt{\mathsf{var}X}+\sqrt{\mathsf{var}Z}} \\
&\le \sqrt{2}\,\|X-Z\|_{2p}(\|X\|_{4q}^2+\|Z\|_{4q}^2) + \\
&\quad (1+2\,\mathsf{snr}\,\mathsf{var}X)\,\|X-Z\|_2, && (182)
\end{aligned}$$

which implies (28) since $\|X-Z\|_2 \le \|X-Z\|_{2p} = W_{2p}(P_X,P_Z)$.

The W_r-Lipschitz continuity of $\mathsf{mmse}(\cdot,\mathsf{snr})$ follows from setting $r = 2p$. Since W_r-distance is finite on $\mathcal{P}_s$ for $s \ge r$, we require $4q = \frac{4r}{r-2} \ge r$, i.e., $2 \le r \le 6$.

■

Appendix E
Proof of Lemma 3

Proof: Let $\lambda = \frac{\sqrt{\mathsf{snr}}}{\sqrt{\mathsf{snr}}+1}$ and

$$Z_{\mathsf{snr}} = \lambda X + (1-\lambda)N. \quad (183)$$

In view of (44), we define a reference probability density q on $\mathbb{R}$ by

$$q(x) = \frac{1}{\int_{\mathbb{R}} \mathrm{e}^{-\psi(|x|)}\mathrm{d}x}\mathrm{e}^{-\psi(|x|)}. \tag{184}$$

Then

$$\begin{aligned}
&I(X, N, \mathsf{snr}) \\
&= h(\sqrt{\mathsf{snr}}X + N) - h(N) && (185)\\
&= \mathbb{E}\left[\log\frac{1}{q(Z_{\mathsf{snr}})}\right] + \log(1+\sqrt{\mathsf{snr}}) - D(p_{Z_{\mathsf{snr}}}\,||\,q) - h(N) && (186)\\
&\le \mathbb{E}\left[\psi(\lambda|X| + (1-\lambda)|N|)\right] + \log(1+\sqrt{\mathsf{snr}}) + \\
&\quad \log\left(\int_{\mathbb{R}} \mathrm{e}^{-\psi(|x|)}\mathrm{d}x\right) - h(N) && (187)\\
&\le a_\lambda \mathbb{E}\left[\psi(|X|)\right] + b_\lambda \mathbb{E}\left[\psi(|N|)\right] + c_\lambda + \log(1+\sqrt{\mathsf{snr}}) \\
&\quad + \log\left(\int_{\mathbb{R}} \mathrm{e}^{-\psi(|x|)}\mathrm{d}x\right) - h(N) && (188)\\
&< \infty, && (189)
\end{aligned}$$

where

- (187): by the monotonicity of ψ and the nonnegativity of relative entropy;
- (188): by (45).
- (189): by (44), (46) and (47).

Next we establish the continuity of $\mathsf{snr} \mapsto I(X, N, \mathsf{snr})$ on $\mathbb{R}_+$. In view of the weak lower-semicontinuity of relative entropy [8], we have

$$\liminf_{\gamma\to\mathsf{snr}} I(X, N, \gamma) \ge I(X, N, \mathsf{snr}). \tag{190}$$

Hence it remains to show

$$\limsup_{\gamma\to\mathsf{snr}} I(X, N, \gamma) \le I(X, N, \mathsf{snr}). \tag{191}$$

By (186), we have

$$\begin{aligned}
I(X, N, \gamma) &= \mathbb{E}\left[\psi(|Z_\gamma|)\right] + \log(1+\sqrt{\mathsf{snr}}) - h(N) + \\
&= \log\left(\int_{\mathbb{R}} \mathrm{e}^{-\psi(|x|)}\mathrm{d}x\right) - D(p_{Z_\gamma}\,||\,q). && (192)
\end{aligned}$$

For any[5] $\gamma \in (\mathsf{snr}-\epsilon, \mathsf{snr}+\epsilon)$,

$$\begin{aligned}
\psi(|Z_\gamma|) &\le \psi\left(\frac{\sqrt{\gamma}\,|X| + |N|}{\sqrt{\gamma}+1}\right) \\
&\le \max\left\{\psi\left(\frac{\sqrt{\mathsf{snr}-\epsilon}\,|X| + |N|}{\sqrt{\mathsf{snr}-\epsilon}+1}\right),\right. \\
&\quad \left.\psi\left(\frac{\sqrt{\mathsf{snr}+\epsilon}\,|X| + |N|}{\sqrt{\mathsf{snr}+\epsilon}+1}\right)\right\}. && (193)
\end{aligned}$$

As reasoned to obtain (188), the right-hand side of (193) is integrable. By the continuity of ψ and the reverse Fatou's lemma,

$$\limsup_{\gamma\to\mathsf{snr}} \mathbb{E}\left[\psi(|Z_\gamma|)\right] \le \mathbb{E}\left[\psi(|Z_{\mathsf{snr}}|)\right]. \tag{194}$$

By the lower semicontinuity of relative entropy, we have

$$\liminf_{\gamma\to\mathsf{snr}} D(p_{Z_\gamma}\,||\,q) \ge D(p_{Z_{\mathsf{snr}}}\,||\,q). \tag{195}$$

Plugging (194) and (195) into (192) yields the desired (191). ∎

[5] If $\mathsf{snr} = 0$, replace $\mathsf{snr} - \epsilon$ by 0.

ACKNOWLEDGMENT

The authors are grateful to Associate Editor Dongning Guo for a careful reading of the draft and many fruitful discussions. Especially, the proof of (8) in the case of infinite mutual information in Remark 4 is due to him. The author also thank Oliver Johnson for stimulating discussions.

REFERENCES

[1] Y. Wu and S. Verdú, "Functional properties of MMSE," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, June 2010, pp. 1453 – 1457.

[2] D. Guo, Y. Wu, S. Shamai (Shitz), and S. Verdú, "Estimation in Gaussian Noise: Properties of the Minimum Mean-square Error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371 – 2385, Apr. 2011.

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

[4] S. Yüksel and T. Linder, "Optimization and Convergence of Observation Channels in Stochastic Control," 2010, submitted to *SIAM Journal on Control and Optimization*.

[5] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual Information and Minimum Mean-Square Error in Gaussian Channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261 – 1283, Apr. 2005.

[6] V. M. Zolotarev, *One-dimensional Stable Distributions*. Providence, RI: American Mathematical Society, 1986.

[7] Y. A. Brychkov and A. P. Prudnikov, *Integral Transforms of Generalized Functions*. New York, NY: Gordon and Breach Science Publishers, 1989.

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

[9] P. J. Huber, *Robust Statistics*. New York, NY: Wiley-Interscience, 1981.

[10] G. L. Wise, "A Note on a Common Misconception in Estimation," *Systems and Control Letters*, vol. 5, pp. 355–0356, Apr. 1985.

[11] Y. Wu and S. Verdú, "MMSE dimension," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4857 – 4879, Aug. 2011.

[12] L. D. Brown, "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *Annals of Mathematical Statistics*, vol. 42, no. 3, pp. 855–903, 1971.

[13] S. Vallender, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory of Probability and its Applications*, vol. 18, pp. 784 – 786, 1974.

[14] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Vol. I: Theory*. Berlin, Germany: Springer-Verlag, 1998.

[15] Y. Wu, "Shannon theory for compressed sensing," Ph.D. dissertation, Department of Electrical Engineering, Princeton University, 2011.

[16] D. Guo, S. Shamai (Shitz), and S. Verdú, "The interplay between information and estimation measures," preprint, 2011.

[17] R. T. Rockafellar, *Convex analysis*. Princeton, NJ: Princeton University Press, 1970.

[18] 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.

[19] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[20] Y. Wu and S. Verdú, "The impact of constellation cardinality on gaussian channel capacity," in *Forty-Eighth Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, September 29 – October 1 2010, pp. 620 – 628.

[21] A. I. Sakhanenko, "Estimates in an invariance principle," *Proc. Inst. Math. Novosibirsk*, vol. 45, no. 5, pp. 27–44, 1985.

[22] E. Rio, "Upper bounds for minimal distances in the central limit theorem," *Annales de l'Institut Henri Poincaré-Probabilités et Statistiques*, vol. 45, no. 3, pp. 802–817, 2009.

[23] A. R. Barron, "Entropy and the central limit theorem," *Annals of probability*, vol. 14, no. 1, pp. 336–342, 1986.

[24] S. Artstein, K. Ball, F. Barthe, and A. Naor, "On the rate of convergence in the entropic central limit theorem," *Probability Theory and Related Fields*, vol. 129, no. 3, pp. 381–390, 2004.

[25] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probability Theory and Related Fields*, vol. 129, no. 3, pp. 391–409, 2004.

[26] A. A. Borovkov and S. A. Utev, "On an inequality and a related characterization of the normal distribution," *Theory of Probability and its Applications*, vol. 28, pp. 219 – 228, 1984.

[27] S. A. Utev, "An application of integro-differential inequalities in probability theory," *Siberian Adv. Math.*, vol. 2, pp. 164 – 199, 1992.

[28] S. G. Bobkov, G. P. Chistyakov, and F. Götze, "Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem," Apr. 2011, preprint. [Online]. Available: arxiv.org/abs/1104.3994

[29] J. M. Stoyanov, *Counterexamples in Probability*, 2nd ed. Chichester, England: Wiley, 1997.

[30] T. Tao, "Szemerédi's regularity lemma revisited," *Contributions to Discrete Mathematics*, vol. 1, no. 1, pp. 8–28, 2006.

[31] R. Ahlswede, "The final form of Tao's inequality relating conditional expectation and conditional mutual information," *Advances in Mathematics of Communications (AMC)*, vol. 1, no. 2, pp. 239–242, 2007.

[32] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., 1982.

[33] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1246–1250, May 1998.

[34] M. N. Ghosh, "Uniform approximation of minimax point estimates," *Annals of Mathematical Statistics*, vol. 35, no. 3, pp. 1031–1047, 1964.

[35] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar Gaussian channels," *Information and Control*, vol. 18, pp. 203 – 219, 1971.

[36] E. Gourdin, B. Jaumard, and B. MacGibbon, "Global optimization decomposition methods for bounded parameter minimax risk evaluation," *SIAM Journal on Scientific Computing*, vol. 15, p. 16, 1994.

[37] G. Casella and W. E. Strawderman, "Estimating a bounded normal mean," *Annals of Statistics*, pp. 870–878, 1981.

[38] M. Raginsky, "On the information capacity of Gaussian channels under small peak power constraints," in *Proceedings of the Forty-sixth Annual Allerton Conference on Communication, Control, and Computing*, 2008, pp. 286–293.

[39] P. J. Bickel, "Minimax estimation of the mean of a normal distribution when the parameter space is restricted," *Annals of Statistics*, pp. 1301–1309, 1981.

[40] P. J. Huber, "Fisher information and spline interpolation," *Annals of Statistics*, pp. 1029–1033, 1974.

[41] E. Uhrmann-Klingen, "Minimal Fisher information distributions with compact-supports," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 360–374, 1995.

[42] S. Shamai (Shitz) and S. Verdú, "Worst-case power-constrained noise for binary-input channels," *IEEE Trans. Inf. Theory*, vol. 38, no. 5, pp. 1494–1511, Sep. 1992.

[43] R. Esposito, "On a Relation between Detection and Estimation in Decision Theory," *Information and Control*, vol. 12, pp. 116–120, 1968.

[44] C. Hatsell and L. Nolte, "Some Geometric Properties of the Likelihood Ratio," *IEEE Trans. Inf. Theory*, vol. 17, no. 5, pp. 616–618, 1971.

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

[46] M. Fozunbal, "On regret of parametric mismatch in minimum mean square error estimation," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, Jun. 2010, pp. 1408–1412.