MMSE Dimension

Yihong Wu, *Student Member, IEEE*, and Sergio Verdú, *Fellow, IEEE*

***Abstract*—If N is standard Gaussian, the minimum mean-square error (MMSE) of estimating a random variable X based on $\sqrt{\mathsf{snr}}X+N$ vanishes at least as fast as $\frac{1}{\mathsf{snr}}$ as $\mathsf{snr}\to\infty$. We define the *MMSE dimension* of X as the limit as $\mathsf{snr}\to\infty$ of the product of snr and the MMSE. MMSE dimension is also shown to be the asymptotic ratio of nonlinear MMSE to linear MMSE. For discrete, absolutely continuous or mixed distribution we show that MMSE dimension equals Rényi's information dimension. However, for a class of self-similar singular X (e.g., Cantor distribution), we show that the product of snr and MMSE oscillates around information dimension periodically in snr (dB). We also show that these results extend considerably beyond Gaussian noise under various technical conditions.**

***Index Terms*—Additive noise, Bayesian statistics, Gaussian noise, high-SNR asymptotics, minimum mean-square error (MMSE), mutual information, non-Gaussian noise, Rényi information dimension.**

I. Introduction

A. Basic Setup

THE minimum mean square error (MMSE) plays a pivotal role in estimation theory and Bayesian statistics. Due to the lack of closed-form expressions for posterior distributions and conditional expectations, exact MMSE formulae are scarce. Asymptotic analysis is more tractable and sheds important insights about how the fundamental estimation-theoretic limits depend on the input and noise statistics. The theme of this paper is the high-SNR scaling law of MMSE of estimating X based on $\sqrt{\mathsf{snr}}X+N$ when N is independent of X.

The MMSE of estimating X based on Y is denoted by[1]

$$\mathsf{mmse}(X\,|\,Y)=\inf_f \mathbb{E}\left[(X-f(Y))^2\right] \tag{1}$$

$$=\mathbb{E}[(X-\mathbb{E}[X\,|\,Y])^2]=\mathbb{E}[\mathsf{var}(X\,|\,Y)], \tag{2}$$

Manuscript received December 22, 2009; revised November 11, 2010; accepted February 18, 2011. Date of current version July 29, 2011. This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1016625 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-0939370. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Austin, TX, June 2010.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: yihongwu@princeton.edu; verdu@princeton.edu).

Communicated by M. Lops, Associate Editor for Detection and Estimation.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2158905

[1]Throughout the paper, the probability space $(\Omega,\mathcal{F},\mathbb{P})$ is fixed. For $p\geq 1$, $L^p(\Omega)$ denotes the collection of all random variables defined on $(\Omega,\mathcal{F},\mathbb{P})$ with finite p^{th} moments. $L^\infty(\Omega)$ denotes the collection of all almost surely (a.s.) bounded random variables. f_X denotes the density of a random variable X whose distribution is absolutely continuous with respect to Lebesgue measure. $\mu\perp\nu$ denotes that μ and ν are mutually singular, i.e., there exists a measurable set E such that $\mu(E)=1$ and $\nu(E)=0$. N_{G} denotes a standard normal random variable. For brevity, natural logarithms are adopted and information units are nats.

where the infimum in (1) is over all Borel measurable f. When Y is related to X through an additive-noise channel with gain $\sqrt{\mathsf{snr}}$, i.e.,

$$Y=\sqrt{\mathsf{snr}}X+N \tag{3}$$

where N is independent of X, we denote

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

and, in particular, when the noise is Gaussian, we simplify

$$\mathsf{mmse}(X,\mathsf{snr})=\mathsf{mmse}(X,N_{\mathsf{G}},\mathsf{snr}). \tag{5}$$

B. Main Contributions

Before defining MMSE dimension, note that

$$0\leq\mathsf{mmse}(X,\mathsf{snr})\leq\frac{1}{\mathsf{snr}}, \tag{6}$$

where the rightmost side is the mean-square error attained by the linear estimator $f(y)=\frac{y}{\sqrt{\mathsf{snr}}}$. Therefore[2] as $\mathsf{snr}\to\infty$,

$$\mathsf{mmse}(X,\mathsf{snr})=O\left(\frac{1}{\mathsf{snr}}\right). \tag{7}$$

Seeking a finer characterization, we are interested in the exact scaling constant in (7), which depends on the distribution X only. To this end, we define the *lower and upper MMSE dimension* of X as

$$\underline{\mathscr{D}}(X)=\liminf_{\mathsf{snr}\to\infty}\mathsf{snr}\cdot\mathsf{mmse}(X,\mathsf{snr}), \tag{8}$$

$$\overline{\mathscr{D}}(X)=\limsup_{\mathsf{snr}\to\infty}\mathsf{snr}\cdot\mathsf{mmse}(X,\mathsf{snr}). \tag{9}$$

When they coincide, the common value is denoted by $\mathscr{D}(X)$, called the *MMSE dimension* of X. This information measure governs the high-SNR scaling law of $\mathsf{mmse}(X,\mathsf{snr})$ and sharpens (7) to

$$\mathsf{mmse}(X,\mathsf{snr})=\frac{\mathscr{D}(X)}{\mathsf{snr}}+o\left(\frac{1}{\mathsf{snr}}\right). \tag{10}$$

[2]We use the following asymptotic notations: $f(x)=O(g(x))$ if $\limsup\frac{|f(x)|}{|g(x)|}<\infty$, $f(x)=\Omega(g(x))$ if $g(x)=O(f(x))$, $f(x)=\Theta(g(x))$ if $f(x)=O(g(x))$ and $f(x)=\Omega(g(x))$, $f(x)=o(g(x))$ if $\lim\frac{|f(x)|}{|g(x)|}=0$, $f(x)=\omega(g(x))$ if $g(x)=o(f(x))$.

As we show in Section II-B, MMSE dimension also characterizes the high-SNR suboptimality of linear estimation.

The MMSE dimension is closely related to the *information dimension* defined by Rényi in [2]:

Definition 1 (Information Dimension): Let X be a real-valued random variable. Denote for a positive integer m the quantized version of X :

$$\langle X\rangle_m = \frac{\lfloor mX\rfloor}{m}, \tag{11}$$

where $\lfloor x\rfloor$ denotes the largest integer not exceeding x. Define

$$\underline{d}(X) = \liminf_{m\to\infty} \frac{H(\langle X\rangle_m)}{\log m} \tag{12}$$

and

$$\bar{d}(X) = \limsup_{m\to\infty} \frac{H(\langle X\rangle_m)}{\log m}, \tag{13}$$

called *lower* and *upper information dimensions* of X respectively, where $H(Z)$ denotes the entropy of a discrete random variable Z. If $\underline{d}(X) = \bar{d}(X)$, the common value is called the *information dimension* of X, denoted by $d(X)$.

Information dimension has many applications, for example, in lossless analog compression [3], quantization [4], rate-distortion theory [5] and fractal geometry [6]. Based on the integral relationship between the MMSE and mutual information in Gaussian channels [7] and the high-SNR behavior of mutual information [8], we show that the information dimensions are sandwiched between the MMSE dimensions

$$\underline{\mathscr{D}}(X) \le \underline{d}(X) \le \bar{d}(X) \le \overline{\mathscr{D}}(X). \tag{14}$$

If X is discrete, absolutely continuous or a mixture thereof, we show that (14) holds with equalities, that is, the MMSE dimension coincides with the information dimension. In view of the fact that information dimensions of discrete and absolutely continuous distributions are zero and one respectively [2, Th. 1 and 3], this implies that

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}} + o\left(\frac{1}{\mathsf{snr}}\right) \tag{15}$$

if X has an absolutely continuous distribution, and

$$\mathsf{mmse}(X,\mathsf{snr}) = o\left(\frac{1}{\mathsf{snr}}\right) \tag{16}$$

if X has a discrete distribution (even if it takes values on a dense subset of the real line).

We define the conditional MMSE dimension $\mathscr{D}(X\,|\,U)$ as the average over P_U of the MMSE dimension of $P_{X\,|\,U=u}$, and show that

$$\mathscr{D}(X) \ge \mathscr{D}(X\,|\,U), \tag{17}$$

with equality if U is a discrete random variable.

If X has a singular distribution, (14) does not hold with equalities. In fact, for self-similar inputs (e.g., the Cantor distribution [6]), we prove that the MMSE dimension does not exist; the function $\mathsf{snr}\cdot\mathsf{mmse}(X,\mathsf{snr})$ fluctuates periodically in snr (dB) around the information dimension. This periodicity originates from the self-similarity of the input distribution, and the period can be computed exactly.

C. Connections to Asymptotic Statistics

The high-SNR behavior of $\mathsf{mmse}(X,\mathsf{snr})$ is equivalent to the behavior of quadratic Bayesian risk for the Gaussian location model in the large sample limit, where P_X is the prior distribution and the sample size n plays the role of snr. To see this, let $\{N_i : i\in\mathbb{N}\}$ be a sequence of i.i.d. standard Gaussian random variables independent of X and denote $Y_i = X + N_i$ and $Y^n = (Y_1,\ldots,Y_n)$. By the sufficiency of sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ in Gaussian location models, we have

$$\mathsf{mmse}(X|Y^n) = \mathsf{mmse}(X\,|\,\bar{Y}) = \mathsf{mmse}(X,n), \tag{18}$$

where the right-hand side is the function $\mathsf{mmse}(X,\cdot)$ defined in (5) evaluated at n. Therefore as sample size grows, the Bayesian risk of estimating X vanishes as $O(\frac{1}{n})$ with the scaling constant given by the *MMSE dimension of the prior*[3]

$$\mathscr{D}(X) = \lim_{n\to\infty} n\,\mathsf{mmse}(X\,|\,Y^n). \tag{19}$$

The asymptotic expansion of $\mathsf{mmse}(X\,|\,Y^n)$ has been studied in [9]–[11] for absolutely continuous priors and general models where X and Y are not necessarily related by additive Gaussian noise. Further comparison to our results is given in Section IV-C.

D. Related Work

The low-SNR asymptotics of $\mathsf{mmse}(X,\mathsf{snr})$ has been studied extensively in [7] and [12]. In particular, it is shown in [12, Proposition 7] that if all moments of X are finite, then $\mathsf{mmse}(X,\cdot)$ is smooth on $\mathbb{R}_+$ and admits a Taylor expansion at $\mathsf{snr}=0$ up to arbitrarily high order. For example, if $\mathbb{E}[X]=0$ and $\mathbb{E}[X^2]=1$, then as $\mathsf{snr}\to 0$,

$$\mathsf{mmse}(X,\mathsf{snr}) = 1 - \mathsf{snr} + (2-(\mathbb{E}X^3)^2)\frac{\mathsf{snr}^2}{2} - (15 - 12(\mathbb{E}X^3)^2 - 6\mathbb{E}X^4 + (\mathbb{E}X^4)^2)\frac{\mathsf{snr}^3}{6} + O(\mathsf{snr}^4). \tag{20}$$

However, the asymptotics in the high-SNR regime remain underexplored in the literature. In [7, p. 1268] it is pointed out that the high-SNR behavior depends on the input distribution: For example, for binary X, $\mathsf{mmse}(X,\mathsf{snr})$ decays exponentially, while for standard Gaussian X, $\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}+1}$. Unlike the low-SNR regime, the high-SNR behavior is considerably more complicated, as it depends on the measure-theoretical structure of the input distribution rather than moments. On the other hand, it can be shown that the high-SNR asymptotics of MMSE is equivalent to the low-SNR asymptotics when the input and noise distributions are switched. As we show in Section II-D, this simple observation yields new low-SNR results for Gaussian input contaminated by non-Gaussian noise.

[3]In view of Lemma 1 in Section V, the limit of (19) is unchanged even if n takes non-integer values.

The asymptotic behavior of Fisher's information (closely related to MMSE when N is Gaussian) is conjectured in [8, p. 755] to satisfy

$$\lim_{\mathsf{snr}\to\infty} J(\sqrt{\mathsf{snr}}X + N_{\mathsf{G}}) = 1 - \bar{d}(X). \tag{21}$$

As a corollary to our results, we prove this conjecture when X has no singular components. The Cantor distribution gives a counterexample to the general conjecture (Section V-A).

Other than our scalar Bayesian setup, the weak-noise asymptotics of optimal estimation/filtering error has been studied in various regimes in statistics. One example is filtering a deterministic signal observed in weak additive white Gaussian noise (AWGN): Pinsker's theorem ([13], [14]) establishes the exact asymptotics of the optimal minimax square error when the signal belongs to a Sobolev class with finite duration. For AWGN channels and stationary Markov input processes that satisfy a stochastic differential equation, it is shown in [15, p. 372] that, under certain regularity conditions, the filtering MMSE decays as $\Theta(\frac{1}{\sqrt{\mathsf{snr}}})$.

E. Organization

Section II states the main definitions, as well as connections to linear estimation, Fisher information, and low-SNR asymptotics of MMSE. Section III gives an overview of Rényi information dimension and its applications in Shannon theory. The relationship between MMSE dimension and information dimension is shown. Section IV gives results on (conditional) MMSE dimension for various input distributions. Results about non-Gaussian noise and second-order expansion involving Fisher information are also presented. Asymptotic tightness of the Bayesian Crámer-Rao bound is discussed. Based on discrete approximation and regularity of the MMSE functional [16], some numerical experiments are shown in Section V. Section VI concludes the paper with remarks about future work. Technical proofs are relegated to the appendix.

II. MMSE Dimension

In this section we define the (conditional) MMSE dimension formally. We focus particular attention on the case of Gaussian noise.

A. Definitions

Let X, U, N be random variables with $0 < \mathsf{var}N < \infty$ and N independent of $\{X, U\}$. Define

$$\mathsf{mmse}(X, N, \mathsf{snr}\,|\,U) = \mathbb{E}[(X - \mathbb{E}[X\,|\,\sqrt{\mathsf{snr}}X + N, U])^2]. \tag{22}$$

We first note the following general inequality:[4]

$$0 \le \mathsf{mmse}(X, N, \mathsf{snr}\,|\,U) \le \frac{\mathsf{var}N}{\mathsf{snr}}. \tag{23}$$

[4]We do not impose the constraint of $\mathsf{var}N = 1$. Therefore $\frac{\mathsf{snr}}{\mathsf{var}N}$ is a dimensionless ratio that takes the role of snr in the original notation of [7, (3)], where N is assumed to be standard Gaussian.

where the rightmost side can be achieved using the affine estimator $f(y) = \frac{y - \mathbb{E}N}{\sqrt{\mathsf{snr}}}$. Therefore as $\mathsf{snr} \to \infty$, it holds that

$$\mathsf{mmse}(X, N, \mathsf{snr}|U) = O\left(\frac{1}{\mathsf{snr}}\right). \tag{24}$$

We are interested in the exact scaling constant, which depends on the distribution of X, U and N. To this end, we introduce the following notion:

Definition 2: Define the *upper* and *lower MMSE dimension* of the pair (X, N) as follows:

$$\overline{\mathscr{D}}(X, N) = \limsup_{\mathsf{snr}\to\infty} \frac{\mathsf{snr}\cdot\mathsf{mmse}(X, N, \mathsf{snr})}{\mathsf{var}N}, \tag{25}$$

$$\underline{\mathscr{D}}(X, N) = \liminf_{\mathsf{snr}\to\infty} \frac{\mathsf{snr}\cdot\mathsf{mmse}(X, N, \mathsf{snr})}{\mathsf{var}N}. \tag{26}$$

If $\overline{\mathscr{D}}(X, N) = \underline{\mathscr{D}}(X, N)$, the common value is denoted by $\mathscr{D}(X, N)$, called the *MMSE dimension* of (X, N). In particular, when N is Gaussian, we denote these limits by $\overline{\mathscr{D}}(X)$, $\underline{\mathscr{D}}(X)$, and $\mathscr{D}(X)$, called the *upper*, *lower*, and *MMSE dimension* of X respectively.

Replacing $\mathsf{mmse}(X, N, \mathsf{snr})$ by $\mathsf{mmse}(X, N, \mathsf{snr}\,|\,U)$, the *conditional MMSE dimension* of (X, N) given U can be defined similarly, denoted by $\overline{\mathscr{D}}(X, N\,|\,U)$, $\underline{\mathscr{D}}(X, N\,|\,U)$, and $\mathscr{D}(X, N\,|\,U)$ respectively. When N is Gaussian, we denote them by $\overline{\mathscr{D}}(X\,|\,U)$, $\underline{\mathscr{D}}(X\,|\,U)$, and $\mathscr{D}(X\,|\,U)$, called the *upper*, *lower*, and *conditional MMSE dimension* of X given U respectively.

The following proposition is a simple consequence of (23):

Theorem 1:

$$0 \le \underline{\mathscr{D}}(X, N\,|\,U) \le \overline{\mathscr{D}}(X, N\,|\,U) \le 1. \tag{27}$$

The next proposition shows that MMSE dimension is invariant to translations and positive scaling of input and noise.

Theorem 2: For any $\alpha, \beta, \gamma, \eta \in \mathbb{R}$, if either

- $\alpha\beta > 0$

or

- $\alpha\beta \neq 0$ and either X or N has a symmetric distribution,

then

$$\underline{\mathscr{D}}(X, N) = \underline{\mathscr{D}}(\alpha X + \gamma, \beta N + \eta), \tag{28}$$

$$\overline{\mathscr{D}}(X, N) = \overline{\mathscr{D}}(\alpha X + \gamma, \beta N + \eta). \tag{29}$$

Proof: Appendix A. ∎

B. Linear MMSE

MMSE dimension characterizes the gain achievable by nonlinear estimation over linear estimation in the high-SNR regime. To see this, define the linear MMSE (LMMSE) as

$$\mathsf{lmmse}(X\,|\,Y) = \min_{a,b\in\mathbb{R}} \mathbb{E}[(aY + b - X)^2]. \tag{30}$$

When $Y=\sqrt{\mathsf{snr}}X+N$, direct optimization over a and b reveals that the best parameters are given by

$$\hat{a}=\frac{\sqrt{\mathsf{snr}}\,\mathsf{var}X}{\mathsf{snr}\,\mathsf{var}X+\mathsf{var}N}, \tag{31}$$

$$\hat{b}=\frac{\mathbb{E}[X]\,\mathsf{var}N-\sqrt{\mathsf{snr}}\,\mathbb{E}[N]\,\mathsf{var}X}{\mathsf{snr}\,\mathsf{var}X+\mathsf{var}N}. \tag{32}$$

Hence

$$\mathsf{lmmse}(X,N,\mathsf{snr})\triangleq\mathsf{lmmse}(X\,|\,\sqrt{\mathsf{snr}}X+N) \tag{33}$$

$$=\frac{\mathsf{var}N\mathsf{var}X}{\mathsf{snr}\,\mathsf{var}X+\mathsf{var}N} \tag{34}$$

$$=\frac{\mathsf{var}N}{\mathsf{snr}}+o\left(\frac{1}{\mathsf{snr}}\right), \tag{35}$$

as $\mathsf{snr}\to\infty$. As long as $\mathsf{var}N<\infty$, the above analysis holds for any input X even if $\mathsf{var}X=\infty$, in which case (34) simplifies to $\mathsf{lmmse}(X|\sqrt{\mathsf{snr}}X+N)=\frac{\mathsf{var}N}{\mathsf{snr}}$. In view of Definition 2, (35) gives a more general definition of MMSE dimension:

$$\mathscr{D}(X,N)=\lim_{\mathsf{snr}\to\infty}\frac{\mathsf{mmse}(X,N,\mathsf{snr})}{\mathsf{lmmse}(X,N,\mathsf{snr})}, \tag{36}$$

which can be easily generalized to random vectors or processes.

C. Asymptotics of Fisher Information

In the special case of Gaussian noise it is interesting to draw conclusions on the asymptotic behavior of Fisher's information based on our results. Recall that the Fisher information (with respect to the location parameter) of a random variable Z is defined as: [17, Definition 4.1]

$$J(Z)=\sup\{|\mathbb{E}[\psi'(Z)]|^2:\psi\in C^1,\mathbb{E}[\psi^2(Z)]=1\} \tag{37}$$

where C^1 denotes the collection of all continuously differentiable functions. When Z has an absolutely continuous density f_Z, we have $J(Z)=\int\frac{f_Z'^2}{f_Z}$. Otherwise, $J(Z)=\infty$.

In view of the representation of MMSE by the Fisher information of the channel output with additive Gaussian noise [18, (1.3.4)], [7, (58)]

$$\mathsf{snr}\cdot\mathsf{mmse}(X,\mathsf{snr})=1-J(\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}) \tag{38}$$

and $J(aZ)=a^{-2}J(Z)$, letting $\epsilon=\frac{1}{\sqrt{\mathsf{snr}}}$ yields

$$\mathsf{mmse}(X,\epsilon^{-2})=\epsilon^2-\epsilon^4J(X+\epsilon N_{\mathsf{G}}). \tag{39}$$

By the lower semicontinuity of Fisher information [17, p. 79], when the distribution of X is *not* absolutely continuous, $J(X+\epsilon N_{\mathsf{G}})$ diverges as ϵ vanishes, but no faster than

$$J(X+\epsilon N_{\mathsf{G}})\le\epsilon^{-2}, \tag{40}$$

because of (39). Similarly to the MMSE dimension, we can define the *Fisher dimension* of a random variable X as follows:

$$\overline{\mathscr{J}}(X)=\limsup_{\epsilon\downarrow0}\epsilon^2\cdot J(X+\epsilon N_{\mathsf{G}}), \tag{41}$$

$$\underline{\mathscr{J}}(X)=\liminf_{\epsilon\downarrow0}\epsilon^2\cdot J(X+\epsilon N_{\mathsf{G}}). \tag{42}$$

Equation (39) shows Fisher dimension and MMSE dimension are complementary of each other:

$$\overline{\mathscr{J}}(X)+\underline{\mathscr{D}}(X)=\underline{\mathscr{J}}(X)+\overline{\mathscr{D}}(X)=1. \tag{43}$$

In [8, p.755] it is conjectured that

$$\underline{\mathscr{J}}(X)=1-\bar{d}(X) \tag{44}$$

or equivalently

$$\overline{\mathscr{D}}(X)=\bar{d}(X). \tag{45}$$

According to Theorem 5, this holds for distributions without singular components but not in general. Counterexamples can be found for singular X. See Section IV-E for more details.

D. Duality to Low-SNR Asymptotics

Note that

$$\mathsf{snr}\cdot\mathsf{mmse}(X,N,\mathsf{snr})=\mathsf{mmse}(\sqrt{\mathsf{snr}}X\,|\,\sqrt{\mathsf{snr}}X+N) \tag{46}$$

$$=\mathsf{mmse}(N\,|\,\sqrt{\mathsf{snr}}X+N) \tag{47}$$

$$=\mathsf{mmse}(N,X,\mathsf{snr}^{-1}), \tag{48}$$

which gives an equivalent definition of the MMSE dimension:

$$\mathscr{D}(X,N)=\lim_{\epsilon\downarrow0}\mathsf{mmse}(N,X,\epsilon). \tag{49}$$

This reveals an interesting duality: the high-SNR MMSE scaling constant is equal to the low-SNR limit of MMSE when the roles of input and noise are switched. Restricted to the Gaussian channel, it amounts to studying the asymptotic MMSE of estimating a Gaussian random variable contaminated with *strong noise with an arbitrary distribution*. On the other end of the spectrum, the asymptotic expansion of the MMSE of an arbitrary random variable contaminated with *strong Gaussian noise* is studied in [12, Sec. V.A]. The asymptotics of other information measures have also been studied: For example, the asymptotic Fisher information of Gaussian (or other continuous) random variables under *weak arbitrary noise* was investigated in [19]. The asymptotics of non-Gaussianness in this regime is studied in [20, Th. 1]. The second-order asymptotics of mutual information under *strong Gaussian noise* is studied in [21, Sec. IV].

Unlike $\mathsf{mmse}(X,\mathsf{snr})$ which is monotonically decreasing with snr, $\mathsf{mmse}(N_{\mathsf{G}}\,|\,\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})$ may be increasing in snr (Gaussian X), decreasing (binary-valued X) or oscillatory (Cantor X) in the high-SNR regime (see Fig. 3). In those cases in which $\mathsf{snr}\mapsto\mathsf{mmse}(N_{\mathsf{G}}\,|\,\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})$ is monotone, MMSE dimension and information dimension exist and coincide, in view of (49) and Theorem 8.

E. Noise With Infinite Variance

Although most of our focus is on square integrable random variables, the functional $\mathsf{mmse}(X,N,\mathsf{snr})$ can be defined for infinite-variance noise. Consider $X\in L^2(\Omega)$ but $N\notin L^2(\Omega)$. Then $\mathsf{mmse}(X,N,\mathsf{snr})$ is still finite but (25) and (26) cease to

make sense. Hence the scaling law in (24) could fail. It is instructive to consider the following example:

Example 1: Let X be uniformly distributed in $[0,1]$ and N have the following density:

$$f_{N_\alpha}(z) = \frac{\alpha-1}{z^\alpha}\mathbf{1}_{\{z>1\}}, \quad 1<\alpha\le 3. \tag{50}$$

Then $\mathbb{E}N_\alpha^2 = \infty$. As α decreases, the tail of N_α becomes heavier and accordingly $\mathsf{mmse}(X,N_\alpha,\mathsf{snr})$ decays slower. For instance, for $\alpha=2$ and $\alpha=3$ we obtain (see Appendix B)

$$\mathsf{mmse}(X,N_2,\mathsf{snr}) = \frac{3+\pi^2}{18}\frac{1}{\sqrt{\mathsf{snr}}} - \frac{\log^2\mathsf{snr}}{4\mathsf{snr}} + \Theta\left(\frac{\log\mathsf{snr}}{\mathsf{snr}}\right), \tag{51}$$

$$\mathsf{mmse}(X,N_3,\mathsf{snr}) = \frac{\log\mathsf{snr}}{\mathsf{snr}} - \frac{2(2+\log 2)}{\mathsf{snr}} + \Theta\left(\frac{1}{\mathsf{snr}^{3/2}}\right) \tag{52}$$

respectively. Therefore in both cases the MMSE decays strictly slower than $\frac{1}{\mathsf{snr}}$, i.e., for $N=N_2$ or N_3,

$$\mathsf{mmse}(X,N,\mathsf{snr}) = \omega\left(\frac{1}{\mathsf{snr}}\right). \tag{53}$$

However, $\mathsf{var}N=\infty$ does not always imply (53). For example, consider an arbitrary integer-valued N (none of whose moments may exist) and $0<X<1$ a.s. Then $\mathsf{mmse}(X,N,\mathsf{snr})=0$ for all $\mathsf{snr}>1$.

III. Relationship Between MMSE Dimension and Information Dimension

In this section we give an overview of Rényi information dimension and its properties, as well as its application in Shannon theory. When the noise is Gaussian, we show that the information dimension is sandwiched between the lower and upper MMSE dimension.

A. Rényi Information Dimension

The lower and upper information dimensions of a random variable are defined in Definition 1. It is shown in [3, Proposition 1] that the information dimension is finite if and only if the mild condition

$$H(\lfloor X\rfloor)<\infty \tag{54}$$

is satisfied. In particular, any X whose Shannon transform [22, Definition 2.12] exists, i.e.,

$$\mathbb{E}[\log(1+|X|)]<\infty, \tag{55}$$

has finite information dimension.

Theorem 3 ([3]): If $H(\lfloor X\rfloor)<\infty$, then

$$0\le \underline{d}(X)\le \bar{d}(X)\le 1. \tag{56}$$

If $H(\lfloor X\rfloor)=\infty$, then

$$\underline{d}(X)=\bar{d}(X)=\infty. \tag{57}$$

The information dimension of X can be understood as the *entropy rate* of the fractional part of X:

Theorem 4 ([3, Sec. III.D]): Assume that $H(\lfloor X\rfloor)<\infty$. For an integer $M\ge 2$, write the M-ary expansion of X as

$$X = \lfloor X\rfloor + \sum_{j\ge 1}(X)_j M^{-j}, \tag{58}$$

where the j^{th} digit $(X)_j = \lfloor M^j X\rfloor - M\lfloor M^{j-1}X\rfloor$ is a discrete random variable taking values in $\{0,\ldots,M-1\}$. Then $\underline{d}(X)$ and $\bar{d}(X)$ coincide with the normalized lower and upper entropy rates of the process $\{(X)_j\}$

$$\underline{d}(X) = \liminf_{m\to\infty}\frac{H((X)_1,\ldots,(X)_m)}{m\log M} \tag{59}$$

$$\bar{d}(X) = \limsup_{m\to\infty}\frac{H((X)_1,\ldots,(X)_m)}{m\log M}. \tag{60}$$

By the Lebesgue Decomposition Theorem [23], any probability distribution can be uniquely represented as the mixture of a discrete, an absolutely continuous and a singular (with respect to Lebesgue measure) probability measure. For nonsingular distributions, the information dimension can be determined as follows:

Theorem 5 ([2]): Let X be a random variable such that $H(\lfloor X\rfloor)<\infty$. Assume the distribution of X can be represented as

$$\nu = (1-\rho)\nu_{\mathrm{d}} + \rho\nu_{\mathrm{c}}, \tag{61}$$

where ν_{d} is a discrete distribution, ν_{c} is an absolutely continuous distribution and $0\le\rho\le 1$. Then

$$d(X)=\rho. \tag{62}$$

Therefore, when X has a discrete-continuous mixed distribution, the information dimension of X is exactly the weight of the continuous part. When the distribution of X has a singular component, its information dimension does *not* admit a simple formula in general. In fact if X has a singular distribution, it is possible that the information dimension does not exist [2]. However, for the important class of *self-similar* singular distributions, the information dimension can be explicitly determined [24], [5]. For example, the information dimension of the Cantor distribution is $\log_3 2$. See Section IV-E.

Let $I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})$ denote the mutual information between X and $\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}$, where X is independent of N_{G}. It is proved in [25] that $I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})<\infty$ if and only if (54) holds. Using [8, Th. 2.7 and 3.1] and [3, Appendix A],

we can relate the scaling law of mutual information under weak noise to Rényi's information dimension:[5]

Theorem 6: Let X be independent of N_{G} with $H(\lfloor X \rfloor) < \infty$. Then

$$\limsup_{\mathsf{snr}\to\infty} \frac{I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})}{\frac{1}{2}\log \mathsf{snr}} = \bar{d}(X), \tag{63}$$

$$\liminf_{\mathsf{snr}\to\infty} \frac{I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}})}{\frac{1}{2}\log \mathsf{snr}} = \underline{d}(X). \tag{64}$$

In [5] Kawabata and Dembo studied the high-rate behavior of rate-distortion function of an arbitrary random variable. Particularized to mean-square distortion, [5, Proposition 3.3] becomes:

Theorem 7: Let

$$R_X(D) = \inf_{P_{\hat{X}|X}:\mathbb{E}[(\hat{X}-X)^2]\le D} I(X;\hat{X}) \tag{65}$$

denote the rate-distortion function of X with mean-square distortion. Then

$$\limsup_{D\downarrow 0} \frac{R_X(D)}{\frac{1}{2}\log\frac{1}{D}} = \bar{d}(X), \tag{66}$$

$$\liminf_{D\downarrow 0} \frac{R_X(D)}{\frac{1}{2}\log\frac{1}{D}} = \underline{d}(X). \tag{67}$$

As a consequence of Theorems 6 and 7, when $d(X)$ exists, it holds that

$$I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}) = \frac{d(X)}{2}\log \mathsf{snr} + o(\log(\mathsf{snr})), \tag{68}$$

$$R_X(D) = \frac{d(X)}{2}\log\frac{1}{D} + o\left(\log\frac{1}{D}\right), \tag{69}$$

when $\mathsf{snr}\to\infty$ and $D\to 0$ respectively. Intuitively, (69) agrees with (68) because when the distortion goes to zero, the optimal backward channel behaves like a Gaussian channel.

B. Bounds on Information Dimension

The following theorem reveals a connection between the MMSE dimension and the information dimension of the input:

Theorem 8: If $H(\lfloor X \rfloor) < \infty$, then

$$\underline{\mathscr{D}}(X) \le \underline{d}(X) \le \bar{d}(X) \le \overline{\mathscr{D}}(X). \tag{70}$$

Therefore, if $\mathscr{D}(X)$ exists, then $d(X)$ exists and

$$\mathscr{D}(X) = d(X), \tag{71}$$

and equivalently, as $\mathsf{snr}\to\infty$,

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{d(X)}{\mathsf{snr}} + o\left(\frac{1}{\mathsf{snr}}\right). \tag{72}$$

In view of (27) and (57), (70) fails when $H(\lfloor X \rfloor) = \infty$. A Cantor-distributed X provides an example in which the inequalities in (70) are strict (see Section IV-E). However, whenever X has a discrete-continuous mixed distribution, (71) holds and the information dimension governs the high-SNR asymptotics of MMSE.

The proof of Theorem 8 hinges on two crucial results:

- The I-MMSE relationship [7]:[6]

$$\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr}) = \frac{1}{2}\mathsf{mmse}(\mathsf{snr}). \tag{73}$$

where we have used the following short-hand notations:

$$I(\mathsf{snr}) = I(X;\sqrt{\mathsf{snr}}X+N_{\mathsf{G}}), \tag{74}$$

$$\mathsf{mmse}(\mathsf{snr}) = \mathsf{mmse}(X,\mathsf{snr}). \tag{75}$$

- The high-SNR scaling law of $I(\mathsf{snr})$ in Theorem 6.

Before proceeding to the proof, we first outline a naïve attempt at proving that MMSE dimension and information dimension coincide. Assuming

$$\lim_{\mathsf{snr}\to\infty} \frac{I(\mathsf{snr})}{\log \mathsf{snr}} = \frac{d(X)}{2}, \tag{76}$$

it is tempting to apply the l'Hôpital's rule to (76) to conclude

$$\lim_{\mathsf{snr}\to\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}\mathsf{snr}} I(\mathsf{snr})}{\frac{1}{\mathsf{snr}}} = \frac{d(X)}{2}, \tag{77}$$

which, combined with (73), would produce the desired result in (71). However, this approach fails because applying l'Hôpital's rule requires establishing the existence of the limit in (77) in the first place. In fact, we show in Section IV-E when X has certain singular (e.g., Cantor) distribution, the limit in (77), i.e., the MMSE dimension, does not exist because of oscillation. Nonetheless, because mutual information is related to MMSE through an integral relation, the information dimension does exist since oscillation in MMSE is smoothed out by the integration.

In fact it is possible to construct a function $I(\mathsf{snr})$ which satisfies all the monotonicity and concavity properties of mutual information [7, Corollary 1]

$$I > 0, \quad I' > 0, \quad I'' < 0. \tag{78}$$

$$I(0) = I'(\infty) = 0, \quad I(\infty) = \infty; \tag{79}$$

yet the limit in (77) does not exist because of oscillation. For instance,

$$I(\mathsf{snr}) = \frac{d}{2}\left[\log(1+\mathsf{snr}) + \frac{1-\cos(\log(1+\mathsf{snr}))}{2}\right] \tag{80}$$

satisfies (76), (78), and (79), but (77) fails, since

$$I'(\mathsf{snr}) = \frac{d\,[2+\sin(\log(1+\mathsf{snr}))]}{4(1+\mathsf{snr})}. \tag{81}$$

[5]The results in [8] are proved under the assumption of (55), which can in fact be weakened to (54).

[6]The previous result in [7, Th. 1] requires $\mathbb{E}[X^2] < \infty$. It is shown in [25] that $H(\lfloor X \rfloor) < \infty$ suffices.

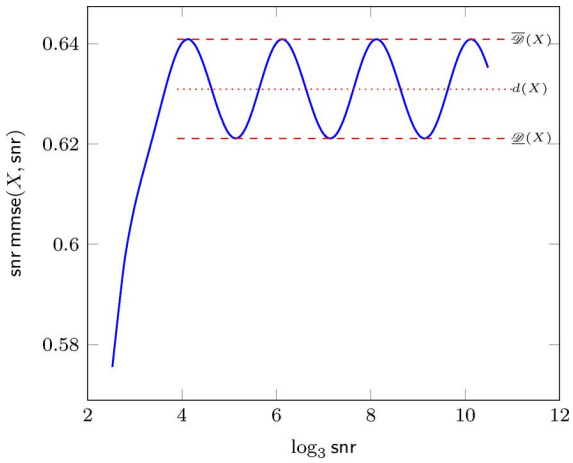

Fig. 1. Plot of $\mathsf{snr}\,\mathsf{mmse}(X,\mathsf{snr})$ against $\log_3 \mathsf{snr}$, where X has a ternary Cantor distribution.

In fact (81) gives a correct quantitative depiction of the oscillatory behavior of $\mathsf{mmse}(X,\mathsf{snr})$ for Cantor-distributed X (see Fig. 1).

Proof of Theorem 8: First we prove (70). By (73), we obtain

$$I(\mathsf{snr}) = \frac{1}{2}\int_0^{\mathsf{snr}} \mathsf{mmse}(\gamma)\mathrm{d}\gamma. \tag{82}$$

By definition of $\underline{\mathscr{D}}(X)$, for all $\delta > 0$, there exists γ_0, such that $\gamma\,\mathsf{mmse}(\gamma) \geq \underline{\mathscr{D}}(X) - \delta$ holds for all $\gamma > \gamma_0$. Then by (82), for any $\mathsf{snr} > \gamma_0$,

$$I(\mathsf{snr}) \geq \frac{1}{2}\int_0^{\gamma_0} \mathsf{mmse}(\gamma)\mathrm{d}\gamma + \frac{1}{2}\int_{\gamma_0}^{\mathsf{snr}} \frac{\underline{\mathscr{D}}(X)-\delta}{\gamma}\mathrm{d}\gamma \tag{83}$$

$$\geq \frac{1}{2}(\underline{\mathscr{D}}(X)-\delta)\log \mathsf{snr} + O(1). \tag{84}$$

In view of (64), dividing both sides by $\frac{1}{2}\log\mathsf{snr}$ and by the arbitrariness of $\delta > 0$, we conclude that $\underline{\mathscr{D}}(X) \leq \underline{d}(X)$. Similarly, $\overline{\mathscr{D}}(X) \geq \bar{d}(X)$ holds.

Next, assuming the existence of $\mathscr{D}(X)$, (71) simply follows from (70), which can also be obtained by applying l'Hôpital's rule to (76) and (77). ∎

IV. Evaluation of MMSE Dimension

In this section we drop the assumption of $H(\lfloor X \rfloor) < \infty$ and proceed to give results for various input and noise distributions.

A. Data Processing Lemma for MMSE Dimension

Theorem 9: For any X, U and any $N \in L^2(\Omega)$,

$$\overline{\mathscr{D}}(X,N) \geq \overline{\mathscr{D}}(X,N\,|\,U), \tag{85}$$

$$\underline{\mathscr{D}}(X,N) \geq \underline{\mathscr{D}}(X,N\,|\,U). \tag{86}$$

When N is Gaussian and U is discrete, (85) and (86) hold with equality.

Proof: Appendix C. ∎

In particular, Theorem 9 states that no *discrete* side information can reduce the MMSE dimension. Consequently, we have

$$\mathscr{D}(X|[X]_m) = \mathscr{D}(X) \tag{87}$$

for any $m \in \mathbb{N}$, that is, knowing arbitrarily finitely many digits X does not reduce its MMSE dimension. This observation agrees with our intuition: when the noise is weak, $[X]_m$ can be estimated with exponentially small error [see (92)], while the fractional part $X - [X]_m$ is the main contributor to the estimation error.

It is possible to extend Theorem 9 to non-Gaussian noise (e.g., uniform, exponential or Cauchy distributions) and general channels. See Remark 4 at the end of Appendix C.

B. Discrete Input

Theorem 10: If X is discrete (finitely or countably infinitely valued), and $N \in L^2(\Omega)$ whose distribution is absolutely continuous with respect to Lebesgue measure, then

$$\mathscr{D}(X,N) = 0. \tag{88}$$

In particular,

$$\mathscr{D}(X) = 0. \tag{89}$$

Proof: Since constants have zero MMSE dimension, $\mathscr{D}(X) = \mathscr{D}(X\,|\,X) = 0$, in view of Theorem 9. The more general result in (88) is proved in Appendix D. ∎

Remark 1: Theorem 10 implies that $\mathsf{mmse}(X,N,\mathsf{snr}) = o(\frac{1}{\mathsf{snr}})$ as $\mathsf{snr} \to \infty$. As observed in Example 4, MMSE can decay much faster than polynomially. Suppose the alphabet of X, denoted by $\mathcal{A} = \{x_i : i \in \mathbb{N}\}$, has no accumulation point. Then

$$d_{\min} \triangleq \inf_{i\neq j} |x_i - x_j| > 0. \tag{90}$$

If N is almost surely bounded, say $|N| \leq A$, then $\mathsf{mmse}(X,N,\mathsf{snr}) = 0$ for all $\mathsf{snr} > (\frac{2A}{d_{\min}})^2$. On the other hand, if X is almost surely bounded, say $|X| \leq \bar{A}$, then $\mathsf{mmse}(X,\mathsf{snr})$ decays exponentially: since the error probability, denoted by $p_{\mathrm{e}}(\mathsf{snr})$, of a MAP estimator for X based on $\sqrt{\mathsf{snr}}X + N_{\mathsf{G}}$ is $O(Q(\frac{\sqrt{\mathsf{snr}}}{2}d_{\min}))$, where $Q(t) = \int_t^\infty \varphi(x)\mathrm{d}x$ and φ is the standard normal density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}. \tag{91}$$

Hence

$$\mathsf{mmse}(\mathsf{snr}) \leq \bar{A}^2 p_{\mathrm{e}}(\mathsf{snr}) = O\left(\frac{1}{\sqrt{\mathsf{snr}}}\mathrm{e}^{-\frac{\mathsf{snr}}{8}d_{\min}^2}\right). \tag{92}$$

If the input alphabet has accumulation points, it is possible that the MMSE decays polynomially. For example, when X

takes values on $\{0, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\}$ and N is uniform distributed on $[0,1]$, it can be shown that $\mathsf{mmse}(X,N,\mathsf{snr}) = \Theta(\mathsf{snr}^{-2})$.

C. Absolutely Continuous Input

Theorem 11: Suppose the density of $N \in L^2(\Omega)$ is bounded and is such that for some $\alpha > 1$,

$$f_N(u) = O(|u|^{-\alpha}), \tag{93}$$

as $|u| \to \infty$. Then

$$\mathscr{D}(X,N) = 1 \tag{94}$$

holds for all X with an absolutely continuous distribution with respect to Lebesgue measure. In particular,

$$\mathscr{D}(X) = 1, \tag{95}$$

i.e.,

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}} + o\left(\frac{1}{\mathsf{snr}}\right). \tag{96}$$

Proof: Appendix E. ∎

If the density of X is sufficiently regular, then (94) holds for all noise distributions:

Theorem 12: Suppose the density of X is continuous and bounded. Then

$$\mathscr{D}(X,N) = 1 \tag{97}$$

holds for all (not necessarily absolutely continuous) $N \in L^2(\Omega)$.

Proof: Appendix E. ∎

In view of Theorem 12 and (36), we conclude that the linear estimator is dimensionally optimal for estimating absolutely continuous random variables contaminated by additive Gaussian noise, in the sense that it achieves the input MMSE dimension.

A refinement of Theorem 12 entails the second-order expansion for $\mathsf{mmse}(X,\mathsf{snr})$ for absolutely continuous input X. This involves the Fisher information of X. Suppose $J(X) < \infty$. Then

$$J(\sqrt{\mathsf{snr}}X + N_{\mathsf{G}}) \le \int \varphi(z) J(\sqrt{\mathsf{snr}}X + z)\mathrm{d}z \tag{98}$$

$$= \int \varphi(z) J(\sqrt{\mathsf{snr}}X)\mathrm{d}z \tag{99}$$

$$= \frac{J(X)}{\mathsf{snr}} < \infty, \tag{100}$$

where (98) and (99) follow from the convexity and translation invariance of Fisher information respectively. In view of (39), we have

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}} + O\left(\frac{1}{\mathsf{snr}^2}\right), \tag{101}$$

which improves (96) slightly.

Under stronger conditions the second-order term can be determined exactly. A result of Prelov and van der Meulen [19] states that: if $J(X) < \infty$ and the density of X satisfies certain regularity conditions [19, (3)–(7)], then

$$J(X + \epsilon N_{\mathsf{G}}) = J(X) + O(\epsilon). \tag{102}$$

Therefore in view of (38), we have

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}} - \frac{J(X)}{\mathsf{snr}^2} + o\left(\frac{1}{\mathsf{snr}^2}\right). \tag{103}$$

This result can be understood as follows: Stam's inequality [26] implies that

$$\frac{1}{J(\sqrt{\mathsf{snr}}X + N_{\mathsf{G}})} \ge \frac{1}{J(\sqrt{\mathsf{snr}}X)} + \frac{1}{J(N_{\mathsf{G}})} \tag{104}$$

$$= \frac{\mathsf{snr}}{J(X)} + 1. \tag{105}$$

Using (38), we have

$$\mathsf{mmse}(X,\mathsf{snr}) \ge \frac{1}{J(X) + \mathsf{snr}} \tag{106}$$

$$= \frac{1}{\mathsf{snr}} - \frac{J(X)}{\mathsf{snr}^2} + o\left(\frac{1}{\mathsf{snr}^2}\right). \tag{107}$$

Inequality (106) is also known as the Bayesian Crámer-Rao bound (or the Van Trees inequality) [27, pp. 72–73], [28, Corollary 2.3]. In view of (103), we see that (106) is asymptotically tight for sufficiently regular densities of X.

Instead of using the asymptotic expansion of Fisher information and Stam's inequality, we can show that (103) holds for a much broader class of densities of X and non-Gaussian noise:

Theorem 13: Suppose $X \in L^3(\Omega)$ with bounded density f_X whose first two derivatives are also bounded and $J(X) < \infty$. Furthermore, assume f_X satisfies the following regularity condition [29, (2.5.16)]:

$$\int_{\mathbb{R}} f_X''(x)\mathrm{d}x = 0. \tag{108}$$

Then for any $N \in L^2(\Omega)$,

$$\mathsf{mmse}(X,N,\mathsf{snr}) = \frac{\mathsf{var}N}{\mathsf{snr}} - J(X)\left(\frac{\mathsf{var}N}{\mathsf{snr}}\right)^2 + o\left(\frac{1}{\mathsf{snr}^2}\right). \tag{109}$$

Proof: Appendix E. ∎

Combined with (107), Theorem 13 implies that the Bayesian Crámer-Rao bound

$$\mathsf{mmse}(X,N,\mathsf{snr}) \ge \frac{1}{J(X) + \mathsf{snr}J(N)} \tag{110}$$

$$= \frac{1}{\mathsf{snr}\,J(N)} - \frac{J(X)}{\mathsf{snr}^2\,J^2(N)} + o\left(\frac{1}{\mathsf{snr}^2}\right) \tag{111}$$

is asymptotically tight if and only if the noise is Gaussian.

Equation (109) reveals a new *operational role* of $J(X)$. The regularity conditions imposed in Theorem 13 are much weaker and easier to check than those in [19]; however, (103) is slightly stronger than (109) because the $o(\mathsf{snr}^{-2})$ term in (103) is in fact $\mathrm{O}(\mathsf{snr}^{-\frac{5}{2}})$, as a result of (102).

It is interesting to compare (109) to results on asymptotic Bayesian risk in the large sample limit [9, Th. 5.1a]:

- When N is Gaussian, recalling (18), we have

$$\mathsf{mmse}(X|Y^n) = \mathsf{mmse}(X,n) \tag{112}$$
$$= \frac{1}{n} - \frac{J(X)}{n^2} + o\left(\frac{1}{n^2}\right). \tag{113}$$

  This agrees with the results in [9] but our proof only requires that X has a density.
- When N is non-Gaussian, under the regularity conditions in [9, Th. 5.1a], we have

$$\mathsf{mmse}(X|Y^n) = \frac{1}{n\,J(N)} - \frac{J(X)}{n^2\,J^2(N)} + o\left(\frac{1}{n^2}\right), \tag{114}$$

  which agrees with the Bayesian Crámer-Rao bound in the product case. On the other hand, by (109) we have

$$\mathsf{mmse}(X\,|\,\bar{Y}) = \mathsf{mmse}(X,n) \tag{115}$$
$$= \frac{\mathsf{var}N}{n} - J(X)\left(\frac{\mathsf{var}N}{n}\right)^2 + o\left(\frac{1}{n^2}\right). \tag{116}$$

  whose first-order term is inferior to (114), due to the Crámer-Rao bound $\mathsf{var}N \geq \frac{1}{J(N)}$. This agrees with the fact that the sample mean is asymptotically suboptimal for non-Gaussian noise, and the suboptimality is characterized by the gap in the Crámer-Rao inequality.

To conclude the discussion of absolutely continuous inputs, we give an example where (109) fails:

Example 2: Consider X and N uniformly distributed on $[0,1]$. It is shown in Appendix B that

$$\mathsf{mmse}(X,N,\mathsf{snr}) = \mathsf{var}N\left(\frac{1}{\mathsf{snr}} - \frac{1}{2\mathsf{snr}^{\frac{3}{2}}}\right), \quad \mathsf{snr} \geq 4. \tag{117}$$

Note that (109) does not hold because X does not have a differentiable density, hence $J(X) = \infty$. This example illustrates that the $o(\frac{1}{\mathsf{snr}})$ term in (96) is not necessarily $O(\frac{1}{\mathsf{snr}^2})$.

D. Mixed Distribution

Next we present results for general mixed distributions, which are direct consequences of Theorem 9. The following result asserts that MMSE dimensions are affine functionals, a fundamental property shared by Rényi information dimension [3, Th. 2].

Theorem 14:

$$\overline{\mathscr{D}}\left(\sum \alpha_i \mu_i\right) = \sum \alpha_i \overline{\mathscr{D}}(\mu_i), \tag{118}$$
$$\underline{\mathscr{D}}\left(\sum \alpha_i \mu_i\right) = \sum \alpha_i \underline{\mathscr{D}}(\mu_i). \tag{119}$$

where $\{\alpha_i : i \in \mathbb{N}\}$ is a probability mass function and each μ_i is a probability measure.

Another application of Theorem 9 is to determine the MMSE dimension of inputs with mixed distributions, which are frequently used in statistical models of *sparse signals* [3], [30]–[32]. According to Theorem 9, knowing the support does not decrease the MMSE dimension.

Corollary 1: Let $X = UZ$ where U is independent of Z, taking values in $\{0,1\}$ with $\mathbb{P}\{U=1\} = \rho$. Then

$$\overline{\mathscr{D}}(X) = \rho\overline{\mathscr{D}}(Z), \tag{120}$$
$$\underline{\mathscr{D}}(X) = \rho\underline{\mathscr{D}}(Z). \tag{121}$$

Obtained by combining Theorems 10, 11, and 14, the next result solves the MMSE dimension of discrete-continuous mixtures completely. Together with Theorem 8, it also provides an alternative proof of Rényi's theorem on information dimension (Theorem 5) via MMSE.

Theorem 15: Let X have a discrete-continuous mixed distribution as defined in (61). Then its MMSE dimension equals the weight of the absolutely continuous part, i.e.,

$$\mathscr{D}(X) = \rho. \tag{122}$$

The above results also extend to non-Gaussian noise or general channels which satisfy the condition in Remark 4 at the end of Appendix C.

To conclude this subsection, we illustrate Theorem 15 by the following examples:

Example 1 (Continuous Input): If $X \sim \mathcal{N}(0,1)$, then

$$\mathsf{mmse}(X,\mathsf{snr}) = \frac{1}{\mathsf{snr}+1} \tag{123}$$

and (95) holds.

Example 4 (Discrete Input): If X is equiprobable on $\{-1,1\}$, then by (92) or [33, Th. 3],

$$\mathsf{mmse}(X,\mathsf{snr}) = O\left(\frac{1}{\sqrt{\mathsf{snr}}}\mathrm{e}^{-\frac{\mathsf{snr}}{2}}\right) \tag{124}$$

and (89) holds.

Example 5 (Mixed Input): Let N be uniformly distributed in $[0,1]$, and let X be distributed according to an equal mixture of a mass at 0 and a uniform distribution on $[0,1]$. Then

$$\frac{\mathsf{mmse}(X,N,\mathsf{snr})}{\mathsf{var}N} = \frac{3}{2} - \frac{3}{4\sqrt{\mathsf{snr}}} + \frac{1}{\mathsf{snr}} - \frac{1}{4\mathsf{snr}^{\frac{3}{2}}} - \frac{3\sqrt{\mathsf{snr}}}{2}\log\frac{1+\sqrt{\mathsf{snr}}}{\sqrt{\mathsf{snr}}} \tag{125}$$
$$= \frac{1}{2\mathsf{snr}} + \frac{1}{4\mathsf{snr}^{\frac{3}{2}}} + o\left(\frac{1}{\mathsf{snr}^{\frac{3}{2}}}\right), \tag{126}$$

which implies $\mathscr{D}(X,N) = \frac{1}{2}$ and verifies Theorem 15 for non-Gaussian noise.

E. Singularly Continuous Distribution

In this subsection we consider atomless input distributions mutually singular with respect to Lebesgue measure. There are two new phenomena regarding MMSE dimensions of singular

inputs. First, the lower and upper MMSE dimensions $\underline{\mathscr{D}}(X,N)$ and $\overline{\mathscr{D}}(X,N)$ depend on the noise distribution, even if the noise is restricted to be absolutely continuous. Second, the MMSE dimension of a singular input need not exist. For an important class of *self-similar* singular distributions (e.g., the Cantor distribution), the function $\mathsf{snr} \mapsto \mathsf{snr}\,\mathsf{mmse}(X,\mathsf{snr})$ oscillates between the lower and upper dimension periodically in $\log \mathsf{snr}$ (i.e., in dB). This periodicity arises from the *self-similarity* of the input, and the period can be determined exactly. Unlike the lower and upper dimension, the period does not depend on the noise distribution.

We focus on a special class of inputs with self-similar distributions [24, p.36]: inputs with i.i.d. digits. Consider $X \in [0,1]$ a.s. For an integer $M \geq 2$, the M-ary expansion of X is defined in (58). Assume that the sequence $\{(X)_j : j \in \mathbb{N}\}$ is i.i.d. with common distribution P supported on $\{0,\ldots,M-1\}$. According to Theorem 4, the information dimension of X exists and is given by the normalized entropy rate of the digits

$$d(X) = \frac{H(P)}{\log M}. \tag{127}$$

For example, if X is Cantor-distributed, then the ternary expansion of X consists of i.i.d. digits, and for each j,

$$\mathbb{P}\{(X)_j = 0\} = \mathbb{P}\{(X)_j = 2\} = 1/2. \tag{128}$$

By (127), the information dimension of the Cantor distribution is $\log_3 2$. The next theorem shows that for such X the scaling constant of MMSE oscillates periodically.

Theorem 16: Let $X \in [0,1]$ a.s., whose M-ary expansion defined in (58) consists of i.i.d. digits with common distribution P. Then for any $N \in L^2(\Omega)$, there exists a $2\log M$-periodic function[7] $\Phi_{X,N} : \mathbb{R} \to [0,1]$, such that as $\mathsf{snr} \to \infty$,

$$\mathsf{mmse}(X,N,\mathsf{snr}) = \frac{\mathsf{var}N}{\mathsf{snr}}\Phi_{X,N}(\log \mathsf{snr}) + o\left(\frac{1}{\mathsf{snr}}\right). \tag{129}$$

The lower and upper MMSE dimension of (X,N) are given by

$$\overline{\mathscr{D}}(X,N) = \limsup_{b\to\infty} \Phi_{X,N}(b) = \sup_{0\leq b\leq 2\log M} \Phi_{X,N}(b), \tag{130}$$

$$\underline{\mathscr{D}}(X,N) = \liminf_{b\to\infty} \Phi_{X,N}(b) = \inf_{0\leq b\leq 2\log M} \Phi_{X,N}(b). \tag{131}$$

Moreover, when $N = N_{\mathsf{G}}$ is Gaussian, the average of $\Phi_{X,N_{\mathsf{G}}}$ over one period coincides with the information dimension of X

$$\frac{1}{2\log M}\int_0^{2\log M} \Phi_{X,N_{\mathsf{G}}}(b)\,\mathrm{d}b = d(X) = \frac{H(P)}{\log M}. \tag{132}$$

Proof: Appendix F. ∎

[7]Let $T > 0$. We say a function $f : \mathbb{R} \to \mathbb{R}$ is T-periodic if $f(x) = f(x+T)$ for all $x \in \mathbb{R}$, and T is called a period of f ([34, p. 183] or [35, Sec. 3.7]). This differs from the definition of the *least period* which is the infimum of all periods of f.

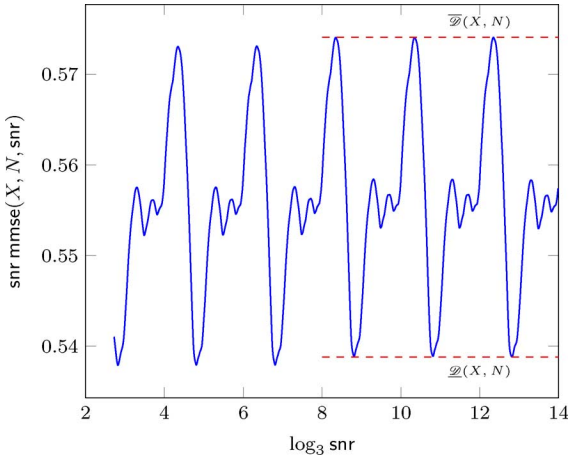

Fig. 2. Plot of $\mathsf{snr}\,\mathsf{mmse}(X,N,\mathsf{snr})$ against $\log_3 \mathsf{snr}$, where X has a ternary Cantor distribution and N is uniformly distributed in $[0,1]$.

Remark 2: Trivial examples of Theorem 16 include $\Phi_{X,N} \equiv 0$ ($X = 0$ or 1 a.s.) and $\Phi_{X,N} \equiv 1$ (X is uniformly distributed on $[0,1]$).

Theorem 16 shows that in the high-SNR regime, the function $\mathsf{snr}\,\mathsf{mmse}(X,N,\mathsf{snr})$ is periodic in snr (dB) with period $20\log M$ dB. Plots are given in Figs. 1–2. Although this reveals the oscillatory nature of $\mathsf{mmse}(X,N,\mathsf{snr})$, we do not have a general formula to compute the lower (or upper) MMSE dimension of (X,N). However, when the noise is Gaussian, Theorem 8 provides a sandwich bound in terms of the information dimension of X, which is reconfirmed by combining (130)–(132).

Remark 3 (Binary-Valued Noise): One interesting case for which we are able to compute the lower MMSE dimension corresponds to binary-valued noise, with which all singular inputs (including discrete distributions) have zero lower MMSE dimension (see Appendix H for a proof). This phenomenon can be explained by the following fact about negligible sets: a set with zero Lebesgue measure can be translated by an arbitrarily small amount to be disjoint from itself. Therefore, if an input is supported on a set with zero Lebesgue measure, we can perform a binary hypothesis test based on its noisy version, which admits a decision rule with zero error probability when SNR is large enough.

V. Numerical Results

A. Approximation by Discrete Inputs

Due to the difficulty of computing conditional expectation and estimation error in closed form, we capitalize on the regularity of the MMSE functional by computing the MMSE of successively finer discretizations of a given X. For an integer m we uniformly quantize X to $[X]_m$. Then we numerically compute $\mathsf{mmse}([X]_m,\mathsf{snr})$ for fixed snr. By the weak continuity of $P_X \mapsto \mathsf{mmse}(X,\mathsf{snr})$ [16, Corollary 3], as the quantization level m grows, $\mathsf{mmse}([X]_m,\mathsf{snr})$ converges to $\mathsf{mmse}(X,\mathsf{snr})$; however, one caveat is that to obtain the value of MMSE within a given accuracy, the quantization level needs to grow as snr grows (roughly as $\log \mathsf{snr}$) so that the quantization error is much

smaller than the noise. Lastly, due to the following result, to obtain the MMSE dimension it is sufficient to only consider integer values of snr.

Lemma 1: It is sufficient to restrict to integer values of snr when calculating $\overline{\mathscr{D}}(X, N \,|\, U)$ and $\underline{\mathscr{D}}(X, N \,|\, U)$ in (25) and (26) respectively.

Proof: Appendix G. ∎

B. Self-Similar Input Distribution

We numerically calculate the MMSE for Cantor distributed X and Gaussian noise by choosing $[X]_m = \lfloor 3^m X \rfloor 3^{-m}$. By (128), $[X]_m$ is equiprobable on the set $\mathcal{A}_m$, which has cardinality 2^m and consists of all 3-adic fractionals whose ternary digits are either 0 or 2. According to Theorem 16, in the high-SNR regime, $\mathsf{snr}\,\mathsf{mmse}(X, \mathsf{snr})$ oscillates periodically in $\log \mathsf{snr}$ with period $2 \log 3$, as plotted in Fig. 1. The lower and upper MMSE dimensions of the Cantor distribution turn out to be (to six decimals)

$$\underline{\mathscr{D}}(X) = 0.621102, \tag{133}$$

$$\overline{\mathscr{D}}(X) = 0.640861. \tag{134}$$

Note that the information dimension $d(X) = \underline{\log}_3 2 = 0.630930$ is sandwiched between $\underline{\mathscr{D}}(X)$ and $\overline{\mathscr{D}}(X)$, according to Theorem 8. From this and other numerical evidence it is tempting to conjecture that

$$d(X) = \frac{\underline{\mathscr{D}}(X) + \overline{\mathscr{D}}(X)}{2} \tag{135}$$

when the noise is Gaussian.

It should be pointed out that the sandwich bounds in (70) need not hold when N is not Gaussian. For example, in Fig. 2 where $\mathsf{snr}\,\mathsf{mmse}(X, N, \mathsf{snr})$ is plotted against $\log_3 \mathsf{snr}$ for X Cantor distributed and N uniformly distributed in $[0, 1]$, it is evident that $d(X) = \log_3 2 > \overline{\mathscr{D}}(X, N) = 0.5741$.

C. Non-Gaussian Noise

Via the input-noise duality in (48), studying high-SNR asymptotics provides insights into the behavior of $\mathsf{mmse}(X, N, \mathsf{snr})$ for non-Gaussian noise N. Combining various results from Sections IV-B, IV-C, and IV-E, we observe that $\mathsf{mmse}(X, N, \mathsf{snr})$ can behave very irregularly in general, unlike in Gaussian channels where $\mathsf{mmse}(X, \mathsf{snr})$ is strictly decreasing in snr. To illustrate this, we consider the case where standard Gaussian input is contaminated by various additive noises. For all N, it is evident that $\mathsf{mmse}(X, N, 0) = \mathsf{var} X = 1$. Due to Theorem 12, the MMSE vanishes as $\frac{\mathsf{var} N}{\mathsf{snr}}$ regardless of the noise. The behavior of MMSE associated with Gaussian, Bernoulli, and Cantor distributed noises is as follows [Fig. 3(a)].

- For standard Gaussian N, $\mathsf{mmse}(X, \mathsf{snr}) = \frac{1}{1+\mathsf{snr}}$ is *continuous* at $\mathsf{snr} = 0$ and decreases monotonically according to $\frac{1}{\mathsf{snr}}$. Recall that [16, Sec. III] this monotonicity is due to the MMSE data-processing inequality [36] and the stability of Gaussian distribution.

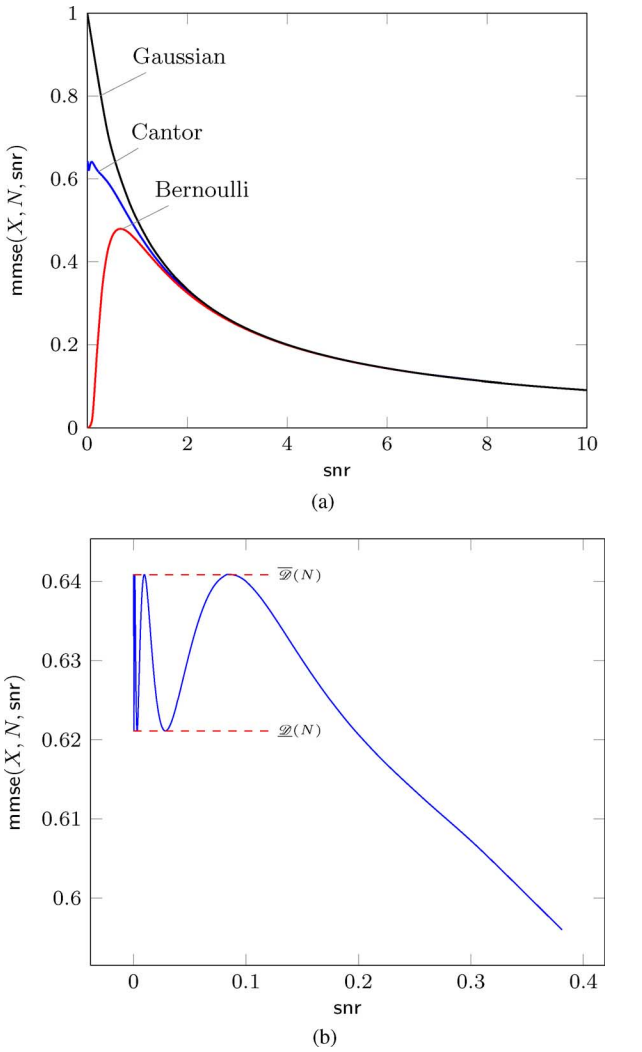

Fig. 3. Plot of $\mathsf{mmse}(X, N, \mathsf{snr})$ against snr, where X is standard Gaussian and N is standard Gaussian, equiprobable Bernoulli or Cantor distributed (normalized to have unit variance). (a) Behavior of MMSE under various noise distributions. (b) Low-SNR plot for Cantor distributed N.

- For equiprobable Bernoulli N, $\mathsf{mmse}(X, N, \mathsf{snr})$ is *discontinuous* at $\mathsf{snr} = 0$, since

$$\mathsf{var} X = \mathsf{mmse}(X, N, 0) > \lim_{\mathsf{snr} \downarrow 0} \mathsf{mmse}(X, N, \mathsf{snr}) = 0. \tag{136}$$

  As $\mathsf{snr} \to 0$, the MMSE vanishes according to $O(\frac{1}{\sqrt{\mathsf{snr}}} \mathrm{e}^{-\frac{1}{2\mathsf{snr}}})$, in view of (48) and (92), and since it also vanishes as $\mathsf{snr} \to \infty$, it is not monotonic with $\mathsf{snr} > 0$.
- For Cantor distributed N, $\mathsf{mmse}(X, N, \mathsf{snr})$ is also *discontinuous* at $\mathsf{snr} = 0$. According to Theorem 16, as $\mathsf{snr} \to 0$, MMSE oscillates relentlessly and does not have a limit [See the zoom-in plot in Fig. 3(b)].

VI. Conclusion

Through the high-SNR asymptotics of MMSE in Gaussian channels, we defined a new information measure called MMSE dimension. Although stemming from estimation-theoretic principles, MMSE dimension shares several important features with

Rényi's information dimension. By Theorem 9 and [3, Th. 2], they are both affine functionals. According to Theorem 8, information dimension is sandwiched between the lower and upper MMSE dimensions. For distributions with no singular components, they coincide to be the weight of the continuous part of the distribution. The high-SNR scaling law of mutual information and MMSE in Gaussian channels are governed by the information dimension and the MMSE dimension respectively. In [3], we have shown that the information dimension plays a pivotal role in almost lossless analog compression, an information-theoretic model for noiseless compressed sensing. In fact we have shown in [37] that the MMSE dimension is closed related to the fundamental limit in noisy compressed sensing with stable reconstruction.

Characterizing the high-SNR suboptimality of linear estimation, (36) provides an alternative definition of MMSE dimension, which enables us to extend our results to random vectors or processes. In these more general setups, it is interesting to investigate how the *causality* constraint of the estimator affects the high-SNR behavior of the optimal estimation error. Another direction of generalization is to study the high-SNR asymptotics of MMSE with a *mismatched* model in the setup of [38] or [39].

Appendix A
Proof of Theorem 2

Proof: Invariance of the MMSE functional under translation is obvious. Hence for any $\alpha, \beta \neq 0$,

$$\begin{aligned}
&\mathsf{mmse}(\alpha X+\gamma, \beta N+\eta, \mathsf{snr}) \\
=&\mathsf{mmse}(\alpha X, \beta N, \mathsf{snr}) && (137)\\
=&\mathsf{mmse}(\alpha X|\alpha\sqrt{\mathsf{snr}}X+\beta N) && (138)\\
=&|\alpha|^2\mathsf{mmse}\left(X\Big|\frac{|\alpha|}{|\beta|}\sqrt{\mathsf{snr}}X+\mathrm{sgn}(\alpha\beta)N\right) && (139)\\
=&|\alpha|^2\mathsf{mmse}\left(X, \mathrm{sgn}(\alpha\beta)N, \frac{|\alpha|}{|\beta|}\sqrt{\mathsf{snr}}\right) && (140)\\
=&|\alpha|^2\mathsf{mmse}\left(\mathrm{sgn}(\alpha\beta)X, N, \frac{|\alpha|}{|\beta|}\sqrt{\mathsf{snr}}\right). && (141)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\underline{\mathscr{D}}(\alpha X+\gamma, \beta N+\eta)\\
=&\liminf_{\mathsf{snr}\to\infty}\frac{\mathsf{snr}\cdot\mathsf{mmse}(\alpha X+\gamma, \beta N+\eta, \mathsf{snr})}{\mathsf{var}(\beta N+\eta)} && (142)\\
=&\liminf_{\mathsf{snr}\to\infty}\frac{\mathsf{snr}\cdot|\alpha|^2\mathsf{mmse}\left(X, \mathrm{sgn}(\alpha\beta)N, \frac{|\alpha|}{|\beta|}\sqrt{\mathsf{snr}}\right)}{|\beta|^2\mathsf{var}N} && (143)\\
=&\underline{\mathscr{D}}(X, \mathrm{sgn}(\alpha\beta)N) && (144)\\
=&\underline{\mathscr{D}}(\mathrm{sgn}(\alpha\beta)X, N), && (145)
\end{aligned}$$

where (144) and (145) follow from (140) and (141) respectively. The claims in Theorem 2 are special cases of (144) and (145). The proof for $\overline{\mathscr{D}}$ follows analogously. ∎

Appendix B
Calculation of (51), (52), and (117)

In this appendix we compute $\mathsf{mmse}(X, N, \mathsf{snr})$ for three different pairs of (X, N).

First we show (52), where X is uniformly distributed in $[0, 1]$ and N has the density in (50) with $\alpha = 3$. Let $\epsilon = \frac{1}{\sqrt{\mathsf{snr}}}$. Then $\mathbb{E}[X|X+\epsilon N = y] = \frac{q_1}{q_0}(y)$, where

$$\begin{aligned}
q_0(y) =&\frac{1}{\epsilon}\mathbb{E}\left[f_N\left(\frac{y-x}{\epsilon}\right)\right] && (146)\\
=&\begin{cases}1-\frac{\epsilon^2}{y^2} & \epsilon < y < 1+\epsilon\\ \frac{\epsilon^2(-1+2y)}{(-1+y)^2y^2} & y > 1+\epsilon.\end{cases} && (147)
\end{aligned}$$

and

$$\begin{aligned}
q_1(y) =&\frac{1}{\epsilon}\mathbb{E}\left[Xf_N\left(\frac{y-x}{\epsilon}\right)\right] && (148)\\
=&\begin{cases}\frac{2\epsilon^{\frac{3}{2}}}{\sqrt{y}}+y-3\epsilon & \epsilon < y < 1+\epsilon\\ \frac{2\epsilon^{\frac{3}{2}}}{\sqrt{y}}+\frac{(3-2y)\epsilon^{\frac{3}{2}}}{(y-1)^{\frac{3}{2}}} & y > 1+\epsilon.\end{cases} && (149)
\end{aligned}$$

Then

$$\begin{aligned}
&\mathsf{mmse}(X, N, \mathsf{snr})\\
=&\mathbb{E}[X^2]-\mathbb{E}[(\mathbb{E}[X\mid X+\epsilon N])^2] && (150)\\
=&\mathbb{E}[X^2]-\int_0^\infty\frac{q_1^2(y)}{q_0(y)}\mathrm{d}y && (151)\\
=&2\epsilon^2\left[\log\left(1+\frac{1}{2\epsilon}\right)-2+8\epsilon\coth^{-1}(1+4\epsilon)\right], && (152)
\end{aligned}$$

where we have used $\mathbb{E}[X^2] = \frac{1}{3}$. Taking Taylor expansion on (152) at $\epsilon = 0$ yields (52). For $\alpha = 2$, (51) can be shown in similar fashion.

To show (117), where X and N are both uniformly distributed in $[0, 1]$, we note that

$$\begin{aligned}
q_0(y) =&\frac{1}{\epsilon}[\min\{y, 1\}-(y-\epsilon)^+] && (153)\\
q_1(y) =&\frac{1}{2\epsilon}[\min\{y^2, 1\}-((y-\epsilon)^+)^2], && (154)
\end{aligned}$$

where $(x)^+ \triangleq \max\{x, 0\}$. Then (117) can be obtained using (151).

Appendix C
Proof of Theorem 9

A. Outline

From

$$\mathsf{mmse}(X, \mathsf{snr}|U) \leq \mathsf{mmse}(X, \mathsf{snr}), \tag{155}$$

we immediately obtain the inequalities in (85) and (86). Next we prove that equalities hold if U is discrete. Let $\mathcal{U} = \{u_i : i = 1, \ldots, n\}$ denote the alphabet of U with $n \in \mathbb{N}\cup\{\infty\}$, $\alpha_i = \mathbb{P}\{U = u_i\}$. Denote by μ_i the distribution of X given

$U = u_i$. Then the distribution of X is given by the following mixture:

$$\mu = \sum_{i=1}^{n} \alpha_i \mu_i. \tag{156}$$

Our goal is to establish

$$\overline{\mathscr{D}}(\mu) = \sum_{i=1}^{n} \alpha_i \overline{\mathscr{D}}(\mu_i), \tag{157}$$

$$\underline{\mathscr{D}}(\mu) = \sum_{i=1}^{n} \alpha_i \underline{\mathscr{D}}(\mu_i). \tag{158}$$

After recalling an important lemma due to Doob in Appendix C-B, we decompose the proof of (157) and (158) into four steps, which are presented in Appendixes C-C–C-F respectively:

1) We prove the special case of $n = 2$ and $\mu_1 \perp \mu_2$;
2) We consider $\mu_1 \ll \mu_2$;
3) Via the Hahn-Lebesgue decomposition and induction on n, the conclusion is extended to any finite mixture;
4) We prove the most general case of countable mixture ($n = \infty$).

B. Doob's Relative Limit Theorem

The following lemma is a combination of [40, Exercise 2.9, p. 243] and [41, Th. 1.6.2, p. 40]:

Lemma 2: Let μ and ν be two Radon measures on $\mathbb{R}^n$. Define the density of μ with respect to ν by

$$\frac{\mathrm{D}\mu}{\mathrm{D}\nu}(x) = \lim_{\epsilon \downarrow 0} \frac{\mu(B(x,\epsilon))}{\nu(B(x,\epsilon))}, \tag{159}$$

where $B(x,\epsilon)$ denotes the open ball of radius ϵ centered at x. If $\mu \perp \nu$, then

$$\frac{\mathrm{D}\mu}{\mathrm{D}\nu} = 0, \quad \nu\text{–a.e.} \tag{160}$$

If $\mu \ll \nu$, then

$$\frac{\mathrm{D}\mu}{\mathrm{D}\nu} = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}, \quad \mu\text{–a.e.} \tag{161}$$

The Lebesgue-Besicovitch differentiation theorem is a direct consequence of Lemma 2:

Lemma 3 ([41, Corollary 1.7.1]): Let ν be a Radon measure on $\mathbb{R}^n$ and $g \in L^1_{\mathrm{loc}}(\mathbb{R}^n, \nu)$. Then

$$\lim_{\epsilon \downarrow 0} \frac{1}{\nu(B(x,\epsilon))} \int_{B(x,\epsilon)} |g(y) - g(x)| \nu(\mathrm{d}y) = 0 \tag{162}$$

holds for ν-a.e. $x \in \mathbb{R}^n$.

It is instructive to reformulate Lemma 2 in a probabilistic context: Suppose μ and ν are probability measures and X and Z are random variables distributed according to μ and ν respectively. Let N be uniformly distributed in $[-1,1]$ and independent of $\{X, Z\}$. Then $X + \epsilon N$ has the following density:

$$f_{X+\epsilon N}(x) = \frac{1}{2\epsilon} \mu(B(x,\epsilon)), \tag{163}$$

hence the density of μ with respect to ν can be written as

$$\frac{\mathrm{D}\mu}{\mathrm{D}\nu}(x) = \lim_{\epsilon \downarrow 0} \frac{f_{X+\epsilon N}(x)}{f_{Z+\epsilon N}(x)}. \tag{164}$$

A natural question is whether (164) still holds if N has a non-uniform distribution. In [42, Th. 4.1], Doob gave a sufficient condition for this to be true, which is satisfied in particular by Gaussian-distributed N [42, Th. 5.2]:

Lemma 4: For any $z \in \mathbb{R}$, let $\varphi_z(\cdot) = \varphi(z + \cdot)$. Under the assumption of Lemma 2, if

$$\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \mu(\mathrm{d}y), \int \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y) \tag{165}$$

are finite for all $\epsilon > 0$ and $x \in \mathbb{R}$, then

$$\frac{\mathrm{D}\mu}{\mathrm{D}\nu}(x) = \lim_{\epsilon \downarrow 0} \frac{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \mu(\mathrm{d}y)}{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y)} \tag{166}$$

holds for ν-a.e. x.

Consequently we have counterparts of Lemmas 2 and 3:

Lemma 5: Under the condition of Lemma 2, if $\mu \perp \nu$, then

$$\lim_{\epsilon \downarrow 0} \frac{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \mu(\mathrm{d}y)}{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y)} = 0, \quad \nu\text{–a.e.} \tag{167}$$

If $\mu \ll \nu$, then

$$\lim_{\epsilon \downarrow 0} \frac{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \mu(\mathrm{d}y)}{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y)} = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}, \quad \mu\text{–a.e.} \tag{168}$$

Lemma 6: Under the condition of Lemma 3,

$$\lim_{\epsilon \downarrow 0} \frac{\int |g(y) - g(x)| \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y)}{\int \varphi_z\left(\frac{y-x}{\epsilon}\right) \nu(\mathrm{d}y)} = 0 \tag{169}$$

holds for ν-a.e. x.

C. Mixture of Two Mutually Singular Measures

We first present a lemma which enables us to truncate the input or the noise. The point of this result is that the error term depends only on the truncation threshold but not on the observation.

Lemma 7: For $K > 0$ define $X_K = X\mathbf{1}_{\{|X| \le K\}}$ and $\bar{X}_K = X - X_K$. Then for all Y,

$$|\mathsf{mmse}(X_K \,|\, Y) - \mathsf{mmse}(X \,|\, Y)| \le 3\|\bar{X}_K\|_2 \|X\|_2. \tag{170}$$

Proof:

$$\begin{aligned}
&\mathsf{mmse}(X_K\,|\,Y)\\
&= \|X_K - \mathbb{E}[X_K\,|\,Y]\|_2^2 && (171)\\
&\le \left(\|X - \mathbb{E}[X\,|\,Y]\|_2 + \left\|\bar{X}_K - \mathbb{E}[\bar{X}_K|Y]\right\|_2\right)^2 && (172)\\
&\le \left(\mathsf{mmse}(X|Y)^{\frac{1}{2}} + \sqrt{\mathsf{var}\bar{X}_K}\right)^2 && (173)\\
&= \mathsf{mmse}(X|Y) + \mathsf{var}\bar{X}_K + 2\sqrt{\mathsf{var}\bar{X}_K \mathsf{mmse}(X|Y)} && (174)\\
&\le \mathsf{mmse}(X|Y) + 3\left\|\bar{X}_K\right\|_2 \|X\|_2. && (175)
\end{aligned}$$

The other direction of (170) follows in entirely analogous fashion. ∎

Let X_i be a random variable with distribution μ_i, for $i = 1, 2$. Let U be a random variable independent of $\{X_1, X_2\}$, taking values on $\{1, 2\}$ with probability $0 < \alpha_1 < 1$ and $\alpha_2 = 1 - \alpha_2$ respectively. Then the distribution of X_U is

$$\mu = \alpha_1\mu_1 + \alpha_2\mu_2. \tag{176}$$

Fixing $M > 0$, we define

$$\begin{aligned}
g_{u,\epsilon}(y) =& \mathbb{E}\left[\varphi\left(\frac{y - X_u}{\epsilon}\right)\right], && (177)\\
f_{u,\epsilon}(y) =& \mathbb{E}\left[\frac{y - X_u}{\epsilon}\varphi\left(\frac{y - X_u}{\epsilon}\right)\mathbf{1}_{\{|\frac{y-X_u}{\epsilon}|\le M\}}\right] && (178)
\end{aligned}$$

and

$$\begin{aligned}
g_\epsilon =& \alpha_1 g_{1,\epsilon} + \alpha_2 g_{2,\epsilon}, && (179)\\
f_\epsilon =& \alpha_1 f_{1,\epsilon} + \alpha_2 f_{2,\epsilon}. && (180)
\end{aligned}$$

Then the densities of $Y_u = X_u + \epsilon N_{\mathsf{G}}$ and $Y_U = X_U + \epsilon N_{\mathsf{G}}$ are respectively given by

$$\begin{aligned}
q_{u,\epsilon}(y) &= \frac{1}{\epsilon} g_{u,\epsilon}(y) && (181)\\
q_\epsilon(y) &= \frac{1}{\epsilon} g_\epsilon(y). && (182)
\end{aligned}$$

We want to show

$$\mathsf{mmse}(X_U, \mathsf{snr}) - \mathsf{mmse}(X_U, \mathsf{snr}|U) = o\left(\frac{1}{\mathsf{snr}}\right), \tag{183}$$

i.e., the benefit of knowing the true distribution is merely $o(\frac{1}{\mathsf{snr}})$ in the high-SNR regime. To this end, let $\epsilon = \frac{1}{\sqrt{\mathsf{snr}}}$ and define

$$\begin{aligned}
W &= N_{\mathsf{G}}\mathbf{1}_{\{|N_{\mathsf{G}}|\le M\}}, && (184)\\
A_M(\epsilon) &= \mathsf{mmse}(W\,|\,Y_U) - \mathsf{mmse}(W\,|\,Y_U, U) \ge 0. && (185)
\end{aligned}$$

By the orthogonality principle,

$$A_M(\epsilon) = \mathbb{E}[(W - \hat{W}(Y_U))^2] - \mathbb{E}[(W - \hat{W}(Y_U, U))^2] \tag{186}$$

$$= \mathbb{E}[(\hat{W}(Y_U) - \hat{W}(Y_U, U))^2] \tag{187}$$

where

$$\hat{W}(y, u) = \mathbb{E}[W|Y_U = y, U = u] = \frac{f_{u,\epsilon}(y)}{g_{u,\epsilon}(y)} \tag{188}$$

and

$$\hat{W}(y) = \mathbb{E}[W|Y_U = y] = \frac{f_\epsilon(y)}{g_\epsilon(y)}. \tag{189}$$

Therefore,

$$\begin{aligned}
&A_M(\epsilon)\\
&= \sum_{u=1}^{2} \alpha_u \mathbb{E}[(\hat{W}(Y_U, u) - \hat{W}(Y_U))^2|U = u] && (190)\\
&= \sum_{u=1}^{2} \alpha_u \mathbb{E}\left[\left(\frac{f_{u,\epsilon}}{g_{u,\epsilon}} - \frac{f_\epsilon}{g_\epsilon}\right)^2 (Y_U)\middle| U = u\right] && (191)\\
&= \alpha_1\alpha_2^2 \mathbb{E}\left[\left(\frac{f_{1,\epsilon}g_{2,\epsilon} - f_{2,\epsilon}g_{1,\epsilon}}{g_{1,\epsilon}g_\epsilon}\right)^2 (Y_1)\right] +\\
&\quad \alpha_2\alpha_1^2 \mathbb{E}\left[\left(\frac{f_{1,\epsilon}g_{2,\epsilon} - f_{2,\epsilon}g_{1,\epsilon}}{g_{2,\epsilon}g_\epsilon}\right)^2 (Y_2)\right] && (192)\\
&= \frac{\alpha_1\alpha_2}{\epsilon}\int \frac{(f_{1,\epsilon}g_{2,\epsilon} - f_{2,\epsilon}g_{1,\epsilon})^2}{g_\epsilon g_{1,\epsilon} g_{2,\epsilon}} \mathrm{d}y && (193)\\
&= \alpha_1\alpha_2 \mathbb{E}\left[\frac{g_{1,\epsilon}(Y_U)g_{2,\epsilon}(Y_U)}{g_\epsilon^2(Y_U)}(\hat{W}(Y_U, 1) - \hat{W}(Y_U, 2))^2\right] && (194)
\end{aligned}$$

where

- (191): by (188) and (189).
- (192): by (179) and (180).
- (193): by (181) and (182).

Next we show that as $\epsilon \to 0$, the quantity defined in (185) vanishes

$$A_M(\epsilon) = o(1). \tag{195}$$

Indeed,

$$\begin{aligned}
A_M(\epsilon) &\le 4M^2\alpha_1\alpha_2 \mathbb{E}\left[\frac{g_{1,\epsilon}g_{2,\epsilon}}{g_\epsilon^2} \circ Y_U\right] && (196)\\
&\le 4M^2\alpha_1\alpha_2 \left(\mathbb{E}\left[\frac{g_{2,\epsilon}}{g_\epsilon} \circ Y_1\right] + \mathbb{E}\left[\frac{g_{1,\epsilon}}{g_\epsilon} \circ Y_2\right]\right), && (197)
\end{aligned}$$

where

- (196): by (194) and

$$|\hat{W}(y, u)| \le M \tag{198}$$

- (197): by (179) and (180), we have for $u = 1, 2$,

$$\frac{g_{u,\epsilon}}{g_\epsilon} \le \frac{1}{\alpha_u}. \tag{199}$$

Write

$$\mathbb{E}\left[\frac{g_{2,\epsilon}}{g_\epsilon}\circ Y_1\right]=\int\varphi(z)\mathrm{d}z\int\frac{g_{2,\epsilon}}{g_\epsilon}(x+\epsilon z)\mu_1(\mathrm{d}x). \quad (200)$$

Fix z. Since

$$\frac{g_{1,\epsilon}}{g_{2,\epsilon}}(x+\epsilon z)=\frac{\mathbb{E}\left[\varphi_z\left(\frac{y-X_1}{\epsilon}\right)\right]}{\mathbb{E}\left[\varphi_z\left(\frac{y-X_2}{\epsilon}\right)\right]}, \quad (201)$$

and $\mu_1\perp\mu_2$, applying Lemma 5 yields

$$\lim_{\epsilon\downarrow0}\frac{g_{2,\epsilon}}{g_{1,\epsilon}}(x+\epsilon z)=0, \quad (202)$$

for μ_1-a.e. x. Therefore,

$$\frac{g_{2,\epsilon}}{g_\epsilon}(x+\epsilon z)=o(1) \quad (203)$$

for μ_1-a.e. x. In view of (200) and (199), we have

$$\mathbb{E}\left[\frac{g_{2,\epsilon}}{g_\epsilon}\circ Y_1\right]=o(1) \quad (204)$$

by the dominated convergence theorem, and in entirely analogous fashion

$$\mathbb{E}\left[\frac{g_{1,\epsilon}}{g_\epsilon}\circ Y_2\right]=o(1). \quad (205)$$

Substituting (204) and (205) into (197), we obtain (195).

By Lemma 7,

$$\begin{aligned}&\limsup_{\epsilon\downarrow0}[\mathsf{mmse}(N_\mathsf{G}\,|\,Y_U)-\mathsf{mmse}(N_\mathsf{G}\,|\,Y_U,U)]\\ &\le\lim_{\epsilon\downarrow0}A_M(\epsilon)+6\|N_\mathsf{G}\|_2\|N_\mathsf{G}\mathbf{1}_{\{|N_\mathsf{G}|>M\}}\|_2 &(206)\\ &=6\left\|N_\mathsf{G}\mathbf{1}_{\{|N_\mathsf{G}|>M\}}\right\|_2. &(207)\end{aligned}$$

By the arbitrariness of M, letting $\epsilon=\frac{1}{\sqrt{\mathsf{snr}}}$ yields

$$\begin{aligned}0&=\limsup_{\epsilon\downarrow0}[\mathsf{mmse}(N_\mathsf{G}\,|\,Y_U)-\mathsf{mmse}(N_\mathsf{G}\,|\,Y_U,U)] &(208)\\ &=\limsup_{\mathsf{snr}\downarrow0}\mathsf{snr}\,(\mathsf{mmse}(X_U,\mathsf{snr})-\mathsf{mmse}(X_U,\mathsf{snr}|U)), &(209)\end{aligned}$$

which gives the desired result (183).

D. Mixture of Two Absolutely Continuous Measures

Now we assume $\mu_1\ll\mu_2$. In view of the proof in Appendix C-C, it is sufficient to show (195). Denote the Radon-Nikodym derivative of μ_1 with respect to μ_2 by

$$h=\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}, \quad (210)$$

where $h:\mathbb{R}\to\mathbb{R}_+$ satisfies $\int h\mathrm{d}\mu_2=1$.

From (193), we have

$$\begin{aligned}&A_M(\epsilon)\\ &=\alpha_1\alpha_2\mathbb{E}\left[\frac{g_{1,\epsilon}}{g_\epsilon}\left(\frac{f_{1,\epsilon}}{g_{1,\epsilon}}-\frac{f_{2,\epsilon}}{g_{2,\epsilon}}\right)^2\circ Y_2\right] &(211)\\ &=\alpha_1\alpha_2\int\varphi(z)\mathrm{d}z\int_{F^\mathrm{c}}\frac{g_{1,\epsilon}}{g_\epsilon}\left(\frac{f_{1,\epsilon}}{g_{1,\epsilon}}-\frac{f_{2,\epsilon}}{g_{2,\epsilon}}\right)^2(x+\epsilon z)\mu_2(\mathrm{d}x) &(212)\\ &+\alpha_1\alpha_2\int\varphi(z)\mathrm{d}z\int_F\frac{(f_{1,\epsilon}g_{2,\epsilon}-f_{2,\epsilon}g_{1,\epsilon})^2}{g_\epsilon g_{1,\epsilon}g_{2,\epsilon}^2}(x+\epsilon z)\mu_2(\mathrm{d}x), &(213)\end{aligned}$$

where

$$F=\{x:h(x)>0\}. \quad (214)$$

If $h(x)=0$, applying Lemma 5 we obtain

$$\frac{g_{1,\epsilon}}{g_\epsilon}(x+\epsilon z)=o(1) \quad (215)$$

for μ_2-a.e. x and every z. In view of (198), we conclude that the integrand in (212) is also $o(1)$.

If $h(x)>0$, then

$$\begin{aligned}&|(f_{1,\epsilon}g_{2,\epsilon}-f_{2,\epsilon}g_{1,\epsilon})(x+\epsilon z)| &(216)\\ &=\left|\int\frac{y-x_1}{\epsilon}\varphi_z\left(\frac{y-x_1}{\epsilon}\right)\mathbf{1}_{\{|\frac{y-x_1}{\epsilon}|\le M\}}h(x_1)\mu_2(\mathrm{d}x_1)\times\right.\\ &\quad\int\varphi_z\left(\frac{y-x_2}{\epsilon}\right)\mu_2(\mathrm{d}x_2)-\int\frac{y-x_1}{\epsilon}\varphi_z\left(\frac{y-x_1}{\epsilon}\right)\times\\ &\quad\left.\mathbf{1}_{\{|\frac{y-x_1}{\epsilon}|\le M\}}\mu_2(\mathrm{d}x_1)\int\varphi_z\left(\frac{y-x_2}{\epsilon}\right)h(x_2)\mu_2(\mathrm{d}x_2)\right| &(217)\\ &\le M\iint|h(x_1)-h(x_2)|\varphi_z\left(\frac{y-x_1}{\epsilon}\right)\times\\ &\quad\varphi_z\left(\frac{y-x_2}{\epsilon}\right)\mu_2(\mathrm{d}x_1)\mu_2(\mathrm{d}x_2) &(218)\\ &\le 2Mg_{2,\epsilon}(x+\epsilon z)\int|h(x_1)-h(x)|\varphi_z\left(\frac{y-x_1}{\epsilon}\right)\mu_2(\mathrm{d}x_1) &(219)\\ &=o\left(g_{2,\epsilon}^2(x+\epsilon z)\right) &(220)\end{aligned}$$

for μ_2-a.e. x and every z, which follows from applying Lemma 6 to $h\in L^1(\mathbb{R},\mu_2)$. By Lemma 5, we have

$$\lim_{\epsilon\downarrow0}\frac{g_{2,\epsilon}^2}{g_{1,\epsilon}g_\epsilon}(x)=\frac{1}{h(x)(\alpha_1+\alpha_2h(x))}. \quad (221)$$

Combining (220) and (221) yields that the integrand in (213) is $o(1)$. Since the integrand is bounded by $\frac{4M^2}{\alpha_1}$, (183) follows from applying the dominated convergence theorem to (211).

E. Finite Mixture

Dealing with the more general case where μ_1 and μ_2 are arbitrary probability measures, we perform the Hahn-Lebesgue decomposition [41, Th. 1.6.3] on μ_1 with respect to μ_2, which yields

$$\mu_1 = \beta_1 \nu_{\mathrm{c}} + \beta_2 \nu_{\mathrm{s}}, \tag{222}$$

where $0 \le \beta_1 = 1 - \beta_2 \le 1$ and ν_{c} and ν_{s} are two probability measures such that $\nu_{\mathrm{c}} \ll \mu_2$ and $\nu_{\mathrm{s}} \perp \mu_2$. Consequently, $\nu_{\mathrm{s}} \perp \nu_{\mathrm{c}}$. Therefore,

$$\begin{aligned} \mu &= \alpha_1 \mu_1 + \alpha_2 \mu_2 && (223)\\ &= \beta_1(\alpha_1 \nu_{\mathrm{c}} + \alpha_2 \mu_2) + \beta_2(\alpha_1 \nu_{\mathrm{s}} + \alpha_2 \mu_2) && (224)\\ &= \beta_1 \mu^* + \beta_2 \nu^* && (225) \end{aligned}$$

where

$$\begin{aligned} \mu^* &= \alpha_1 \nu_{\mathrm{c}} + \alpha_2 \mu_2, && (226)\\ \nu^* &= \alpha_1 \nu_{\mathrm{s}} + \alpha_2 \mu_2. && (227) \end{aligned}$$

Then

$$\begin{aligned} \overline{\mathscr{D}}(\mu) &= \beta_1 \overline{\mathscr{D}}(\mu^*) + \beta_2 \overline{\mathscr{D}}(\nu^*) && (228)\\ &= \beta_1[\alpha_1 \overline{\mathscr{D}}(\nu_{\mathrm{c}}) + \alpha_2 \overline{\mathscr{D}}(\mu_2)] + \beta_2[\alpha_1 \overline{\mathscr{D}}(\nu_{\mathrm{s}}) + \alpha_2 \overline{\mathscr{D}}(\mu_2)] && (229)\\ &= \alpha_1[\beta_1 \overline{\mathscr{D}}(\nu_{\mathrm{c}}) + \beta_2 \overline{\mathscr{D}}(\nu_{\mathrm{s}})] + \alpha_2 \overline{\mathscr{D}}(\mu_2) && (230)\\ &= \alpha_1 \overline{\mathscr{D}}(\beta_1 \nu_{\mathrm{c}} + \beta_2 \nu_{\mathrm{s}}) + \alpha_2 \overline{\mathscr{D}}(\mu_2) && (231)\\ &= \alpha_1 \overline{\mathscr{D}}(\mu_1) + \alpha_2 \overline{\mathscr{D}}(\mu_2), && (232) \end{aligned}$$

where

- (238): applying the results in Appendix C-D to (225), since by assumption $\alpha_2 > 0$, we have $\mu^* \ll \nu^*$.
- (229), (231): applying the results in Appendix C-C to (226), (227) and (222), since $\nu_{\mathrm{c}} \ll \mu_2$, $\nu_{\mathrm{s}} \perp \mu_2$ and $\nu_{\mathrm{s}} \perp \nu_{\mathrm{c}}$.

Similarly, $\underline{\mathscr{D}}(\mu) = \alpha_1 \underline{\mathscr{D}}(\mu_1) + \alpha_2 \underline{\mathscr{D}}(\mu_2)$. This completes the proof of (157) and (158) for $n = 2$.

Next we proceed by induction on n: Suppose that (157) holds for $n = N$. For $n = N + 1$, assume that $\alpha_{N+1} < 1$, then

$$\begin{aligned} &\overline{\mathscr{D}}\left(\sum_{i=1}^{N+1} \alpha_i \mu_i\right)\\ &= \overline{\mathscr{D}}\left((1-\alpha_{N+1}) \sum_{i=1}^{N} \frac{\alpha_i}{(1-\alpha_{N+1})} \mu_i + \alpha_{N+1} \mu_{N+1}\right) && (233)\\ &= (1-\alpha_{N+1}) \overline{\mathscr{D}}\left(\sum_{i=1}^{N} \frac{\alpha_i}{(1-\alpha_{N+1})} \mu_i\right) + \alpha_{N+1} \overline{\mathscr{D}}(\mu_{N+1}) && (234)\\ &= \sum_{i=1}^{N+1} \alpha_i \overline{\mathscr{D}}(\mu_i), && (235) \end{aligned}$$

where (234) and (235) follow from the induction hypothesis. Therefore, (157) and (158) hold for any $n \in \mathbb{N}$.

F. Countable Mixture

Now we consider $n = \infty$: without loss of generality, assume that $\sum_{i=1}^{N} \alpha_i < 1$ for all $N \in \mathbb{N}$. Then

$$\begin{aligned} \overline{\mathscr{D}}(\mu) &= \sum_{i=1}^{N} \alpha_i \overline{\mathscr{D}}(\mu_i) + \left(1 - \sum_{i=1}^{N} \alpha_i\right) \overline{\mathscr{D}}(\nu_N) && (236)\\ &\le \sum_{i=1}^{N} \alpha_i \overline{\mathscr{D}}(\mu_i) + \left(1 - \sum_{i=1}^{N} \alpha_i\right), && (237) \end{aligned}$$

where

- (236): we have denoted $\nu_N = \frac{\sum_{i=N+1}^{\infty} \alpha_i \mu_i}{1 - \sum_{i=1}^{N} \alpha_i}$.
- (237): by Theorem 1.

Sending $N \to \infty$ yields $\overline{\mathscr{D}}(\mu) \le \sum_{i=1}^{\infty} \alpha_i \overline{\mathscr{D}}(\mu_i)$, and in entirely analogous fashion, $\underline{\mathscr{D}}(\mu) \ge \sum_{i=1}^{\infty} \alpha_i \underline{\mathscr{D}}(\mu_i)$. This completes the proof of Theorem 9.

Remark 4: Theorem 9 also generalizes to non-Gaussian noise. From the above proof, we see that (183) holds for all noise densities that satisfy Doob's relative limit theorems, in particular, those meeting the conditions in [42, Th. 4.1], e.g., uniform (by (164)) and exponential and Cauchy density ([42, Th. 5.1]).

More generally, notice that Lemma 4 deals with convolutional kernels which correspond to additive-noise channels. In [42, Th. 3.1], Doob also gave a result for general kernels. Therefore, it is possible to extend the results in Theorem 9 to general channels.

Appendix D
Proof of Theorem 10

Proof: Let $p_i = \mathbb{P}\{X = x_i\}$, $\epsilon = \frac{1}{\sqrt{\mathsf{snr}}}$ and $Y_\epsilon = X + \epsilon N$. In view of (49), it is equivalent to show that

$$\mathsf{mmse}(N \,|\, Y_\epsilon) = o(1). \tag{238}$$

Fix $\delta > 0$. Since $N \in L^2$, there exists $K > 0$, such that

$$\mathbb{E}\left[N^2 \mathbf{1}_{\{|N| > K\}}\right] < \delta/2. \tag{239}$$

Since $f_N(N) < \infty$ a.s., we can choose $J > 0$ such that

$$\mathbb{P}\{f_N(N) > J\} < \frac{\delta}{2K^2}. \tag{240}$$

Define

$$E_\delta = \{z : f_N(z) \le J, |z| \le K\}. \tag{241}$$

and $N_\delta = N \mathbf{1}_{E_\delta}(N)$. Then we have

$$\begin{aligned} \mathbb{E}[(N - N_\delta)^2] &\le K^2 \mathbb{P}\{f_N(N) > J\} + \mathbb{E}\left[N^2 \mathbf{1}_{\{|N| > K\}}\right] && (242)\\ &\le \delta, && (243) \end{aligned}$$

where (243) follows from (239) and (240).

The optimal estimator for N_δ based on Y_ϵ is given by

$$\hat{N}_\delta(y) = \frac{\sum_j p_j \frac{y-x_j}{\epsilon} \mathbf{1}_{E_\delta}\left(\frac{y-x_j}{\epsilon}\right) f_N\left(\frac{y-x_j}{\epsilon}\right)}{\sum_j p_j f_N\left(\frac{y-x_j}{\epsilon}\right)} \tag{244}$$

Then the MMSE of estimating N_δ based on Y_ϵ is

$$\begin{aligned}
&\mathsf{mmse}(N_\delta \,|\, Y_\epsilon) \\
&= \sum_i p_i \mathbb{E}\left[(N_\delta - \hat{N}_\delta(x_i+N))^2\right] && (245)\\
&= \sum_i p_i \int_{E_\delta} (z - \hat{N}_\delta(x_i+z))^2 f_N(z)\mathrm{d}z \\
&\quad + \sum_i p_i \int_{E_\delta^{\mathrm{c}}} \hat{N}_\delta(x_i+z)^2 f_N(z)\mathrm{d}z && (246)\\
&\le \sum_i p_i \int_{E_\delta} g_{z,i}(\epsilon) f_N(z)\mathrm{d}z + K^2 \mathbb{P}\{N \notin E_\delta\} && (247)\\
&\le J \sum_i p_i \int_{-K}^{K} g_{z,i}(\epsilon)\mathrm{d}z + \delta && (248)
\end{aligned}$$

where

- (247): by $|\hat{N}_\delta(y)| \le K$, since $N_\delta \le K$ a.s. We have also defined

$$g_{z,i}(\epsilon) = (z - \hat{N}_\delta(x_i+z))^2. \tag{249}$$

- (248): by (241) and

$$\begin{aligned}
&K^2\mathbb{P}\{N \notin E_\delta\} \\
&\le K^2\mathbb{P}\{f_N(N) > J\} + K^2\mathbb{P}\{|N| > K\} && (250)\\
&\le K^2\mathbb{P}\{f_N(N) > J\} + \mathbb{E}\left[N^2 \mathbf{1}_{\{|N|>K\}}\right] && (251)\\
&\le \delta, && (252)
\end{aligned}$$

where (252) follows from (239) and (240).

Next we show that for all i and $z \in E_\delta$, as $\epsilon \to 0$, we have

$$g_{z,i}(\epsilon) = o(1). \tag{253}$$

Indeed, using (244),

$$\begin{aligned}
&g_{z,i}(\epsilon) \\
&= \left[\frac{\sum_j p_j \frac{x_i-x_j}{\epsilon} \mathbf{1}_{E_\delta}\left(\frac{x_i-x_j}{\epsilon}+z\right) f_N\left(\frac{x_i-x_j}{\epsilon}+z\right)}{\sum_j p_j f_N\left(\frac{x_i-x_j}{\epsilon}+z\right)}\right]^2 && (254)\\
&\le \frac{\left[\sum_{j\neq i} p_j \frac{x_i-x_j}{\epsilon} \mathbf{1}_{E_\delta}\left(\frac{x_i-x_j}{\epsilon}+z\right) f_N\left(\frac{x_i-x_j}{\epsilon}+z\right)\right]^2}{p_i^2 f_N(z)^2} && (255)\\
&\le \frac{J^2(K+|z|)^2}{p_i^2 f_N(z)^2}\left[\sum_{j\neq i} p_j \mathbf{1}_{E_\delta}\left(\frac{x_i-x_j}{\epsilon}+z\right)\right]^2 && (256)\\
&\le \left[\frac{J(K+|z|)}{p_i f_N(z)}\mathbb{P}\left\{X \neq x_i, \left|\frac{x_i - X}{\epsilon} + z\right| \le K\right\}\right]^2 && (257)\\
&= o(1), && (258)
\end{aligned}$$

where

- (256): by (241).
- (258): by the boundedness of E_δ and the dominated convergence theorem.

By definition in (249), we have

$$g_{z,i}(\epsilon) \le (z+K)^2. \tag{259}$$

In view of (248) and dominated convergence theorem, we have

$$\limsup_{\epsilon\to 0} \mathsf{mmse}(N_\delta \,|\, Y_\epsilon) \le \delta. \tag{260}$$

Then

$$\begin{aligned}
&\limsup_{\epsilon\to 0} \sqrt{\mathsf{mmse}(N|Y_\epsilon)} \\
&\le \limsup_{\epsilon\to 0} \left\| N - \hat{N}_\delta(Y_\epsilon)\right\|_2 && (261)\\
&\le \limsup_{\epsilon\to 0} \sqrt{\mathsf{mmse}(N_\delta|Y_\epsilon)} + \|N - N_\delta\|_2 && (262)\\
&\le 2\sqrt{\delta} && (263)
\end{aligned}$$

where

- (261): by the suboptimality of $\hat{N}_\delta$.
- (261): by the triangle inequality.
- (263): by (243) and (260).

By the arbitrariness of δ, the proof of (238) is completed. ■

APPENDIX E
PROOF OF THEOREMS 11–13

We first compute the optimal MMSE estimator under absolutely continuous noise N. Let $\epsilon = \frac{1}{\sqrt{\mathsf{snr}}}$. The density of $Y_\epsilon = X + \epsilon N$ is

$$q_0(y) = \frac{1}{\epsilon}\mathbb{E}\left[f_N\left(\frac{y-X}{\epsilon}\right)\right]. \tag{264}$$

Denote

$$q_1(y) = \frac{1}{\epsilon}\mathbb{E}\left[X f_N\left(\frac{y-X}{\epsilon}\right)\right]. \tag{265}$$

Then the optimal MSE estimator of X given Y_ϵ is given by

$$\hat{X}(y) = \mathbb{E}[X \,|\, Y_\epsilon = y] = \frac{q_1(y)}{q_0(y)} = \frac{\mathbb{E}\left[X f_N\left(\frac{y-X}{\epsilon}\right)\right]}{\mathbb{E}\left[f_N\left(\frac{y-X}{\epsilon}\right)\right]}. \tag{266}$$

A. Proof of Theorem 12

Proof: By (48), we have

$$\mathsf{snr}\cdot\mathsf{mmse}(X,N,\mathsf{snr}) = \mathsf{mmse}(N \,|\, Y_\epsilon). \tag{267}$$

Due to Theorem 1, we only need to show

$$\underline{\mathscr{D}}(X,N) \ge 1, \tag{268}$$

which, in view of (267), is equivalent to

$$\liminf_{\epsilon\downarrow 0} \mathsf{mmse}(N \,|\, Y_\epsilon) \ge \mathsf{var}N. \tag{269}$$

The optimal estimator for N given Y_ϵ is given by

$$\mathbb{E}[N \,|\, Y_\epsilon = y] = \frac{\mathbb{E}[N f_X(y-\epsilon N)]}{\mathbb{E}[f_X(y-\epsilon N)]}. \tag{270}$$

Fix an arbitrary positive δ. Since $N \in L^2(\Omega)$, there exists $M > 0$, such that

$$\mathbb{E}\left[N^2 \mathbf{1}_{\{|N|>M\}}\right] < \delta^2. \tag{271}$$

Then

$$\begin{aligned}
&\sqrt{\mathsf{mmse}(N \,|\, Y_\epsilon)} \\
&= \sqrt{\mathbb{E}[(N - \mathbb{E}[N \,|\, Y_\epsilon])^2]} && (272)\\
&\geq \sqrt{\mathbb{E}[(N - \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon])^2]} - \sqrt{\mathbb{E}[(\mathbb{E}[N\mathbf{1}_{\{|N|>M\}}|Y_\epsilon])^2]} && (273)\\
&\geq \sqrt{\mathbb{E}[(N - \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon])^2]} - \sqrt{\mathbb{E}[N^2\mathbf{1}_{\{|N|>M\}}]} && (274)\\
&\geq \sqrt{E_{M,\epsilon}} - \delta && (275)
\end{aligned}$$

where

- (273): by writing $\mathbb{E}[N \,|\, Y_\epsilon] = \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon] + \mathbb{E}[N\mathbf{1}_{\{|N|>M\}}|Y_\epsilon]$ and the triangle inequality.
- (274): by $\mathbb{E}[(\mathbb{E}[U|V])^2] \leq \mathbb{E}[U^2]$ for all $U, V \in L^2(\Omega)$.
- (275): by (271) and

$$E_{M,\epsilon} \triangleq \mathbb{E}[(N - \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon])^2]. \tag{276}$$

Define

$$\begin{aligned}
p_M(y;\epsilon) &= \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}} f_X(y - \epsilon N)] && (277)\\
q(y;\epsilon) &= \mathbb{E}[f_X(y - \epsilon N)] && (278)\\
\hat{N}_M(y;\epsilon) &= \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon = y] = \frac{p_M(y;\epsilon)}{q(y;\epsilon)}. && (279)
\end{aligned}$$

Suppose $f \in C^b$. Then by the bounded convergence theorem,

$$\begin{aligned}
\lim_{\epsilon \downarrow 0} p_M(x + \epsilon z;\epsilon) &= \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}] f_X(x), && (280)\\
\lim_{\epsilon \downarrow 0} q(x + \epsilon z;\epsilon) &= f_X(x) && (281)
\end{aligned}$$

hold for all $x, z \in \mathbb{R}$. Since $f_X(X) > 0$ a.s.,

$$\begin{aligned}
\lim_{\epsilon \downarrow 0} \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}|Y_\epsilon] &= \lim_{\epsilon \downarrow 0} \hat{N}_M(X + \epsilon N;\epsilon) && (282)\\
&= \mathbb{E}\left[N\mathbf{1}_{\{|N|\leq M\}}\right] && (283)
\end{aligned}$$

holds a.s. Then by Fatou's lemma

$$\liminf_{\epsilon \downarrow 0} E_{M,\epsilon} \geq \mathbb{E}\left[N - \mathbb{E}[N\mathbf{1}_{\{|N|\leq M\}}]\right]^2 \geq \mathsf{var}N. \tag{284}$$

By (275),

$$\begin{aligned}
\liminf_{\epsilon \downarrow 0} \sqrt{\mathsf{mmse}(N|Y_\epsilon)} &\geq \liminf_{\epsilon \downarrow 0} \sqrt{E_{M,\epsilon}} - \delta && (285)\\
&\geq \sqrt{\mathsf{var}N} - \delta. && (286)
\end{aligned}$$

By the arbitrariness of δ, we conclude that

$$\liminf_{\epsilon \downarrow 0} \mathsf{mmse}(N|Y_\epsilon) \geq \mathsf{var}N, \tag{287}$$

hence (268) holds. ∎

B. Proof of Theorem 11

Now we are dealing with X whose density is not necessarily continuous or bounded. In order to show that (280) and (281) continue to hold under the assumptions of the noise density in Theorem 11, we need the following lemma from [43, Sec. 3.2]:

Lemma 8 ([43, Th. 3.2.1]): Suppose the family of functions $\{K_\epsilon : \mathbb{R}^d \to \mathbb{R}\}_{\epsilon>0}$ satisfies the following conditions: for some constant $\eta > 0$ and $C \in \mathbb{R}$,

$$\begin{aligned}
\int_{\mathbb{R}^d} K_\epsilon(x)\mathrm{d}x &= C && (288)\\
\sup_{x \in \mathbb{R}^d, \epsilon > 0} \epsilon |K_\epsilon(x)| &< \infty && (289)\\
\sup_{x \in \mathbb{R}^d, \epsilon > 0} \frac{|x|^{1+\eta}}{\epsilon^\eta} |K_\epsilon(x)| &< \infty && (290)
\end{aligned}$$

hold for all $\epsilon > 0$ and $x \in \mathbb{R}^d$. Then for all $f \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$,

$$\lim_{\epsilon \downarrow 0} f * K_\epsilon(x) = C f(x) \tag{291}$$

holds for Lebesgue-a.e. x.

Note that in the original version of Lemma 8 in [43, Sec. 3.2] $C = 1$, and K_ϵ is dubbed *approximation of the identity*. For $C \neq 0$ or 1, the same conclusion follows from scaling. The case of $C = 0$ can be shown as follows: take some kernel G_ϵ which is an approximation to the identity. Then $G + K$ is also an approximation to the identity. Then the conclusion for K follows by applying Lemma 8 to both G and $G + K$ and then subtracting the corresponding (291) limits.

Proof of Theorem 11: Based on the proof of Theorem 12, it is sufficient to show that (280) and (281) hold for Lebesgue-a.e. x and z. Fix z. First look at (281): introduce the following kernel which corresponds to the density of $\epsilon(N - z)$

$$K_\epsilon(x) = \frac{1}{\epsilon} f_N\left(\frac{x}{\epsilon} + z\right). \tag{292}$$

We check that K_ϵ is an approximation to the identity by verifying:

- (288): $\int_{\mathbb{R}} K_\epsilon(x)\mathrm{d}x = \int_{\mathbb{R}} f_N(u)\mathrm{d}u = 1$.
- (289): $\sup_{x \in \mathbb{R}, \epsilon > 0} \epsilon |K_\epsilon(x)| = \sup_{u \in \mathbb{R}} f_N(u) < \infty$, since f_N is bounded.
- (290): by boundedness of f_N and (93), we have: for some $\eta > 0$,

$$\sup_{u \in \mathbb{R}} |u|^{1+\eta} f_N(u) < \infty \tag{293}$$

then

$$\begin{aligned}
&\left(\sup_{x \in \mathbb{R}, \epsilon > 0} \frac{|x|^{1+\eta}}{\epsilon^\eta} |K_\epsilon(x)|\right)^{\frac{1}{1+\eta}} \\
&= \sup_{u \in \mathbb{R}} |u| f_N^{\frac{1}{1+\eta}}(u + z) && (294)\\
&\leq \sup_{u \in \mathbb{R}} |u| f_N^{\frac{1}{1+\eta}}(u) + |z| \sup_{u \in \mathbb{R}} f_N^{\frac{1}{1+\eta}}(u) && (295)\\
&< \infty. && (296)
\end{aligned}$$

Note that

$$q(x+\epsilon z;\epsilon) = f_X * K_\epsilon(x). \tag{297}$$

Since $f_X \in L^1(\mathbb{R})$, by Lemma 8, (281) holds for Lebesgue-a.e. x. Similarly, we define

$$G_\epsilon(x) = \left(\frac{x}{\epsilon}+z\right)\mathbf{1}_{\{\left|\frac{x}{\epsilon}+z\right|\le M\}}\frac{1}{\epsilon}f_N\left(\frac{x}{\epsilon}+z\right). \tag{298}$$

Then it is verifiable that G_ϵ satisfies (288)–(290), with

$$\int G_\epsilon(x)\mathrm{d}x = \mathbb{E}\left[N\mathbf{1}_{\{|N|\le M\}}\right]. \tag{299}$$

Since

$$p_M(x+\epsilon z;\epsilon) = f_X * G_\epsilon(x), \tag{300}$$

Lemma 8 implies that (280) holds for Lebesgue-a.e. x. ∎

C. Proof of Theorem 13

Proof: Without loss of generality, assume $\mathbb{E}N = 0$. Following the notation in the proof of Theorem 12, similar to $q(y;\epsilon)$ in (278), we define

$$p(y;\epsilon) = \mathbb{E}[Nf_X(y-\epsilon N)]. \tag{301}$$

Then

$$\mathbb{E}[N|Y_\epsilon = y] = \frac{p(y;\epsilon)}{q(y;\epsilon)}. \tag{302}$$

Fix $x, z \in \mathbb{R}$. Since f_X, f_X' and f_X'' are all bounded, we can interchange derivatives with integrals [44, Th. 2.27] and write

$$\begin{aligned} &q(x+\epsilon z;\epsilon)\\ &= \mathbb{E}[f_X(x+\epsilon(z-N))] &(303)\\ &= f_X(x)+\epsilon f_X'(x)\mathbb{E}(z-N)+\frac{\epsilon^2}{2}f_X''(x)\mathbb{E}(z-N)^2+o(\epsilon^2) &(304)\\ &= f_X(x)+\epsilon f_X'(x)z+\frac{\epsilon^2}{2}f_X''(x)(z^2+\mathsf{var}N)+o(\epsilon^2). &(305)\end{aligned}$$

Similarly, since $\mathbb{E}[|N|^3] < \infty$, we have

$$\begin{aligned} &p(x+\epsilon z;\epsilon)\\ &= \mathbb{E}[Nf_X(x+\epsilon(z-N))] &(306)\\ &= -\epsilon f_X'(x)\,\mathsf{var}N+\frac{\epsilon^2}{2}f_X''(x)(\mathbb{E}[N^3]-2z\,\mathsf{var}N)+o(\epsilon^2). &(307)\end{aligned}$$

Define the score function $\rho(x) = \frac{f_X'}{f_X}(x)$. Since $f_X(X) > 0$ a.s., by (302),

$$\mathbb{E}[N|Y_\epsilon] = \frac{\epsilon^2}{2}\left[\frac{f_X''(X)}{f_X(X)}(\mathbb{E}[N^3]-2N\mathsf{var}N)+2N\rho^2(X)\mathsf{var}N\right] - \epsilon\rho(X)\mathsf{var}N+o(\epsilon^2). \tag{308}$$

holds a.s. Then

$$\begin{aligned} &\mathsf{mmse}(N|Y_\epsilon)\\ &= \mathbb{E}[N-\mathbb{E}[N|Y_\epsilon]]^2 &(309)\\ &= \mathsf{var}N-\epsilon\mathbb{E}[N]\mathbb{E}[\rho(X)]\mathsf{var}N+\epsilon^2(\mathsf{var}N)^2\mathbb{E}[\rho^2(X)]-\\ &\quad 2\epsilon^2(\mathsf{var}N)^2J(X)+\epsilon^2(\mathbb{E}[N]\mathbb{E}[N^3]-2\mathsf{var}^2N)\mathbb{E}\left[\frac{f_X''(X)}{f_X(X)}\right]\\ &\quad +o(\epsilon^2) &(310)\\ &= \mathsf{var}N-\epsilon^2(\mathsf{var}N)^2J(X)+o(\epsilon^2), &(311)\end{aligned}$$

where

- (310): by the bounded convergence theorem, because the ratio between the $o(\epsilon^2)$ term in (308) and ϵ^2 is upper bounded by $\rho^2(X)$ and $\frac{f_X''(X)}{f_X(X)}$, which are integrable by assumption.
- (311): by $\mathbb{E}[\rho(X)] = 0$, $\mathbb{E}[\rho^2(X)] = J(X)$ and $\mathbb{E}[\frac{f_X''(X)}{f_X(X)}] = 0$, in view of (108).

∎

Appendix F
Proof of Theorem 16

By Remark 2, Theorem 16 holds trivially if $\mathbb{P}\{X=0\}=1$ or $\mathbb{P}\{X=1\}=1$. Otherwise, since $X=0$ if and only if $(X)_j = 0$ for all $j \in \mathbb{N}$, but $\{(X)_j\}$ are i.i.d., therefore $\mathbb{P}\{X=0\}=0$. Similarly, $\mathbb{P}\{X=1\}=0$. Hence $0 < X < 1$ a.s.

To prove (129), we define the function $G : \mathbb{R} \to \mathbb{R}_+$

$$\begin{aligned} G(b) &= M^{2b}\cdot\mathsf{mmse}(X,N,M^{2b}) &(312)\\ &= \mathsf{mmse}(N,X,M^{-2b}), &(313)\end{aligned}$$

where (313) follows from (48). The oscillatory behavior of G is given in the following lemma:

Lemma 9:

1) For any $b \in \mathbb{R}$, $\{G(k+b) : k \in \mathbb{N}\}$ is an nondecreasing nonnegative sequence bounded from above by $\mathsf{var}N$.
2) Define a function $\Psi : \mathbb{R} \to \mathbb{R}_+$ by

$$\Psi(b) = \lim_{k\to\infty} G(k+b). \tag{314}$$

Then Ψ is a 1-periodic function, and the convergence in (314) is *uniform* in $b \in [0,1]$. Therefore as $b \to \infty$,

$$G(b) = \Psi(b)+o(1) \tag{315}$$

In view of Lemma 9, we define a $2\log M$-periodic function $\Phi_{X,N} : \mathbb{R} \to [0,1]$ as follows:

$$\Phi_{X,N}(t) = \frac{1}{\mathsf{var}N}\Psi\left(\frac{t}{2\log M}\right). \tag{316}$$

Having defined $\Phi_{X,N}$, (129), (130), and (131) readily follow from (315).

Next we prove (132) in the case of Gaussian N: in view of (316), it is equivalent to show

$$\int_0^1 \Psi(b)\mathrm{d}b = d(X). \tag{317}$$

Denote

$$\mathsf{snr} = M^{2b}. \tag{318}$$

Recalling $G(b)$ defined in (313), we have

$$\int_0^b G(\tau)\mathrm{d}\tau = \int_1^{\mathsf{snr}} \gamma\, \mathsf{mmse}(X,\gamma)\frac{1}{2\gamma\log M}\mathrm{d}\gamma \tag{319}$$

$$= \frac{I(\mathsf{snr}) - I(1)}{2\log M}, \tag{320}$$

where (320) follows from (73) and $I(\mathsf{snr})$ is defined in (74).

Since $d(X)$ exists, in view of (63) and (64), we have

$$\lim_{b\to\infty}\frac{1}{b}\int_0^b G(\tau)\mathrm{d}\tau = \lim_{\mathsf{snr}\to\infty}\frac{2I(\mathsf{snr})}{\log\mathsf{snr}} = d(X). \tag{321}$$

Since the convergence in (314) is uniform in $b\in[0,1]$, for all $\epsilon>0$, there exists k_0 such that for all $k\geq k_0$ and $b\in[0,1]$,

$$|G(k+b) - \Phi_{X,N_{\mathrm{G}}}(b)| \leq \epsilon. \tag{322}$$

Then for all integers $k\geq k_0$, we have

$$\left|\frac{1}{k}\int_0^k G(\tau)\mathrm{d}\tau - \int_0^1 \Psi(\tau)\mathrm{d}\tau\right|$$

$$\leq \frac{2k_0}{k} + \frac{1}{k}\sum_{j=k_0}^{k-1}\int_0^1 |G(j+\tau)-\Psi(\tau)|\mathrm{d}\tau \tag{323}$$

$$\leq \frac{2k_0}{k} + \frac{k-k_0}{k}\epsilon \tag{324}$$

where

- (323): G and Ψ map into $[0,1]$.
- (324): by (322).

By the arbitrariness of ϵ and (321), we have

$$\int_0^1 \Psi(\tau)\mathrm{d}\tau = \lim_{k\to\infty}\frac{1}{k}\int_0^k G(\tau)\mathrm{d}\tau = d(X). \tag{325}$$

To finish the proof, we prove Lemma 9. Note that

$$G(k+1+b) = \mathsf{mmse}(N,X,M^{-2(k+1+b)}) \tag{326}$$

$$= \mathsf{mmse}(N|N+\sqrt{\mathsf{snr}}M^{k+1}X) \tag{327}$$

$$= \mathsf{mmse}(N|N+\sqrt{\mathsf{snr}}M^{k}(U+V)) \tag{328}$$

$$\geq \mathsf{mmse}(N|N+\sqrt{\mathsf{snr}}M^{k}V) \tag{329}$$

$$= \mathsf{mmse}(N|N+\sqrt{\mathsf{snr}}M^{k}X) \tag{330}$$

$$= G(k+b) \tag{331}$$

where

- (326): by (313);
- (327): by (47) and (318);
- (328): by the M-ary expansion of X in (58) and we have defined

$$U = (X)_1, \tag{332}$$

$$V = \sum_{j=1}^{\infty}(X)_{j+1}M^{-j}; \tag{333}$$

- (329): by the data-processing inequality of MMSE [36], since U is independent of V;
- (330): by $V\overset{\mathrm{D}}{=}X$ since $\{(X)_j\}$ is an i.i.d. sequence.

Therefore, for fixed b, $G(k+b)$ is an nondecreasing sequence in k. By (326),

$$G(k+b)\leq \mathsf{var}N, \tag{334}$$

hence $\lim_{k\to\infty}G(k+b)$ exists, denoted by $\Psi(b)$. The 1-periodicity of Ψ readily follows.

To prove (315), we show that there exist two functions $c_1,c_2:\mathbb{R}\to\mathbb{R}_+$ depending on the distribution of X and N only, such that

$$\lim_{b\to\infty} c_i(b) = 0, \quad i=1,2 \tag{335}$$

and

$$0\leq \Psi(b) - G(b) \tag{336}$$

$$\leq c_1^2(b) + c_2(b) + 2c_1(b)\sqrt{\mathsf{var}N + c_2(b)}. \tag{337}$$

Then (315) follows by combining (335)–(337) and sending $b\to\infty$. Inequalities (336) and (337) also show that the convergence in (314) is uniform in $b\in[0,1]$.

To conclude this proof, we proceed to construct the desired functions c_1 and c_2 and prove (336) and (337): by monotonicity, for all $b\in\mathbb{R}$,

$$G(b)\leq\Psi(b). \tag{338}$$

Hence (336) follows. To prove (337), fix $k\in\mathbb{N}$ and $K>0$ to be specified later. Define

$$N_K = N\mathbf{1}_{\{|N|\leq K\}} \tag{339}$$

$$\bar{N}_K = N - N_K \tag{340}$$

We use a suboptimal estimator to bound $G(k+b)-G(b)$. To streamline the proof, introduce the following notation:

$$W = M^k[X]_k \tag{341}$$

$$Z = M^k(X-[X]_k) \tag{342}$$

$$Y = \sqrt{\mathsf{snr}}X + N \tag{343}$$

$$Y' = \sqrt{\mathsf{snr}}M^kX + N \tag{344}$$

$$= \sqrt{\mathsf{snr}}W + \sqrt{\mathsf{snr}}Z + N, \tag{345}$$

where W is integer-valued. Note that since X has i.i.d. M-ary expansion, $\{N,W,Z\}$ are independent, and

$$Z\overset{\mathrm{D}}{=}X, \tag{346}$$

hence

$$Y'\overset{\mathrm{D}}{=}Y + \sqrt{\mathsf{snr}}W. \tag{347}$$

Based on Y', we use the following two-stage suboptimal estimator $\tilde{N}_K$ for N_K: first estimate W (the first k bits of X) based on Y' according to

$$\tilde{w}(y) = \left[\frac{y}{\sqrt{\mathsf{snr}}}\right]. \tag{348}$$

Then peel off $\tilde{W}=\tilde{w}(Y')$ from Y' and plug it into the optimal estimator for X based on Y

$$\hat{N}_K(y)=\mathbb{E}[N_K\,|\,Y=y] \tag{349}$$

to estimate X, i.e.,

$$\tilde{N}_K(y)=\hat{N}_K(y-\sqrt{\mathsf{snr}}\tilde{w}(y)). \tag{350}$$

Next we bound the probability of choosing the wrong W

$$\begin{aligned}
&\mathbb{P}\{\tilde{W}\neq W\}\\
&=1-\mathbb{P}\left\{\left[\frac{N}{\sqrt{\mathsf{snr}}}+W+Z\right]=W\right\} &(351)\\
&=1-\mathbb{P}\left\{W\le\frac{N}{\sqrt{\mathsf{snr}}}+W+Z<W+1\right\} &(352)\\
&=1-\mathbb{P}\{-\sqrt{\mathsf{snr}}X\le N<\sqrt{\mathsf{snr}}(1-X)\} &(353)\\
&=\int_{(0,1)}P_X(\mathrm{d}x)[1-\mathbb{P}\{-\sqrt{\mathsf{snr}}x\le N<\sqrt{\mathsf{snr}}(1-x)\}] &(354)\\
&\triangleq\kappa(\mathsf{snr})\to 0, &(355)
\end{aligned}$$

where

- (351): by (345) and (348);
- (352): by the fact that $[x]=n$ if and only if $n\le x<n+1$;
- (353): by (346);
- (354): by the assumption that $X\in(0,1)$ a.s.;
- (355): by the bounded convergence theorem.

Note that $\kappa(\mathsf{snr})$ is a nonincreasing nonnegative function depending only on the distribution of X and N.

Finally we choose K and analyze the performance of $\tilde{N}_K$. Observe from (351) that the probability of choosing the wrong W does *not* depend k. This allows us to choose K independent of k

$$K=[\kappa(\mathsf{snr})]^{-\frac{1}{3}}. \tag{356}$$

Therefore,

$$\begin{aligned}
\sqrt{G(k+b)}&=\sqrt{\mathsf{mmse}(N|Y')} &(357)\\
&\le\|N-N_K+N_K-\tilde{N}_K(Y')\|_2 &(358)\\
&\le\|\bar{N}_K\|_2+\|N_K-\hat{N}_K(Y'-\sqrt{\mathsf{snr}}\tilde{W})\|_2 &(359)
\end{aligned}$$

where

- (357): by (313) and (345);
- (358): by the suboptimality of $\tilde{N}_K$;
- (359): by (350).

Now

$$\begin{aligned}
&\mathbb{E}[(N_K-\hat{N}_K(Y'-\sqrt{\mathsf{snr}}\tilde{W})^2]\\
&\le\mathbb{E}[N_K-\hat{N}_K(Y)]^2+\\
&\mathbb{E}\left[(N_K-\hat{N}_K(Y'-\sqrt{\mathsf{snr}}\tilde{W}))^2\mathbf{1}_{\{\tilde{W}\neq W\}}\right] &(360)\\
&\le\mathsf{mmse}(N_K|Y)+4K^2\mathbb{P}\{\tilde{W}\neq W\} &(361)\\
&\le G(b)+3\|\bar{N}_K\|_2\|N\|_2+4\kappa(\mathsf{snr})^{\frac{1}{3}} &(362)
\end{aligned}$$

where

- (360): by (347);
- (361): by $|\hat{N}_K(y)|=|\mathbb{E}[\hat{N}_K|Y=y]|\le K$ for all y, since $|N_K|\le K$ a.s.;
- (362): by Lemma 7, (355), (356), and $\mathsf{mmse}(N\,|\,Y)=G(b)$.

Define

$$\begin{aligned}
c_1(b)&=\|\bar{N}_K\|_2 &(363)\\
c_2(b)&=3\|\bar{N}_K\|_2\|N\|_2+4\kappa(\mathsf{snr})^{\frac{1}{3}} &(364)
\end{aligned}$$

where b and K are related to snr through (318) and (356), respectively. Then substituting (362) into (359) yields

$$\sqrt{G(k+b)}\le c_1(b)+\sqrt{G(b)+c_2(b)}. \tag{365}$$

Note that the right hand side of (365) does not depend on k. By (314), sending $k\to\infty$ we obtain

$$\sqrt{\Psi(b)}\le c_1(b)+\sqrt{G(b)+c_2(b)}. \tag{366}$$

In view of (338), squaring[8] (366) on both sides and noticing that $G(b)\le\mathsf{var}N$ for all b, we have

$$\Psi(b)-G(b)\le c_1^2(b)+c_2(b)+2c_1(b)\sqrt{\mathsf{var}N+c_2(b)}. \tag{367}$$

By (355) and (356), $K\to\infty$ as $b\to\infty$. Since $\mathbb{E}[N^2]<\infty$, $c_1(b)$ and $c_2(b)$ both tend to zero as $b\to\infty$, and (337) follows.

Appendix G
Proof of Lemma 1

Proof: For any $\mathsf{snr}>0$, there exists $n\in\mathbb{Z}_+$, such that $n\le\mathsf{snr}<n+1$. Since the function $\mathsf{mmse}(\mathsf{snr})=\mathsf{mmse}(X,N,\mathsf{snr}\,|\,U)$ is monotonically decreasing, we have

$$\mathsf{mmse}(n+1)\le\mathsf{mmse}(\mathsf{snr})\le\mathsf{mmse}(n), \tag{368}$$

hence

$$\begin{aligned}
n\,\mathsf{mmse}(n+1)&\le\mathsf{snr}\,\mathsf{mmse}(\mathsf{snr}) &(369)\\
&\le n\,\mathsf{mmse}(n)+\mathsf{mmse}(n). &(370)
\end{aligned}$$

Since $\lim_{n\to\infty}\mathsf{mmse}(n)=0$, the claim of Lemma 1 follows.■

Appendix H
Proof for Remark 3

We show that for any singular X and any binary-valued N,

$$\underline{\mathscr{D}}(X,N)=0. \tag{371}$$

To this end, we need the following auxiliary results:

Lemma 10 (Mutually Singular Hypothesis): The optimal test for the binary hypothesis testing problem

$$\begin{cases}H_0: & P\\ H_1: & Q\end{cases} \tag{372}$$

[8]In this step it is essential that $G(b)$ be bounded, because in general $\sqrt{a_n}=\sqrt{b_n}+o(1)$ does *not* imply that $a_n=b_n+o(1)$. For instance, $a_n=n+1, b_n=n$.

with prior $\mathbb{P}\{H_0\} \in (0,1)$ has zero error probability if and only if $P \perp Q$.

Proof: By definition, $P \perp Q$ if and only if there exists an event A such that $P(A) = 1$ and $Q(A) = 0$. Then the test that decides H_0 if and only if A occurs yields zero error probability. ■

Lemma 11 ([45, Th. 10]): Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$ that is mutually singular with respect to Lebesgue measure. Then there exists a Borel set E with $\mu(E) = 1$ and a non-empty perfect set C such that $\{c + E : c \in C\}$ is a family of disjoint sets.

Proof of (371): In view of Theorem 2, we may assume that N is $\{0,1\}$-valued without loss of generality. For any input X whose distribution μ is mutually singular to the Lebesgue measure, we show that there exists a vanishing sequence $\{\epsilon_n\}$, such that for all n,

$$\mathsf{mmse}(X \,|\, X + \epsilon_n N) = 0. \tag{373}$$

which implies that $\underline{\mathscr{Z}}(X, N) = 0$.

By Lemma 11, there exists a Borel set E and a perfect set C, such that $\mu(E) = 1$ and $\{c + E : c \in C\}$ is a family of disjoint sets. Pick any $a \in C$. Since C is perfect, there exists $\{a_n\} \subset C$, such that $a_n \to a$ and $\epsilon_n \triangleq a_n - a > 0$. Since X and $X + \epsilon_n$ are supported on disjoint subsets E and $E + \epsilon_n$ respectively, their distributions are mutually singular. By Lemma 10, the optimal test for N based on $X + \epsilon_n N$ succeeds with probability one, which implies (373). ■

ACKNOWLEDGMENT

The paper has benefited from thorough suggestions by the anonymous reviewers.

REFERENCES

[1] Y. Wu and S. Verdú, "MMSE dimension," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*, Austin, TX, Jun. 2010, pp. 1463–1467.

[2] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Hungarica*, vol. 10, no. 1–2, Mar. 1959.

[3] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721–3748, Aug. 2010.

[4] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[5] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1564–1572, Sep. 1994.

[6] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 2nd ed. New York: Wiley, 2003.

[7] 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.

[8] A. Guionnet and D. Shlyakhtenko, "On classical analogues of free entropy dimension," *J. Funct. Anal.*, vol. 251, no. 2, pp. 738–771, Oct. 2007.

[9] J. K. Ghosh, *Higher Order Asymptotics*. Hayward, CA: Inst. Math. Stat., 1994.

[10] J. K. Ghosh, B. K. Sinha, and S. N. Joshi, "Expansions for posterior probability and integrated Bayes risk," in *Statistical Decision Theory and Related Topics III*. New York: Academic, 1982, pp. 403–456.

[11] M. V. Burnašev, "Investigation of second order properties of statistical estimators in a scheme of independent observations," *Math. USSR Izvestiya*, vol. 18, no. 3, pp. 439–467, 1982.

[12] 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.

[13] M. S. Pinsker, "Optimal filtering of square-integrable signals in Gaussian noise," *Problemy Peredachi Informatsii*, vol. 16, no. 2, pp. 52–68, 1980.

[14] M. Nussbaum, "Minimax risk: Pinsker bound," in *Encyclopedia of Statistical Sciences*. New York: Wiley, 1999, pp. 451–460.

[15] B. Bobrovsky and M. Zakai, "Asymptotic a priori estimates for the error in the nonlinear filtering problem," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 371–376, 1982.

[16] Y. Wu and S. Verdú, "Functional properties of MMSE," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*, Austin, TX, Jun. 2010, pp. 1453–1457.

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

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

[19] V. V. Prelov and E. C. van der Meulen, "Asymptotics of Fisher information under weak perturbation," *Probl. Inf. Transm.*, vol. 31, no. 1, pp. 14–22, 1995.

[20] M. S. Pinsker, V. V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1877–1888, Nov. 1995.

[21] V. V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.

[22] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Inf. Theory*, vol. 1, no. 1, June 2004.

[23] E. Çinlar, *Probability and Stochastics*. New York: Springer, 2011.

[24] K. Falconer, *Techniques in Fractal Geometry*. Chichester, U.K.: Wiley, 1997.

[25] Y. Wu and S. Verdú, "Functional properties of MMSE and mutual information," *IEEE Trans. Inf. Theory*, 2010, submitted for publication.

[26] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Control*, vol. 2, no. 2, pp. 101–112, 1959.

[27] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[28] L. D. Brown and L. Gajek, "Information inequalities for the Bayes risk," *Ann. Stat.*, vol. 18, no. 4, pp. 1578–1594, 1990.

[29] E. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1998.

[30] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[31] D. Guo, D. Baron, and S. Shamai (Shitz), "A single-letter characterization of optimal noisy compressed sensing," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, Oct. 2009.

[32] C. Weidmann and M. Vetterli, "Rate distortion behavior of sparse sources," *IEEE Trans. Inf. Theory*, Aug. 2010, submitted for publication.

[33] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3033–3051, Jul. 2006.

[34] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[35] K. Knopp and F. Bagemihl, *Theory of Functions, Parts I and II*. Mineola, NY: Dover, 1996.

[36] 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, Aug. 1998.

[37] Y. Wu and S. Verdú, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, Jun. 2011, submitted for publication.

[38] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3712–3720, Jul. 2010.

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

[40] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[41] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. Boca Raton, FL: CRC, 1992.

[42] J. Doob, "Relative limit theorems in analysis," *Journal d'Analyse Mathématique*, vol. 8, no. 1, pp. 289–306, 1960.

[43] E. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton, NJ: Princeton Univ. Press, 2005.

[44] G. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. New York: Wiley-Interscience, 1999.

[45] V. Prokaj, "A characterization of singular measures," *Real Analysis Exchange*, vol. 29, pp. 805–812, 2004.

Yihong Wu (S'10) received the B.E. and M.A. degrees from Tsinghua University, Beijing, China, in 2006 and Princeton University in 2008, respectively, both in electrical engineering. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, Princeton University, Princeton, NJ.

His research interests are in information theory, signal processing, mathematical statistics, approximation theory, optimization, and distributed algorithms.

Mr. Wu is a recipient of the Princeton University Wallace Memorial Honorific Fellowship in 2010.

Sergio Verdú (S'80-M'84-SM'88-F'93) is the Eugene Higgins Professor of Electrical Engineering at Princeton University.

A member of the National Academy of Engineering, Verdú is the recipient of the 2007 Claude Shannon Award and the 2008 IEEE Richard Hamming Medal. He was awarded a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya in 2005.

His research has received several awards including the 1998 Information Theory Outstanding Paper Award, the Information Theory Golden Jubilee Paper Award, and the 2006 Joint Communications/Information Theory Paper Award.

Sergio Verdú is currently Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.