Measuring Information from Moments

Wael Alghamdi, *Graduate Student Member, IEEE,* and Flavio P. Calmon, *Member, IEEE*

***Abstract*—We investigate the problem of representing information measures in terms of the moments of the underlying random variables. First, we derive polynomial approximations of the conditional expectation operator. We then apply these approximations to bound the best mean-square error achieved by a polynomial estimator—referred to here as the PMMSE. In Gaussian channels, the PMMSE coincides with the minimum mean-square error (MMSE) if and only if the input is either Gaussian or constant, i.e., if and only if the conditional expectation of the input of the channel given the output is a polynomial of degree at most 1. By combining the PMMSE with the I-MMSE relationship, we derive new formulas for information measures (e.g., differential entropy, mutual information) that are given in terms of the moments of the underlying random variables. As an application, we introduce estimators for information measures from data via approximating the moments in our formulas by sample moments. These estimators are shown to be asymptotically consistent and possess desirable properties, e.g., invariance to affine transformations when used to estimate mutual information.**

***Index Terms*—Polynomials, Hilbert space, AWGN channels, Estimation theory, Probability.**

I. Introduction

A FUNDAMENTAL formula in information theory is the I-MMSE relation [1], which shows that in Gaussian channels the mutual information is the integral of the minimum mean-square error (MMSE):

$$I(X;\sqrt{\gamma}X+N)=\frac{1}{2}\int_0^{\gamma}\operatorname{mmse}\left(X\mid\sqrt{t}X+N\right)\,dt. \quad (1)$$

Here, X has finite variance and N is a standard normal random variable independent of X. In this paper, we build on this relation to express information measures of two random variables X and Y as functions of their moments. For example, whenever X and Y are continuous with finite moment-generating functions around the origin, there is a sequence of rational functions $\{\rho_n\}_{n\in\mathbb{N}}$—each completely determined by finitely many moments of X and Y—such that the mutual information is

$$I(X;Y)=\lim_{n\to\infty}\int_{\mathbb{R}}\rho_n(t)\,dt. \quad (2)$$

We derive the new expression (2) and a similar formula for differential entropy in three steps. First, we produce polynomial approximations of conditional expectations. Second, we apply these approximations to bound the mean-square error of reconstructing a hidden variable X from an observation Y using an estimator that is a polynomial in Y. We call this approximation the PMMSE, in short for Polynomial MMSE. Finally, we use the PMMSE in the I-MMSE relation (1) to approximate mutual information (as in (2)) and differential entropy.

W. Alghamdi and F. P. Calmon are with the School of Engineering and Applied Science, Harvard University (emails: alghamdi@g.harvard.edu, flavio@seas.harvard.edu).

This material is based upon work supported by the National Science Foundation under Grant Nos. CAREER-1845852, CIF-1900750, and IIS-2040880. The authors also acknowledge Oracle Research for a gift that supported this work. This paper was presented in part at the IEEE International Symposium on Information Theory (ISIT) in 2019 and in 2021.

A. Overview of Main Results

The crux of our work is the study of polynomial approximations of conditional expectations. A surprising result that motivates this study is a negative answer to the question: If X and $N\sim\mathcal{N}(0,1)$ are independent random variables, can $y\mapsto\mathbb{E}[X\mid X+N=y]$ be a nonlinear polynomial? Proposition 1, stated below, shows that if X is integrable (i.e., $\mathbb{E}[|X|]<\infty$), the only way that $\mathbb{E}[X\mid X+N]$ can be a polynomial is if X is Gaussian or constant. In other words, if $y\mapsto\mathbb{E}[X\mid X+N=y]$ is a polynomial, then it is of degree at most 1.

Proposition 1 ([2, Theorem 1])**.** *For $Y=X+N$ where X is an integrable random variable and $N\sim\mathcal{N}(0,1)$ independent of X, the conditional expectation $\mathbb{E}[X\mid Y]$ cannot be a polynomial in Y with degree greater than* 1*. Therefore, the MMSE estimator in a Gaussian channel with finite-variance input is a polynomial if and only if the input is Gaussian or constant.*

Despite the negative result in Proposition 1, we produce a sequence of polynomials converging to the conditional expectation $\mathbb{E}[X\mid Y]$, provided that X has finite variance and Y is light-tailed. For each $n\in\mathbb{N}$, we consider the orthogonal projection of X onto the subspace[1] $\mathscr{P}_n(Y)\subset L^2(P_Y)$ of polynomials in Y with real coefficients and of degree at most n, where it is assumed that $\mathbb{E}[X^2],\mathbb{E}[Y^{2n}]<\infty$. The standard theory of orthogonal projections in Hilbert spaces yields that the orthogonal projection of X onto $\mathscr{P}_n(Y)$, which we denote by $E_n[X\mid Y]$, exists and is unique; indeed, being finite-dimensional, the subspace $\mathscr{P}_n(Y)$ is closed. Further, it is well-known that the orthogonal projection $E_n[X\mid Y]$ is the unique best polynomial approximation of both X and $\mathbb{E}[X\mid Y]$ in the $L^2(P_Y)$ norm (see, e.g., [3, Section 4.4]). From an estimation-theoretic point of view, the operators E_n are natural generalizations of the linear minimum mean-square error (LMMSE) estimate. Hence, we call this process polynomial minimum mean-square (PMMSE) estimation. We

[1]Throughout, we fix a probability space $(\Omega,\mathcal{F},P)$, over which random variables are defined. For $q\geq 1$, the Banach space $L^q(P)$ consists of all q-integrable random variables Z, i.e., $\|Z\|_q:=\left(\int_\Omega|Z|^q\,dP\right)^{1/q}<\infty$. The inner product of the Hilbert space $L^2(P)$ is denoted by $\langle\,\cdot\,,\,\cdot\,\rangle$. The Borel probability measure on $\mathbb{R}$ induced by Y is denoted by P_Y. The Banach subspace $L^q(P_Y)\subset L^q(P)$ consists of $\sigma(Y)$-measurable and q-integrable random variables.

collect these observations in the following definition, in which we denote the random vector $\boldsymbol{Y}^{(n)} := (1, Y, \cdots, Y^n)^T$.

Definition 1 (Polynomial MMSE)**.** Fix $n \in \mathbb{N}$ and two random variables X and Y satisfying $\mathbb{E}\left[X^2\right] < \infty$ and $\mathbb{E}\left[Y^{2n}\right] < \infty$. We define the n-th order *polynomial minimum mean-square error* (PMMSE) for estimating X given Y by

$$\operatorname{pmmse}_n(X \mid Y) := \min_{\boldsymbol{c} \in \mathbb{R}^{n+1}} \mathbb{E}\left[\left(X - \boldsymbol{c}^T \boldsymbol{Y}^{(n)}\right)^2\right]. \tag{3}$$

We define the n-th order PMMSE *estimate* of X given Y by $E_n[X \mid Y] := \boldsymbol{c}^T \boldsymbol{Y}^{(n)} \in \mathscr{P}_n(Y)$ for any minimizer $\boldsymbol{c} \in \mathbb{R}^{n+1}$ in (3).

The PMMSE estimate is the unique minimizer (in $L^2(P_Y)$) of the following two minimization problems

$$E_n[X \mid Y] = \operatorname*{argmin}_{q(Y) \in \mathscr{P}_n(Y)} \mathbb{E}\left[(q(Y) - \mathbb{E}[X \mid Y])^2\right] \tag{4}$$

$$= \operatorname*{argmin}_{q(Y) \in \mathscr{P}_n(Y)} \mathbb{E}\left[(q(Y) - X)^2\right]. \tag{5}$$

Furthermore, we have that the PMMSE satisfies the equality $\operatorname{pmmse}_n(X \mid Y) = \mathbb{E}[(X - E_n[X \mid Y])^2]$. We show in the following result that the PMMSE indeed converges to the MMSE, provided that Y is light-tailed, and we also give an explicit formula for the PMMSE. Recall that Y is said to satisfy Carleman's condition if $\sum_{n=1}^{\infty} \mathbb{E}\left[Y^{2n}\right]^{-1/(2n)} = \infty$, which holds if, e.g., Y has a moment-generating function (MGF) [4, Sec. 4.2]. For $n \in \mathbb{N}$, we denote the n-th order Hankel matrix[2] of moments of Y by $\boldsymbol{M}_{Y,n} := \left(\mathbb{E}\left[Y^{i+j}\right]\right)_{0 \le i,j \le n}$.

Theorem 1. *If X has finite variance and Y satisfies Carleman's condition, then, as $n \to \infty$, we have the convergences $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$ in $L^2(P_Y)$-norm and $\operatorname{pmmse}_n(X \mid Y) \searrow \operatorname{mmse}(X \mid Y)$. Further, for each $n \in \mathbb{N}$, if $|\operatorname{supp}(Y)| > n$ then $E_n[X \mid Y] = \mathbb{E}\left[(X, XY, \cdots, XY^n)\right] \boldsymbol{M}_{Y,n}^{-1} (1, Y, \cdots, Y^n)^T$.*

Proof. We may assume Y has infinite support, for otherwise we would have $\mathbb{E}[X \mid Y] \in \mathscr{P}_{|\operatorname{supp}(Y)|-1}(Y)$ and $\mathbb{E}[X \mid Y] = E_n[X \mid Y]$ for every $n \ge |\operatorname{supp}(Y)| - 1$. Since Y satisfies Carleman's condition, polynomials are dense in $L^2(P_Y)$ [4, Sec. 4.2]. Let $\{p_j(Y) \in \mathscr{P}_j(Y)\}_{j \in \mathbb{N}}$ be the complete orthonormal set in $L^2(P_Y)$ that results from applying Gram-Schmidt orthonormalization to the monomials $\{Y^j\}_{j \in \mathbb{N}}$. By definition of $E_n[X \mid Y]$ as the orthogonal projection of $\mathbb{E}[X \mid Y]$ onto $\mathscr{P}_n(Y)$, we have that $E_n[X \mid Y] = \sum_{j=0}^{n} \langle \mathbb{E}[X \mid Y], p_j(Y) \rangle p_j(Y)$. The $L^2(P_Y)$-norm convergence $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$ follows. Furthermore, by the orthogonality principle of $\mathbb{E}[X \mid Y]$, we have that

$$\begin{aligned} &\operatorname{pmmse}_n(X, t) - \operatorname{mmse}(X, t) \\ &\qquad = \mathbb{E}\left[(E_n[X \mid Y] - \mathbb{E}[X \mid Y])^2\right]. \end{aligned} \tag{6}$$

Since $\mathscr{P}_0(Y) \subset \mathscr{P}_1(Y) \subset \cdots$, we deduce the monotone convergence $\operatorname{pmmse}_n(X \mid Y) \searrow \operatorname{mmse}(X \mid Y)$ from the $L^2(P_Y)$ convergence $E_n[X \mid Y] \to \mathbb{E}[X \mid Y]$. Finally, the formula for $E_n[X \mid Y]$ is shown in Lemma 1. □

[2]Hankel matrices are square matrices with constant skew diagonals.

Remark 1. The convergences in Theorem 1 are stated for Y that is not necessarily a Gaussian perturbation of X. In general, when stating the results of this paper we do not make an implicit assumption on the relationship between X and Y.

We investigate the PMMSE in more detail in the case when Y is the output of a Gaussian channel whose input is X, i.e., $Y = \sqrt{t}X + N$ where $N \sim \mathcal{N}(0, 1)$ is independent of X and $t \ge 0$ is constant. In this case, we show the following rationality of the PMMSE in signal-to-noise ratio (SNR), t. We use the shorthand

$$\operatorname{pmmse}_n(X, t) := \operatorname{pmmse}_n(X \mid \sqrt{t}X + N). \tag{7}$$

Theorem 2. *Fix $n \in \mathbb{N}_{>0}$ and a random variable X satisfying $\mathbb{E}\left[X^{2n}\right] < \infty$. The mapping $t \mapsto \operatorname{pmmse}_n(X, t)$ over $[0, \infty)$ is a rational function, with leading coefficients given by*

$$\operatorname{pmmse}_n(X, t) = \frac{\sigma_X^2 G(n+2) + \cdots + (\det \boldsymbol{M}_{X,n}) t^{d_n - 1}}{G(n+2) + (\sigma_X^2 G(n+2) d_n)\, t + \cdots + (\det \boldsymbol{M}_{X,n}) t^{d_n}}, \tag{8}$$

where $d_n := \binom{n+1}{2}$ and $G(k) := \prod_{j=1}^{k-2} j!$ (for integers $k \ge 1$) is the Barnes G-function [5]. Further, each coefficient in the numerator or denominator of $\operatorname{pmmse}_n(X, t)$ is a multivariate polynomial in $(\mathbb{E}[X], \cdots, \mathbb{E}[X^{2n}])$.

Proof. See Section III-A and Appendix B. □

Remark 2. The PMMSE definition naturally generalizes to random vectors, where orthogonal projection is then done over spaces of multivariate polynomials. In this case, if $\boldsymbol{X}$ is an m-dimensional random vector that is independent of $\boldsymbol{N} \sim \mathcal{N}(0, \boldsymbol{I}_m)$, the leading terms in the PMMSE formula become

$$\operatorname{pmmse}_n(\boldsymbol{X}, t) = \frac{(\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{X}}) \det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + (\operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{N}}) (\det \boldsymbol{M}_{\boldsymbol{X},n})\ t^{d_{n,m}-1}}{\det \boldsymbol{M}_{\boldsymbol{N},n} + \cdots + (\det \boldsymbol{M}_{\boldsymbol{X},n})\ t^{d_{n,m}}}, \tag{9}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{N}}$ are the covariance matrices and $d_{n,m} = m\binom{n+m}{m+1}$; the matrices $\boldsymbol{M}_{\boldsymbol{X},n}$ and $\boldsymbol{M}_{\boldsymbol{N},n}$ are also natural generalizations of the real-valued case, see Appendix D for the details.

The intermediate terms in the rational function $\operatorname{pmmse}_n(X, t)$ can also be given explicitly via Theorem 1. For example, if X is zero-mean and unit-variance, denoting $\mathcal{X}_k = \mathbb{E}[X^k]$, we have the formula

$$\operatorname{pmmse}_2(X, t) = \frac{2 + 4t + (\mathcal{X}_4 - \mathcal{X}_3^2 - 1)t^2}{2 + 6t + (\mathcal{X}_4 + 3)t^2 + (\mathcal{X}_4 - \mathcal{X}_3^2 - 1)t^3}. \tag{10}$$

For a general $n \in \mathbb{N}$, the coefficients in both the numerator and denominator of the PMMSE in (8) are "homogeneous" polynomials in the moments of X (i.e., for a single coefficient $c(X)$ there is a $k_c \in \mathbb{N}$ such that $c(\alpha X) = \alpha^{k_c} c(X)$).

The expression (8) of the PMMSE in terms of moments gives a simple yet powerful method for approximating the MMSE. Figure 1 shows an example of how the PMMSE approximates the MMSE for a random variable X that takes the values 1 and -1 equiprobably, where we are also using

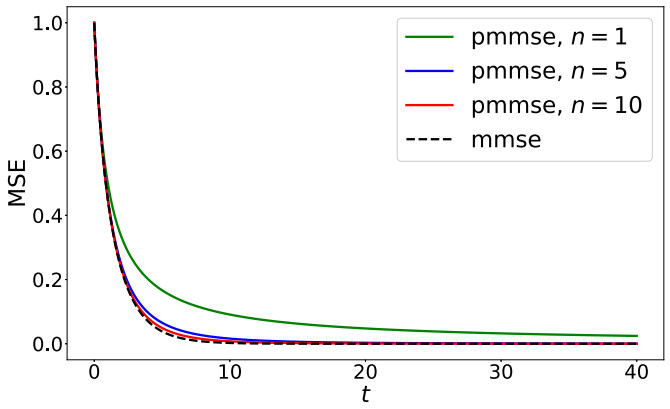

Figure 1. Comparison of the graphs of the functions $t \mapsto \mathrm{pmmse}_n(X,t)$ (solid lines) against the function $t \mapsto \mathrm{mmse}(X,t)$ (dashed black line) for $n \in \{1,5,10\}$ and $X \sim \mathrm{Unif}(\{\pm 1\})$.

the shorthand $\mathrm{mmse}(X,t) := \mathrm{mmse}(X \mid \sqrt{t}X + N)$ for $N \sim \mathcal{N}(0,1)$ independent of X. In this case, the MMSE is given by

$$\mathrm{mmse}(X,t) = 1 - \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \tanh(z\sqrt{t})^2 e^{-(z+\sqrt{t})^2/2}\, dz, \quad (11)$$

whereas the functions $\mathrm{pmmse}_n(X,t)$ are rational in t, e.g., for $n = 1$ we have the LMMSE $\mathrm{pmmse}_1(X,t) = 1/(1+t)$, and for $n = 5$ we have the 5-th degree PMMSE[3]

$$\mathrm{pmmse}_5(X,t) = \frac{45 + 360t + 675t^2 + 300t^3}{45 + 405t + 1035t^2 + 1005t^3 + 450t^4 + 96t^5 + 8t^6}. \quad (12)$$

Comparing the curves in Figure 1 hints that the convergence $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ is uniform in the SNR t; note that the corresponding pointwise convergence is an immediate corollary of Theorem 1. We show in the following result that the PMMSE indeed converges uniformly to the MMSE provided that X has a MGF (i.e., its MGF is finite over a neighborhood of the origin). We also show, under additional assumptions on the distribution of X, that for fixed t the pointwise-convergence rate of $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ is faster than any polynomial in n.

Theorem 3. *If the MGF of a random variable X exists,[4] then we have the uniform and monotone convergence*

$$\sup_{t \geq 0} \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) \searrow 0 \quad (13)$$

as $n \to \infty$. If, in addition, X has a probability density function or a probability mass function p_X that is compactly-supported, even, and decreasing over $[0,\infty) \cap \mathrm{supp}(p_X)$, then for all $k, t \geq 0$ we have that

$$\lim_{n\to\infty} n^k \cdot (\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)) = 0. \quad (14)$$

Proof. See Section III-B and Appendix C. □

Remark 3. By the orthogonality property of the conditional expectation, we have the equality of approximation errors

$$\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) = \mathbb{E}\left[\left(E_n[X \mid \sqrt{t}X + N] - \mathbb{E}[X \mid \sqrt{t}X + N]\right)^2\right], \quad (15)$$

where $N \sim \mathcal{N}(0,1)$ is independent of X. Thus, the convergence rate (14) is equivalent to

$$\lim_{n\to\infty} n^k \mathbb{E}\left[\left(E_n[X \mid \sqrt{t}X + N] - \mathbb{E}[X \mid \sqrt{t}X + N]\right)^2\right] = 0. \quad (16)$$

Equipped with the PMMSE functional, we are able to derive new formulas for differential entropy and mutual information in terms of moments. A corollary of the I-MMSE relation states that the differential entropy of a finite-variance continuous random variable X can be expressed in terms of the MMSE as [1]

$$h(X) = \frac{1}{2} \int_0^\infty \mathrm{mmse}(X,t) - \frac{1}{2\pi e + t}\, dt. \quad (17)$$

Naturally, we consider the functionals obtained by replacing the MMSE with the PMMSE, which we show converge to the differential entropy monotonically from above.

Theorem 4. *Let $\boldsymbol{X}$ be a continuous m-dimensional random vector whose MGF exists. Consider the functionals*

$$h_n(\boldsymbol{X}) := \frac{1}{2} \int_0^\infty \mathrm{pmmse}_n(\boldsymbol{X},t) - \frac{m}{2\pi e + t}\, dt \quad (18)$$

for each $n \in \mathbb{N}_{>0}$. Then, we have a decreasing sequence

$$h(\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{X}})) = \frac{1}{2} \log\left((2\pi e)^m \det \boldsymbol{\Sigma}_{\boldsymbol{X}}\right) \quad (19)$$

$$= h_1(\boldsymbol{X}) \geq h_2(\boldsymbol{X}) \geq \cdots \geq h(\boldsymbol{X}) \quad (20)$$

converging to the differential entropy, $h_n(\boldsymbol{X}) \searrow h(\boldsymbol{X})$.

Proof. See Section IV and Appendix E-B. □

Figure 2 illustrates how $h_n(X)$ approximates $h(X)$, where X has a chi distribution with two degrees of freedom (commonly denoted by χ_2). It is evident from the figure that $h_n(X)$ approximates the differential entropy of X monotonically more accurately as n grows; indeed, this is true in general in view of the monotonicity of the convergence $\mathrm{pmmse}_n(X \mid Y) \searrow \mathrm{mmse}(X \mid Y)$ as in Theorem 1.

A noteworthy implication of Theorem 4 is that it gives a formula for the differential entropy $h(X)$ that, in view of Theorem 2, is entirely in terms of the moments of X. Furthermore, closure properties of polynomial subspaces under affine transformations imply that the PMMSE behaves under affine transformations exactly as the MMSE does: if $\mathbb{E}[X^2], \mathbb{E}[Y^{2n}] < \infty$ then

$$\mathrm{pmmse}_n(aX + b \mid cY + d) = a^2\, \mathrm{pmmse}_n(X \mid Y) \quad (21)$$

[3] In general, $\mathrm{pmmse}_5(Z,t)$ is a ratio of a degree-14 polynomial by a degree-15 polynomial as in equation (8). In the special case of a Rademacher random variable, significant cancellations occur and we obtain equation (12).

[4] The assumption that the MGF of X exists is imposed so that $\sqrt{t}X + N$ satisfies Carleman's condition (for $N \sim \mathcal{N}(0,1)$ independent of X and $t \geq 0$ fixed), which holds because $\sqrt{t}X + N$ will then have a MGF. It is not true in general that Carleman's condition is satisfied by the sum of two independent random variables each satisfying Carleman's condition, see [6, Proposition 3.1].

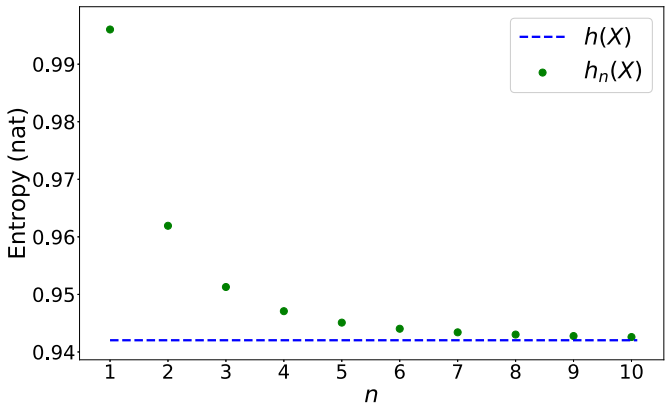

Figure 2. Comparison of the values of $h_n(X)$ (green dots) against the true value $h(X)$ (dashed blue line) for $n \in \{1, \cdots, 10\}$ and $X \sim \chi_2$. We have that $h(X) < h_{10}(X) < h(X) + 6 \cdot 10^{-4}$.

for all constants a, b, c, and d such that $c \neq 0$ (Lemma 2). Thus, the distribution functionals h_n behave under affine transformations exactly as differential entropy does, namely, if $\mathbb{E}[X^{2n}] < \infty$ then

$$h_n(aX + b) = h_n(X) + \log |a| \tag{22}$$

for $a \neq 0$ (Corollary 2).

The moment-based differential entropy formula in Theorem 4 gives rise to formulas of mutual information primarily in terms of moments.

Theorem 5. *If the mutual information $I(X;Y)$ exists (but possibly infinite), then it can be written in terms of the underlying moments in the following two cases:*

1) *Suppose X is discrete with finite support, and Y is continuous whose MGF exists and that satisfies $h(Y) > -\infty$. Then, letting $Y^{(x)}$ denote the random variable obtained from Y by conditioning on $\{X = x\}$, we have*

$$I(X;Y) = \frac{1}{2} \lim_{n\to\infty} \int_0^\infty \operatorname{pmmse}_n(Y,t) - \mathbb{E}_X\left[\operatorname{pmmse}_n(Y^{(X)},t)\right] dt. \tag{23}$$

2) *Suppose that X and Y are continuous whose MGFs exist and that satisfy $h(X), h(Y) > -\infty$. Suppose also that $I(X;Y) < \infty$ or else (X,Y) is not continuous. Then,*

$$I(X;Y) = \frac{1}{2} \lim_{n\to\infty} \int_0^\infty \operatorname{pmmse}_n(X,t) + \operatorname{pmmse}_n(Y,t) - \operatorname{pmmse}_n((X,Y),t)\, dt. \tag{24}$$

Proof. See Section IV and Appendix E-C. □

One result that helps in the proof of Theorem 5 in the second scenario is the following generalization of the MMSE dimension to random vectors.

Theorem 6. *Fix two square-integrable continuous m-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{N}$ that are independent. Suppose that $p_{\boldsymbol{N}}$ is bounded and*[5] *$p_{\boldsymbol{N}}(\boldsymbol{z}) = O\left(\|\boldsymbol{z}\|^{-(m+2)}\right)$ as $\|\boldsymbol{z}\| \to \infty$. Then, we have that*

$$\lim_{t\to\infty} t \cdot \operatorname{mmse}\left(\boldsymbol{X} \mid \sqrt{t}\boldsymbol{X} + \boldsymbol{N}\right) = \operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{N}}. \tag{25}$$

Proof. This result follows by a straightforward extension of the proof for the one-dimensional case given in [7]; see [8, Appendix I] for the full details. □

We introduce new estimators of information measures by approximating the PMMSE in (8) via plugging in sample moments in place of moments. If $\{X_j\}_{j=1}^m$ are i.i.d. samples taken from the distribution of X, then a uniform random variable over the samples $U \sim \operatorname{Unif}(\{X_j\}_{j=1}^m)$ provides an estimate $\operatorname{pmmse}_n(U,t)$ for $\operatorname{pmmse}_n(X,t)$. The moments of U converge to the moments of X by the law of large numbers. Further, using $\operatorname{pmmse}_n(U,t)$ to estimate $\operatorname{pmmse}_n(X,t)$ is a consistent estimator by the continuous mapping theorem, as the PMMSE is a continuous function of the moments. The same can be said of $h_n(U)$ as an estimate of $h_n(X)$, or of $I_n(U;V)$ as an estimate of $I(X;Y)$ when $(U,V) \sim \operatorname{Unif}(\{(X_j,Y_j)\}_{j=1}^m)$ where $\{(X_j,Y_j)\}_{j=1}^m$ are i.i.d. samples drawn according to the distribution of (X,Y) (where I_n is the functional given by the expressions inside the limits in Theorem 5). These estimators also satisfy some desirable properties. For example, the behavior of the PMMSE under affine transformations (21) implies that the estimate of the PMMSE from data is robust to (injective) affine transformations, the functionals h_n behave under affine transformations exactly as differential entropy does, and the same is true for I_n and I.

The rest of the paper is organized as follows. We introduce the PMMSE, provide an explicit formula for it, prove its convergence to the MMSE (Theorem 1), and exhibit some of its properties in Section II. A more detailed treatment of the Gaussian-channel case occupies Section III. Specifically, we show rationality of the PMMSE (Theorem 2) in Section III-A, then prove the uniform convergence of the PMMSE to the MMSE and bound the pointwise-convergence rate (Theorem 3) in Section III-B. Building on the derived results about the PMMSE, we prove new moments-based formulas for differential entropy and mutual information in Section IV. Our formulas then give rise to a new estimator that we introduce in Section V, where simulations also illustrate the estimator's performance.

B. Related Literature

The mutual information between the input and output of the Gaussian channel is known to have an integral relation with the MMSE, referred to in the literature as the I-MMSE relation. This connection was made in the work of Guo, Shamai, and Verdú in [1]. Extensions of the I-MMSE relation were investigated in [9–17], and applications have been established, e.g., in optimal power allocation [18] and monotonicity of non-Gaussianness [19]. Our work is inscribed within this literature.

We introduce the PMMSE approximation of the MMSE, derive new representations of distribution functionals in terms

[5] The exponent $m+2$ in the decay rate may be replaced with $m+1+\varepsilon$ for any $\varepsilon > 0$, see [3, Section 3.2]

of moments, and introduce estimators based on these new representations. We note that utilizing higher-order polynomials as proxies of the MMSE has appeared, e.g., in approaches to denoising [20]. Works such as [21] and [22] show some impossibility results for estimating the MMSE in the general case. Recent work by Diaz et al. [23] gives lower bounds for the MMSE via estimating by neural networks. Also, studying smoothed distributions, e.g., via convolutions with Gaussians, has generated recent interest in the context of information theory [24, 25] and learning theory [26, 27].

At the heart of our work is the Bernstein approximation problem, on which a vast literature exists within approximation theory. The original Bernstein approximation problem extends Weierstrass approximation to the whole real line by investigating whether polynomials are dense in $L^\infty(\mu)$ for a measure μ that is absolutely continuous with respect to the Lebesgue measure. Works such as those by Carleson [28] and Freud [29], and eventually the more comprehensive solution given by Ditzian and Totik [30]—which introduces moduli of smoothness, a natural extension of the modulus of continuity—show that tools used to solve the Bernstein approximation problem can be useful for the more general question of denseness of polynomials in $L^p(\mu)$ for all $p \geq 1$ (see [31] for a comprehensive survey). In particular, the case $p = 2$ has a close relationship with the Hamburger moment problem, described next.

The Hamburger moment problem asks whether a countably-infinite sequence of real numbers corresponds uniquely to the moments of a positive Borel measure on $\mathbb{R}$. A connection between this problem and the Bernstein approximation problem is that if the Hamburger moment problem has a positive answer for the sequence of moments of μ then polynomials are dense in $L^2(\mu)$, see [32]. In the context of information theory, the application of the Bernstein approximation problem and the Hamburger moment problem has appeared in [33].

The denominator of the PMMSE in Gaussian channels, which is given by $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$, as well as the leading coefficient of both the numerator and the denominator, $\det \boldsymbol{M}_{X,n}$, can be seen as generalizations of the Selberg integral. Denote

$$\mathcal{I}_n(\varphi) = \int_{\mathbb{R}^{n+1}} \prod_{0\leq i<j\leq n} (y_i - y_j)^2 \prod_{i=0}^{n} \varphi(y_i)\, dy_0 \cdots dy_n. \quad (26)$$

If φ is the PDF of a Beta distribution or a standard normal distribution, then $\mathcal{I}_n(\varphi)$ is the Selberg integral or the Mehta integral, respectively (both with parameter $\gamma = 1$) [34]. For a continuous random variable Y whose PDF is p_Y,

$$\det \boldsymbol{M}_{Y,n} = \frac{1}{(n+1)!}\mathcal{I}_n(p_Y). \quad (27)$$

The Vandermonde-determinant power $\prod_{i<j}(y_i - y_j)^2$ in the integrand in (26) bears a close connection with the quantum Hall effect [35, 36]. The connection arises via expanding powers of the Vandermonde determinant and investigating which of the ensuing monomials have nonzero coefficients.

We quantify the rate of convergence of the PMMSE to the MMSE in Theorem 3, for which the key ingredient is the bound in Lemma 9 on the derivatives of the conditional expectation. The first-order derivative of the conditional expectation in Gaussian channels has been treated in [37]. We note that in parallel to this work the authors were made aware that the higher-order derivative expressions in Proposition 3 were also derived in [38]. We also extend the proofs for the MMSE dimension in the continuous case as given in [7] to higher dimensions.

Distribution functionals, such as mutual information, are popular metrics for quantifying associations between data (e.g., [39–41]), yet reliably estimating distributional functions directly from samples is a non-trivial task. The naive route of first estimating the underlying distribution is generally impractical and imprecise. To address this challenge, a growing number of distribution functionals' estimators have recently been proposed within the information theory and computer science communities (see, e.g., [42–46]). The estimators proposed in this paper satisfy desirable properties, such as shift invariance and scale resiliency, without the need to estimate the underlying distributions.

C. Notation

Throughout, we fix a probability space $(\Omega, \mathcal{F}, P)$. For $q \geq 1$, the Banach space $L^q(P)$ consists of all q-times integrable real-valued random variables with norm denoted by $\| \cdot \|_q$. The Borel probability measure induced by a random variable Y is denoted by P_Y. The subspace $L^q(P_Y) \subset L^q(P)$ consists of q-times integrable and $\sigma(Y)$-measurable random variables. The inner product of $L^2(P)$ is denoted by $\langle \cdot\,, \cdot \rangle$. The Banach space $L^q(\mathbb{R})$ consists of all q-times Lebesgue integrable functions from $\mathbb{R}$ to itself, with norm denoted by $\| \cdot \|_{L^q(\mathbb{R})}$. We say that Y has a moment-generating function (MGF) if $\mathbb{E}[e^{tY}] < \infty$ over some nonempty interval $t \in (-\delta, \delta)$. We let $\operatorname{supp}(Y)$ denote the support of Y. We denote the cardinality of a set S by $|S|$, and say that Y has *infinite support* if $|\operatorname{supp}(Y)| = \infty$. If $\mathbb{E}\left[Y^{2n}\right] < \infty$, we denote the Hankel matrix of moments by $\boldsymbol{M}_{Y,n} := \left(\mathbb{E}\left[Y^{i+j}\right]\right)_{0\leq i,j\leq n}$. We denote the random vector $\boldsymbol{Y}^{(n)} := (1, Y, \cdots, Y^n)^T$. Note that $\boldsymbol{M}_{Y,n}$ is the expectation of the outer product of $\boldsymbol{Y}^{(n)}$, i.e., $\boldsymbol{M}_{Y,n} = \mathbb{E}\left[\boldsymbol{Y}^{(n)}\left(\boldsymbol{Y}^{(n)}\right)^T\right]$. Therefore, $\boldsymbol{M}_{Y,n}$ is a rank-1 perturbation of the covariance matrix of $\boldsymbol{Y}^{(n)}$, denoted $\boldsymbol{\Sigma}_{\boldsymbol{Y}^{(n)}}$. We let $\mathscr{P}_n$ denote the collection of all polynomials of degree at most n with real coefficients, and we set $\mathscr{P}_n(Y) := \{q(Y)\ ;\ q \in \mathscr{P}_n\}$. For $n \in \mathbb{N}$, we set $[n] := \{0, 1, \cdots, n\}$. Vectors are denoted by boldface letters, in which case subscripted regular letters refer to the entries. The $n \times n$ identity matrix is denoted by $\boldsymbol{I}_n$. The closure of a set S will be denoted by $\overline{S}$. We use the shorthand $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$, and the notation $\mathcal{Y}_k$ is defined analogously.

II. Polynomial MMSE

We give in this section a brief overview of the Polynomial MMSE (PMMSE). The PMMSE, introduced in Definition 1, can be characterized in two equivalent ways: it is the orthogonal projection onto subspaces of polynomials of bounded degree, and it is also a natural generalization of the Linear MMSE (LMMSE) to higher-degree polynomials. Recall that standard results on orthogonal projections in Hilbert spaces

(see, e.g., [3, Section 4.4]) yield that the minimum in (3) is always attained, and that the polynomials $\boldsymbol{c}^T\boldsymbol{Y}^{(n)}$ represent the same element of $\mathscr{P}_n(Y)$ for all minimizers $\boldsymbol{c}$ of (3). In other words, the PMMSE estimate $E_n[X \mid Y]$ as given by Definition 1 is a well-defined element in $\mathscr{P}_n(Y) \subset L^2(P_Y)$.[6]

Unlike the case of the MMSE, working with the PMMSE is tractable and allows for explicit formulas. For instance, the PMMSE in Gaussian channels is a rational function in the SNR; more precisely, the formula for $t \mapsto \mathrm{pmmse}_n(X \mid \sqrt{t}X + N)$ stated in Theorem 2 reveals that this mapping is a rational function of t (where $N \sim \mathcal{N}(0,1)$ is independent of X). In addition, as shown in Theorem 1, we have the strong convergence (i.e., in the strong operator topology) of orthogonal projection operators $E_n[\,\cdot \mid Y] \to \mathbb{E}[\,\cdot \mid Y]$ provided that polynomials in Y are dense in $L^2(P_Y)$.

A. PMMSE Formula

We show next explicit PMMSE formulas. We build on these formulas in the next section to prove the rationality of $t \mapsto \mathrm{pmmse}_n(X,t)$ stated in Theorem 2, which in turn will simplify the proof of consistency of the estimators for information measures introduced in Section V.

Lemma 1. *Fix $n \in \mathbb{N}$, and let X and Y be random variables such that $\mathbb{E}\left[X^2\right], \mathbb{E}\left[Y^{2n}\right] < \infty$. We have that $\boldsymbol{M}_{Y,n}$ is invertible if and only if $|\mathrm{supp}(Y)| > n$. Further, if it is the case that $|\mathrm{supp}(Y)| > n$, then the PMMSE estimator is given by*

$$E_n[X \mid Y] = \mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\,\boldsymbol{Y}^{(n)}, \tag{28}$$

and the PMMSE by

$$\mathrm{pmmse}_n(X \mid Y) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right], \tag{29}$$

which then satisfy the relation

$$\mathrm{pmmse}_n(X \mid Y) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[XE_n[X \mid Y]\right]. \tag{30}$$

Proof. See Appendix A-A. □

To expound on the formulas given by Lemma 1, we instantiate them next for the cases $n \in \{1,2\}$. By definition of the PMMSE, these expressions recover the LMMSE and "quadratic" MMSE. Polynomial regression is also shown below to be an instantiation of the PMMSE.

Example 1. For $n = 1$, if $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$ and $|\mathrm{supp}(Y)| > 1$, we have from (28) that

$$E_1[X \mid Y] = (\mathbb{E}[X], \mathbb{E}[XY]) \begin{pmatrix} 1 & \mathbb{E}[Y] \\ \mathbb{E}[Y] & \mathbb{E}\left[Y^2\right] \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ Y \end{pmatrix}. \tag{31}$$

[6] Uniqueness of the minimizing polynomial $E_n[X \mid Y]$ should not be confused with the possible non-uniqueness of the vector $\boldsymbol{c} \in \mathbb{R}^{n+1}$ in the relation $E_n[X \mid Y] = \boldsymbol{c}^T\boldsymbol{Y}^{(n)}$. For example, if Y is binary and $n = 2$, then $Y^2 = Y$, so for any $c_0, c_1, c_2 \in \mathbb{R}$ for which $E_2[X \mid Y] = c_0 + c_1Y + c_2Y^2$ we also have $E_2[X \mid Y] = c_0 + (c_1 - 1)Y + (c_2 + 1)Y^2$. In particular, there is no unique quadratic $p \in \mathscr{P}_2$ for which $E_2[X \mid Y] = p(Y)$. Nevertheless, in the problems of interest to us, uniqueness of $\boldsymbol{c}$ is also attained (e.g., if Y is continuous); in fact, $\boldsymbol{c}$ is unique if and only if $|\mathrm{supp}(Y)| > n$ holds.

Computing the matrix inverse and multiplying out, we obtain

$$E_1[X \mid Y] = \mathbb{E}[X] + \frac{\mathrm{cov}(X,Y)}{\sigma_Y^2}\left(Y - \mathbb{E}[Y]\right), \tag{32}$$

where $\mathrm{cov}(X,Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ is the covariance between X and Y. Formula (32) indeed gives the LMMSE estimate. Via the relation in (30), we recover

$$\mathrm{pmmse}_1(X \mid Y) = \sigma_X^2 - \frac{\mathrm{cov}(X,Y)^2}{\sigma_Y^2} = \sigma_X^2 \cdot (1 - \rho_{X,Y}^2), \tag{33}$$

with $\rho_{X,Y} := \mathrm{cov}(X,Y)/(\sigma_X\sigma_Y)$ the Pearson correlation coefficient between X and Y (when $\sigma_X \neq 0$). Formula (33) verifies that $\mathrm{pmmse}_1(X \mid Y)$ is the LMMSE. □

Example 2. We will use the notation $\mathcal{Y}_k := \mathbb{E}\left[Y^k\right]$ for short. For $n = 2$, and assuming $\mathbb{E}[X^2], \mathbb{E}[Y^4] < \infty$ and $|\mathrm{supp}(Y)| > 2$, Lemma 1 gives the quadratic $E_2[X \mid Y] = \frac{\alpha_0}{\delta} + \frac{\alpha_1}{\delta}Y + \frac{\alpha_2}{\delta}Y^2$ where

$$\begin{aligned} \alpha_0 = {} & (\mathcal{Y}_2\mathcal{Y}_4 - \mathcal{Y}_3^2)\mathbb{E}[X] + (\mathcal{Y}_2\mathcal{Y}_3 - \mathcal{Y}_1\mathcal{Y}_4)\mathbb{E}[XY] \\ & + (\mathcal{Y}_1\mathcal{Y}_3 - \mathcal{Y}_2^2)\mathbb{E}[XY^2] \end{aligned} \tag{34}$$

$$\begin{aligned} \alpha_1 = {} & (\mathcal{Y}_2\mathcal{Y}_3 - \mathcal{Y}_1\mathcal{Y}_4)\mathbb{E}[X] + (\mathcal{Y}_4 - \mathcal{Y}_2^2)\mathbb{E}[XY] \\ & + (\mathcal{Y}_1\mathcal{Y}_2 - \mathcal{Y}_3)\mathbb{E}[XY^2] \end{aligned} \tag{35}$$

$$\begin{aligned} \alpha_2 = {} & (\mathcal{Y}_1\mathcal{Y}_3 - \mathcal{Y}_2^2)\mathbb{E}[X] + (\mathcal{Y}_1\mathcal{Y}_2 - \mathcal{Y}_3)\mathbb{E}[XY] \\ & + (\mathcal{Y}_2 - \mathcal{Y}_1^2)\mathbb{E}[XY^2] \end{aligned} \tag{36}$$

and

$$\delta = \mathcal{Y}_2\mathcal{Y}_4 - \mathcal{Y}_1^2\mathcal{Y}_4 - \mathcal{Y}_2^3 - \mathcal{Y}_3^2 + 2\mathcal{Y}_1\mathcal{Y}_2\mathcal{Y}_3. \tag{37}$$

Note that $\delta = \det \boldsymbol{M}_{Y,2} \neq 0$ by Lemma 1. Relation (30) then yields the formula

$$\mathrm{pmmse}_2(X \mid Y) = \mathbb{E}\left[X^2\right] - \delta^{-1}\sum_{k=0}^{2}\alpha_k\mathbb{E}\left[XY^k\right]. \tag{38}$$

□

Example 3. Finding the PMMSE estimate can be seen as a generalization of modeling via polynomial regression. The goal of single-variable polynomial regression is to model a random variable X as a polynomial in a random variable Y, i.e., $X = \beta_0 + \beta_1Y + \cdots + \beta_nY^n + \varepsilon$ for a modeling-error random variable ε and constants β_j to be determined from data. Given access to samples $\{(x_i, y_i)\}_{i=1}^m$, this model leads to the equation $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X} = (x_1, \cdots, x_m)^T$, $\boldsymbol{Y} = (y_i^j)_{i\in\{1,\cdots,m\}, j\in[n]}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_m)^T$ where the ε_j are samples from ε, and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_n)^T$. It is assumed that the number of distinct y_i is strictly larger than n, so $\boldsymbol{Y}$ has full column-rank. A value of $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{\varepsilon}\|$ is known from polynomial regression to be $\boldsymbol{\beta}^T = \boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}$. This formula follows from the PMMSE estimate formula in Lemma 1. Indeed, minimizing $\|\boldsymbol{\varepsilon}\|$ in polynomial regression amounts to finding the PMMSE estimate $E_n[U \mid V]$, where $(U,V) \sim \mathrm{Unif}(\{(x_i, y_i)\}_{i=1}^m)$. By the PMMSE formula in Lemma 1, we have that

$$\boldsymbol{\beta}^T = \mathbb{E}\left[U\boldsymbol{V}^{(n)}\right]^T \boldsymbol{M}_{V,n}^{-1} \tag{39}$$

By definition of (U,V), we also have that $\boldsymbol{X}^T\boldsymbol{Y} = m\mathbb{E}\left[U\boldsymbol{V}^{(n)}\right]^T$ and $(\boldsymbol{Y}^T\boldsymbol{Y})^{-1} = \frac{1}{m}\boldsymbol{M}_{V,n}^{-1}$. Multiplying the latter two equations together, we obtain $\boldsymbol{\beta}^T =$

$\boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}$ in view of (39). To sum up, the polynomial regression approach solves the restricted problem of finding the PMMSE $E_n[X' \mid Y']$ when both X' and Y' are discrete with PMFs that evaluate to rational numbers, i.e., when the distribution of (X', Y') is uniform over a finite dataset $\{(x'_i, y'_i)\}_{i=1}^m$. □

Remark 4. We note that $E_n[\,\cdot \mid Y]$ is not in general a conditional expectation operator, i.e., there are some $n \in \mathbb{N}$ and $Y \in L^{2n}(P)$ such that for every sub-σ-algebra $\Sigma \subset \mathcal{F}$ we have $E_n[\,\cdot \mid Y] \neq \mathbb{E}[\,\cdot \mid \Sigma]$. One way to see this is that $E_n[\,\cdot \mid Y]$ might not preserve positivity. For example, if $X \sim \mathrm{Unif}(0,1)$ and $Y = X + N$ for $N \sim \mathcal{N}(0,1)$ independent of X, we have that $E_1[X \mid Y] = (Y+6)/13$ (see (32)). Therefore, the probability that $E_1[X \mid Y] < 0$ is $P_Y((-\infty, -6)) > 0$. In other words, although X is non-negative, $E_1[X \mid Y]$ is not; in contrast, $\mathbb{E}[X \mid \Sigma]$ is non-negative for every sub-σ-algebra $\Sigma \subset \mathcal{F}$.

Remark 5. We may define the pointwise PMMSE estimate $E_n[X \mid Y = y]$ for $y \in \mathrm{supp}(Y)$ by the equation $E_n[X \mid Y = y] := \sum_{j\in[n]} c_j y^j$ where $\boldsymbol{c} = (c_0, \cdots, c_n)^T$ is any minimizer in (3), and a direct verification shows that this makes $E_n[X \mid Y = y]$ well-defined.

Remark 6. The PMMSE, as we introduce it in Definition 1, can be equivalently written in vector LMMSE notation as $\mathrm{pmmse}_n(X \mid Y) = \mathrm{lmmse}(X \mid \boldsymbol{Y}^{(n)})$. However, even when the channel producing Y from X is additive, the same might not be true of that producing $\boldsymbol{Y}^{(n)}$ from X. For example, if $Y = X + N$, then Y^2 contains the cross term XN. For this reason, we use the introduced PMMSE notation in place of the vector LMMSE notation.

B. PMMSE Properties

We investigate next the behavior of the PMMSE under affine transformations, and exhibit a few additional properties of the PMMSE that parallel those of the MMSE. The behavior of the PMMSE under affine transformations, shown in Lemma 2 below, has desirable implications on the moments-based approximations of differential entropy and mutual information that we introduce in Section IV. For example, recall that differential entropy satisfies $h(aY + b) = h(Y) + \log|a|$ for any $a, b \in \mathbb{R}$ with $a \neq 0$. Because of Lemma 2, the same property holds for the approximations h_n (as given by (18)), i.e., $h_n(aY + b) = h_n(Y) + \log|a|$.

For random variables X and Y such that $\mathbb{E}[X^2] < \infty$ and constants $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ such that $\gamma \neq 0$, one has $\mathrm{mmse}(\alpha X + \beta \mid \gamma Y + \delta) = \alpha^2 \mathrm{mmse}(X \mid Y)$ (see, e.g., [1]). This property of the MMSE holds because $\mathrm{mmse}(\,\cdot \mid Y)$ measures the distance to $L^2(P_Y)$, which is a space that is invariant under (injective) affine transformations of Y. A similar reasoning yields an analogous property for the PMMSE.

Lemma 2. *Let X and Y be two random variables and $n \in \mathbb{N}$, and assume that both $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^{2n}]$ are finite. For any $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ such that $\gamma \neq 0$, we have that $\mathrm{pmmse}_n(\alpha X + \beta \mid \gamma Y + \delta) = \alpha^2 \mathrm{pmmse}_n(X \mid Y)$.*

Proof. This property follows from the fact that $\mathscr{P}_n(aY+b) - c = \mathscr{P}_n(Y)$ for $a, b, c \in \mathbb{R}$ with $a \neq 0$. □

We show next that the operator $E_n[\,\cdot \mid Y]$ satisfies several properties analogously to the conditional expectation $\mathbb{E}[\,\cdot \mid Y]$. Note that the properties we derive for the PMMSE cannot be straightforwardly deduced from analogous properties that the conditional expectation satisfies, since $E_n[\,\cdot \mid Y]$ is not in general a conditional expectation operator (see Remark 4). Nevertheless, we have the following PMMSE operator properties.

Lemma 3. *For $n \in \mathbb{N}$ and random variables X, Y, and Z such that $\sigma_X, \sigma_Y, \mathbb{E}[Z^{2n}] < \infty$, the following hold:*

(i) Linearity: $E_n[aX + bY \mid Z] = aE_n[X \mid Z] + bE_n[Y \mid Z]$ for any $a, b \in \mathbb{R}$.
(ii) Invariance: $E_n[p(Z) \mid Z] = p(Z)$ for any $p \in \mathscr{P}_n$.
(iii) Idempotence: $E_n[E_n[X \mid Z] \mid Z] = E_n[X \mid Z]$.
(iv) Contractivity: $\|E_n[X \mid Z]\|_2 \leq \|X\|_2$.
(v) Self-Adjointness: $\mathbb{E}[E_n[X \mid Z]Y] = \mathbb{E}[XE_n[Y \mid Z]]$, i.e., $E_n[\,\cdot \mid Z]$ is self-adjoint.
(vi) Orthogonality: $\mathbb{E}[(X - E_n[X \mid Z])p(Z)] = 0$ for $p \in \mathscr{P}_n$, and $E_n[Y \mid Z] = 0$ if and only if $Y \in \mathscr{P}_n(Z)^\perp$.
(vii) Total expectation: $\mathbb{E}[E_n[X \mid Z]] = \mathbb{E}[X]$.
(viii) Independence: If X and Z are independent, then $E_n[X \mid Z] = \mathbb{E}[X]$.
(ix) Markov Chain: If X—Y—Z forms a Markov chain, then $E_n[\mathbb{E}[X \mid Y] \mid Z] = E_n[X \mid Z]$.

Proof. Properties (i)–(vi) follow immediately from the characterization of $E_n[\,\cdot \mid Z]$ as an orthogonal projection from $L^2(P)$ onto $\mathscr{P}_n(Z)$. Property (vii) follows from the first part of (vi) via linearity of expectation by choosing the constant polynomial $p \equiv 1$. If X and Z are independent, then $X - \mathbb{E}[X] \in \mathscr{P}_n(Z)^\perp$, so we deduce (viii) from the second part of (vi) by choosing $Y = X - \mathbb{E}[X]$. Finally, (ix) is a restatement of $X - \mathbb{E}[X \mid Y] \in \mathscr{P}_n(Z)^\perp$, which can be easily seen to hold when X—Y—Z forms a Markov chain. □

Remark 7. In view of properties (vii)–(viii), one may define the unconditional version of E_n as $E_n[X] := \mathbb{E}[X]$ for $X \in L^2(P)$. With this definition, the total expectation property (vii) becomes $E_n[E_n[X \mid Z]] = E_n[X]$, and the independence property (viii) becomes $E_n[X \mid Z] = E_n[X]$ for independent X and Z. This definition of $E_n[X]$ is also consistent with defining it as $E_n[X \mid 1]$, because $\mathbb{E}[X]$ is the closest constant to X in $L^2(P)$.

We also show that the PMMSE estimate satisfies the "tower property" similarly to the conditional expectation. This property is relegated Proposition 6 in Appendix D-B, where we extend our results on the PMMSE to multiple dimensions.

Next we show that, if X and Z are symmetric random variables[7] that are independent, then the polynomial in $X + Z$ closest to X is always of odd degree or is a constant.

Lemma 4. *For $k \in \mathbb{N}_{\geq 1}$ and symmetric and independent random variables X and Z satisfying $\mathbb{E}[Z^2], \mathbb{E}[X^{4k}] < \infty$*

[7] A random variable Y is symmetric if $P_{Y-a} = P_{-(Y-a)}$ for some $a \in \mathbb{R}$.

and $|\mathrm{supp}(X+Z)| > 2k$*, we have that* $E_{2k}[X \mid X+Z] = E_{2k-1}[X \mid X+Z]$.

Proof. See Appendix A-B. □

Finally, we show that the pointwise PMMSE estimate $E_n[X \mid Y = y]$ (see Remark 5) satisfies the following convergence theorems.

Lemma 5 (Convergence Theorems)**.** *Fix a sequence of square-integrable random variables* $\{X_k\}_{k\in\mathbb{N}}$*, and let* $n \in \mathbb{N}$ *and the random variable* Y *be such that* $\mathbb{E}\left[Y^{2n}\right] < \infty$ *and* $|\mathrm{supp}(Y)| > n$*. For every* $y \in \mathbb{R}$*, the following hold:*

(i) Monotone Convergence: If $\{X_k\}_{k\in\mathbb{N}}$ *is monotone with square-integrable pointwise limit* $X = \lim_{k\to\infty} X_k$*, and either* $Y \geq 0$ *or* $Y \leq 0$ *holds almost surely, then*

$$E_n[X \mid Y = y] = \lim_{k\to\infty} E_n\left[X_k \mid Y = y\right]. \tag{40}$$

(ii) Dominated Convergence: If there is a square-integrable random variable M *such that* $\sup_{k\in\mathbb{N}} |X_k| \leq M$*, and if the pointwise limit* $X := \lim_{k\to\infty} X_k$ *exists, then*

$$E_n[X \mid Y = y] = \lim_{k\to\infty} E_n\left[X_k \mid Y = y\right]. \tag{41}$$

Proof. See Appendix A-C. □

III. PMMSE in Gaussian Channels

We focus in this section on the case $Y = \sqrt{t}X + N$ for $t \geq 0$ and $N \sim \mathcal{N}(0,1)$ independent of X. We prove rationality of the PMMSE (Theorem 2), uniform convergence of the PMMSE to the MMSE, and a pointwise-convergence rate bound (Theorem 3). Investigating the PMMSE in Gaussian channels allows us to extrapolate—via the I-MMSE relation—new formulas for differential entropy and mutual information primarily in terms of moments in the next section, which then pave the way for new estimators for these information measures in Section V. We write

$$\mathrm{pmmse}_n(X,t) := \mathrm{pmmse}_n(X \mid \sqrt{t}X + N), \tag{42}$$

$$\mathrm{mmse}(X,t) := \mathrm{mmse}(X \mid \sqrt{t}X + N), \tag{43}$$

$$\mathrm{lmmse}(X,t) := \mathrm{lmmse}(X \mid \sqrt{t}X + N). \tag{44}$$

Omitted proofs of results stated in this section can be found in Appendix B (for Section III-A) and Appendix C (for Section III-B).

A. Rationality of the PMMSE: Proof of Theorem 2

Fix an integer $n > 0$, and let X be a random variable such that $\mathbb{E}[X^{2n}] < \infty$. We denote the moments of X by $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$, where $\mathcal{X}_0 := 1$. We begin by rewriting the PMMSE as

$$\mathrm{pmmse}_n(X,t) = \frac{\mathrm{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n}}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}}, \tag{45}$$

where $N \sim \mathcal{N}(0,1)$ is independent of X. With some algebra, one can show that the above expresses the PMMSE as a rational function.

Lemma 6. *Fix* $n \in \mathbb{N}$*,* $N \sim \mathcal{N}(0,1)$*, and a random variable* X *that is independent of* N *and which satisfies* $\mathbb{E}\left[X^{2n}\right] < \infty$*. Over* $t \in [0,\infty)$*, the function* $t \mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ *is a polynomial of degree at most* $d_n := \binom{n+1}{2}$*, whose coefficient of* t^{d_n} *is* $\det \boldsymbol{M}_{X,n}$*, coefficient of* t *is* $\sigma_X^2 G(n+2) d_n$*, and constant term is* $G(n+2)$*, where* $G(n+2) := \prod_{k=1}^{n} k!$ *is the Barnes G-function. In addition, over* $t \in [0,\infty)$*, the function* $t \mapsto \mathrm{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ *is a polynomial of degree at most* $d_n - 1$*, whose constant term is* $\sigma_X^2 G(n+2)$*. Furthermore, each coefficient in either of these two polynomials stays unchanged if* X *is shifted by a constant.*

Proof. See Appendix B-A. □

According to Lemma 6, we may define constants $a_X^{n,j}$ and $b_X^{n,j}$ by the polynomial identities

$$\mathrm{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n-1]} a_X^{n,j}\ t^j, \tag{46}$$

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j}\ t^j, \tag{47}$$

and taking the ratio of these two polynomials yields the following rational expression for the PMMSE

$$\mathrm{pmmse}_n(X,t) = \frac{\sum_{j\in[d_n-1]} a_X^{n,j}\ t^j}{\sum_{j\in[d_n]} b_X^{n,j}\ t^j}. \tag{48}$$

Lemma 6 also derives a subset of the desired coefficients values[8]

$$\begin{aligned}&\left(a_X^{n,0}, b_X^{n,0}, b_X^{n,1}, b_X^{n,d_n}\right) = \\ &\left(\sigma_X^2 G(n+2), G(n+2), \sigma_X^2 G(n+2) d_n, \det \boldsymbol{M}_{X,n}\right),\end{aligned} \tag{49}$$

so it only remains to derive the value of a_X^{n,d_n-1}.

Remark 8. We give fully-expanded formulas for each of the $a_X^{n,j}$ and $b_X^{n,j}$ in Appendix B-B, expressing them as integer-coefficient multivariate polynomials in the first $2n$ moments of X. Examining these expressions gives a strengthening of Theorem 2 in which the specific moments that could appear in any of the $a_X^{n,j}$ or $b_X^{n,j}$ are further restricted.

To complete the proof, we show that the value of the leading term in the numerator in (48) is given by

$$a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}. \tag{50}$$

We prove (50) next for continuous X, then generalize for every random variable X.

Assume for now that X is continuous. In particular, $|\mathrm{supp}(X)| = \infty$, so $\det \boldsymbol{M}_{X,n} \neq 0$ according to Lemma 1. In view of $b_X^{n,d_n} = \det \boldsymbol{M}_{X,n}$ (see (49)), showing $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ becomes equivalent to showing $\mathrm{pmmse}_n(X,t) \sim 1/t$ as $t \to \infty$ (see (48)). In addition, the PMMSE is bounded by the LMMSE and the MMSE,

$$\mathrm{mmse}(X,t) \leq \mathrm{pmmse}_n(X,t) \leq \mathrm{lmmse}(X,t). \tag{51}$$

We have that $\mathrm{lmmse}(X,t) \sim 1/t$ as $t \to \infty$. Further, the assumption of continuity of X implies that $\mathrm{mmse}(X,t) \sim 1/t$

[8] Note that Lemma 6 also shows that $a_{X+s}^{n,j} = a_X^{n,j}$ and $b_{X+s}^{n,\ell} = b_X^{n,\ell}$ for each $(j,\ell,s) \in [d_n - 1] \times [d_n] \times \mathbb{R}$, which is a stronger result than shift-invariance of the PMMSE (see Lemma 2); however, we do not utilize this fact in the remainder of the proof.

too [7]. Thus, by (51), we obtain $\mathrm{pmmse}_n(X,t) \sim 1/t$ as $t \to \infty$. We have thus shown the desired equation (50) when X is continuous.

We now return to the general case (i.e., not necessarily continuous X). Note that the quantity $a_X^{n,d_n-1} - \det \boldsymbol{M}_{X,n}$ is a multivariate polynomial in the first $2n$ moments of X. By showing $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ in the previous paragraph for every continuous X, we have established the vanishing of a multivariate polynomial in the first $2n$ moments of every continuous $2n$-times integrable random variable X. We show in Proposition 2 below that such a set of zeros is in fact too large to be contained in the zero-locus of any nonzero polynomial, i.e., that such a polynomial must vanish identically (equivalently, $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ must hold even when X is not continuous). For the proof of the latter claim, we first derive a moment-approximation intermediate result.

Lemma 7. *Fix $m \in \mathbb{N}_{>0}$, set $\ell = \lfloor m/2 \rfloor$ and $\mu_0 = 1$, and let $(\mu_1, \cdots, \mu_m) \in \mathbb{R}^m$ be such that $(\mu_{i+j})_{(i,j)\in[\ell]^2}$ is positive definite. For every $\varepsilon > 0$, there exists a continuous random variable Z such that $\left|\mathbb{E}\left[Z^k\right] - \mu_k\right| < \varepsilon$ for every $k \in [m]$.*

Proof. Since $(\mu_{i+j})_{(i,j)\in[\ell]^2}$ is assumed to be positive definite, the solution to the truncated Hamburger moment problem implies that there is a finitely-supported discrete random variable W such that $\mathbb{E}\left[W^k\right] = \mu_k$ for each $k \in [2\ell+1]$ (see [47, Theorem 3.1, items (iii) and (v)]). Let $U \sim \mathrm{Unif}(0,1)$ be independent of W, and consider the continuous random variables $Z_\eta = W + \eta U$ for $\eta > 0$. For each $k \in [m]$, $Z_\eta^k \to W^k$ in distribution as $\eta \to 0^+$. Further, the set $\{Z_\eta^k\}_{0<\eta\le 1}$ is uniformly integrable since $|Z_\eta^k| \le (|W|+1)^k \in L^1(P)$. By the Lebesgue-Vitali theorem [48, Theorem 4.5.4], we get $\mathbb{E}[Z_\eta^k] \to \mathbb{E}[W^k] = \mu_k$ for each $k \in [m]$ as $\eta \to 0^+$. Hence, for each $\varepsilon > 0$, we may choose $\eta > 0$ small enough so that $|\mathbb{E}[Z_\eta^k] - \mu_k| < \varepsilon$ for every $k \in [m]$, completing the proof. □

In the other direction, if $\mu_0 = 1$ and $(\mu_1, \cdots, \mu_{2\ell}) \in \mathbb{R}^{2\ell}$ come from a continuous random variable Z, i.e., $\mathbb{E}\left[Z^k\right] = \mu_k$ for each $k \in [2\ell]$, then it must be that the Hankel matrix $\boldsymbol{H} = (\mu_{i+j})_{(i,j)\in[\ell]^2}$ is positive definite. Indeed, since $|\mathrm{supp}(Z)| = \infty$, we have that $\boldsymbol{v}^T \boldsymbol{H} \boldsymbol{v} = \left\|\sum_{k\in[\ell]} v_k Z^k\right\|_2^2 > 0$ for every nonzero real vector $\boldsymbol{v} = (v_0, \cdots, v_\ell)^T$.

For each integer $m \ge 2$, let $\mathcal{R}_m \subset L^m(P)$ be the set of all continuous random variables X such that $\mathbb{E}[|X|^m] < \infty$. Consider the set $\mathcal{C}_m \subset \mathbb{R}^m$ defined by $\mathcal{C}_m = \{(\mathbb{E}[X], \cdots, \mathbb{E}[X^m]) \; ; \; X \in \mathcal{R}_m\}$. We have the following result.

Proposition 2. *Let p be a polynomial in m variables with real coefficients. If $p\left(\mathbb{E}[X], \cdots, \mathbb{E}[X^m]\right) = 0$ for every continuous random variable X satisfying $\mathbb{E}[|X|^m] < \infty$, then p is the zero polynomial.*

Proof. See Appendix B-C. □

Proposition 2 completes the proof of Theorem 2. Indeed, since $a_X^{n,d_n-1} - \det \boldsymbol{M}_{X,n} = p\left(\mathbb{E}[X], \cdots, \mathbb{E}[X^{2n}]\right)$ for some multivariate polynomial p, and since we have shown above that p vanishes over $\mathcal{C}_m$, we conclude from Proposition 2 that p vanishes identically. In other words, the equation $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ holds for any random variable X satisfying $\mathbb{E}[X^{2n}] < \infty$ (regardless of whether X is continuous).[9] This completes the proof of Theorem 2. □

We note the following corollary of Theorem 2.

Corollary 1. *For a random variable X satisfying $\mathbb{E}\left[X^{2n}\right] < \infty$, we have that $\mathrm{pmmse}_n(X,0) = \sigma_X^2$, for every $t > 0$ we have the inequalities*

$$\mathrm{pmmse}_n(X,t) \le \frac{\sigma_X^2}{1+\sigma_X^2 t} < \frac{1}{t}, \tag{52}$$

and the function $t \mapsto \mathrm{pmmse}_n(X,t)$ is real-analytic at each $t \in [0,\infty)$. If X also satisfies $|\mathrm{supp}(X)| > n$, then as $t \to \infty$ we have the asymptotic

$$\mathrm{pmmse}_n(X,t) = \frac{1}{t} + O(t^{-2}). \tag{53}$$

Proof. That $\mathrm{pmmse}_n(X,0) = \sigma_X^2$ follows by setting $t = 0$ in (8) or in the definition of the PMMSE. The inequalities in (52) follow since $\mathrm{pmmse}_n(X,t) \le \mathrm{lmmse}(X,t) = \sigma_X^2/(1+\sigma_X^2 t)$. In addition, a rational function is analytic at each point in its domain. For each $t \ge 0$, $|\mathrm{supp}(\sqrt{t}X + N)| = \infty$ where $N \sim \mathcal{N}(0,1)$ independent of X. Therefore, $\boldsymbol{M}_{\sqrt{t}X+N}$ is invertible for every $t \ge 0$, i.e., the denominator in (8) is never zero for $t \ge 0$, so we infer analyticity of $\mathrm{pmmse}_n(X,t)$. Finally, if $|\mathrm{supp}(X)| > n$ then $\det \boldsymbol{M}_{X,n} \neq 0$, so (53) follows from (8). □

B. Convergence of PMMSE to MMSE: Proof of Theorem 3

In Appendix C-A we give the proof of the uniform convergence in (13), namely, that as $n \to \infty$ we have

$$\sup_{t\ge 0} \; \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) \searrow 0 \tag{54}$$

for X having a MGF. In a nutshell, the proof follows from Cantor's intersection theorem in view of continuity of the PMMSE and the MMSE in the SNR, t, and monotonicity of the PMMSE in the polynomial degree, n.

In this subsection, we prove the asymptotic convergence rate stated in (14). Specifically, let $\mathscr{D}$ denote the set of all PDFs or PMFs p that are compactly-supported, even, and decreasing over $[0,\infty) \cap \mathrm{supp}(p)$. Suppose that X is continuous or discrete, with PDF or PMF $p_X \in \mathscr{D}$. We prove next that for any fixed $k, t \ge 0$ we have

$$\lim_{n\to\infty} \; n^k \cdot (\mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)) = 0. \tag{55}$$

Let $N \sim \mathcal{N}(0,1)$ be independent of X, and set $Y = X + N$.

The proof of the convergence rate in (14) relies on results on the Bernstein approximation problem in weighted L^p spaces. In particular, we consider the Freud case [31, Definition 3.3], where the weight is of the form e^{-Q} for Q of polynomial growth, e.g., a Gaussian weight.

Definition 2 (Freud Weight, [31, Definition 3.3])**.** A function $W : \mathbb{R} \to (0,\infty)$ is called a *Freud Weight*, and we write $W \in \mathscr{F}$, if it is of the form $W = e^{-Q}$ for $Q : \mathbb{R} \to \mathbb{R}$ satisfying:

[9] In [8, Appendix L], an alternative proof of $a_X^{n,d_n-1} = \det \boldsymbol{M}_{X,n}$ is given via a self-contained algebraic argument.

(1) Q is even,
(2) Q is differentiable, and $Q'(y) > 0$ for $y > 0$,
(3) $y \mapsto yQ'(y)$ is strictly increasing over $(0, \infty)$,
(4) $yQ'(y) \to 0$ as $y \to 0^+$, and
(5) there exist $\lambda, a, b, c > 1$ such that for every $y > c$ we have $a \le \frac{Q'(\lambda y)}{Q'(y)} \le b$.

One may associate to each Freud weight $W = e^{-Q}$ its Mhaskar–Rakhmanov–Saff numbers $a_n(Q)$, defined next.

Definition 3. If $Q : \mathbb{R} \to \mathbb{R}$ satisfies conditions (2)–(4) in Definition 2, and if $yQ'(y) \to \infty$ as $y \to \infty$, then the *n-th Mhaskar–Rakhmanov–Saff (MRS) number* $a_n(Q)$ *of* Q is defined as the unique positive root a_n of the equation

$$n = \frac{2}{\pi} \int_0^1 \frac{a_n t Q'(a_n t)}{\sqrt{1-t^2}}\, dt. \tag{56}$$

Remark 9. The condition $yQ'(y) \to \infty$ as $y \to \infty$ in Definition 3 is satisfied if e^{-Q} is a Freud weight. Indeed, in view of properties (2)–(3) in Definition 2, the quantity $\ell := \lim_{y\to\infty} yQ'(y)$ is well-defined and it belongs to $(0, \infty]$. If $\ell \ne \infty$, then because $\lim_{y\to\infty} \lambda y Q'(\lambda y) = \ell$ too, property (5) would imply that $a \le 1/\lambda \le b$ contradicting that $\lambda, a > 1$. Therefore, $\ell = \infty$.

For example, the Gaussian weight $W(y) = e^{-y^2}$ is a Freud weight for which $Q(y) = y^2$, and it has the MRS numbers $a_n(Q) = \sqrt{n}$ since $\int_0^1 t^2/\sqrt{1-t^2}\, dt = \frac{\pi}{4}$.

We apply the following Jackson-type theorem.

Theorem 7 ([31, Corollary 3.6])**.** *Fix $W \in \mathscr{F}$, and let u be an r-times continuously differentiable function such that $u^{(r)}$ is absolutely continuous. Let $a_n = a_n(Q)$ where $W = e^{-Q}$, and fix $1 \le s \le \infty$. Then, for some constant $D(W, r, s)$ and every $n \ge \max(r-1, 1)$*

$$\inf_{q \in \mathscr{P}_n} \|(q-u)W\|_{L^s(\mathbb{R})} \le D(W, r, s) \left(\frac{a_n}{n}\right)^r \|u^{(r)} W\|_{L^s(\mathbb{R})}. \tag{57}$$

We will apply the polynomial approximation result stated in Theorem 7 for the $L^2(P_Y)$ norm, i.e., we set $s = 2$, $W = \sqrt{p_Y}$, and $u(y) = \mathbb{E}[X \mid Y = y]$ in Theorem 7. To this end, we will establish the following three facts:

(i) $\sqrt{p_Y} \in \mathscr{F}$,
(ii) $a_n(-\frac{1}{2}\log p_Y) = O_{p_X}(\sqrt{n})$, and
(iii) $\|(d^r/dy^r)\mathbb{E}[X \mid Y = y]\|_2 = O_r(1)$.

The former two facts are established in the following lemma.

Lemma 8. *If $X \sim p$ for some $p \in \mathscr{D}$, and $N \sim \mathcal{N}(0,1)$ is independent of X, then p^s_{X+N} is a Freud weight for any fixed constant $s > 0$. Further, suppose $M > 0$ is such that $\mathrm{supp}(p) \subset [-M, M]$, and denote $Q = -\log p_{X+N}$. Then, for each integer $n \ge 1$ and real $s > 0$, we have the bound*

$$a_n(sQ) \le \left(2M + \sqrt{2}\right)\sqrt{n/s}. \tag{58}$$

Proof. See Appendix C-B. □

Next, we derive a bound on $\|(d^r/dy^r)\mathbb{E}[X \mid Y = y]\|_2$ that depends only on r. We will need the following result showing that the higher-order derivatives of the conditional expectation are given by the conditional cumulants.

Proposition 3 ([2, Proposition 1], [38, Proposition 7])**.** *Fix an integrable random variable X and an independent $N \sim \mathcal{N}(0,1)$, and let $Y = X + N$. For each integer $r \ge 1$ and real y, we have the formula*

$$\frac{d^{r-1}}{dy^{r-1}}\, \mathbb{E}\left[X \mid Y = y\right] = \kappa_r(X \mid Y = y), \tag{59}$$

where $\kappa_r(X \mid Y = y) := \frac{\partial^r}{\partial \tau^r} \log \mathbb{E}\left[e^{\tau X} \mid Y = y\right]\big|_{\tau=0}$ is the r-th conditional cumulant of X given $\{Y = y\}$.

Using Proposition 3, we obtain the following bound on the second moment of the derivatives of the conditional expectation via Hölder's inequality.

Lemma 9. *Fix an integrable random variable X and an independent $N \sim \mathcal{N}(0,1)$, let $Y = X + N$, and fix an integer $r \ge 2$. Denote the constants $q_r := \lfloor(\sqrt{8r+9} - 3)/2\rfloor$, $\gamma_r := (2rq_r)!^{1/(4q_r)}$, and*

$$C_r = \sum_{k=1}^{r} (k-1)! \sum_{j=0}^{k} (-1)^j \binom{r}{j} \left\{ {r-j \atop k-j} \right\}, \tag{60}$$

where $\left\{ {r \atop k} \right\}$ denotes Stirling's number of the second kind.[10] *We have the bound*

$$\left\| \frac{d^{r-1}}{dy^{r-1}}\, \mathbb{E}[X \mid Y = y] \right\|_2 \le\ 2^r C_r\ \min\left(\gamma_r, \|X\|^r_{2rq_r}\right). \tag{61}$$

Proof. See Appendix C-C. □

Remark 10. For $2 \le r \le 7$, we obtain the first few values of q_r as $1, 1, 1, 2, 2, 2$, and we have $q_r \sim \sqrt{2r}$ as $r \to \infty$ (see Remark 18 at the end of the proof in Appendix C-C for a way to reduce q_r). The first few values of C_r (for $2 \le r \le 7$) are given by $1, 1, 4, 11, 56, 267$, and as $r \to \infty$ we have the asymptotic $C_r \sim (r-1)!/\alpha^r$ for some constant $\alpha \approx 1.146$ (see [49]). The crude bound $C_r < r^r$ can also be seen by a combinatorial argument.

We now apply the results of Lemmas 8–9 in Theorem 7 to complete the proof of the convergence rate in (14). Fix a real $k \ge 0$, set $r = \lceil k + 1 \rceil$, and let $n \ge \max(r-1, 1)$ be an integer. We apply Theorem 7 for the conditional expectation function $u(y) = \mathbb{E}[X \mid Y = y]$, the weight $W = \sqrt{p_Y}$, and the exponent $s = 2$. By our choice of weight, $\|vW\|_{L^2(\mathbb{R})} = \|v(Y)\|_2$ for any Borel function $v : \mathbb{R} \to \mathbb{R}$; in particular, this holds for the choice $v(y) = q(y) - \mathbb{E}[X \mid Y = y]$ for any $q \in \mathscr{P}_n$, and also for $v(y) = \frac{d^r}{dy^r}\mathbb{E}[X \mid Y = y]$. Recall from (4) that $E_n[X \mid Y]$ minimizes $\|q(Y) - \mathbb{E}[X \mid Y\|_2$ over $q(Y) \in \mathscr{P}_n(Y)$. Hence, with our choice of W and u, we have

$$\|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2 = \inf_{q \in \mathscr{P}_n} \|(q-u)W\|_{L^2(\mathbb{R})}. \tag{62}$$

By Lemma 8, $W = \sqrt{p_Y}$ is a Freud weight, and we have a bound $a_n(Q) = O_{p_X}(\sqrt{n})$ where $W = e^{-Q}$. In addition, by Lemma 9, we have a bound $\|\frac{d^r}{dy^r}\mathbb{E}[X \mid Y = y]\|_2 = O_r(1)$.

[10] The integer $\left\{ {r \atop k} \right\}$ equals the number of unordered set-partitions of an r-element set into k nonempty subsets. The integer C_r equals the number of cyclically-invariant ordered set-partitions of an r-element set into subsets of sizes at least 2, see sequence A032181 at [49].

Therefore, by Theorem 7, we obtain a constant $D'(p_X,k)$ (depending on $D(\sqrt{p_Y},r,2)$, see (57)) such that

$$\|E_n[X\mid Y]-\mathbb{E}[X\mid Y]\|_2\le\frac{D'(p_X,k)}{n^{\lceil k+1\rceil/2}}. \tag{63}$$

From (63), we conclude

$$n^k\ \|E_n[X\mid Y]-\mathbb{E}[X\mid Y]\|_2^2\le\frac{D'(p_X,k)^2}{n}. \tag{64}$$

Further, by the orthogonality principle of $\mathbb{E}[X\mid Y]$, we have that (see (6))

$$\mathrm{pmmse}_n(X,1)-\mathrm{mmse}(X,1)=\|E_n[X\mid Y]-\mathbb{E}[X\mid Y]\|_2^2. \tag{65}$$

Hence, we conclude from (64) that

$$\lim_{n\to\infty}n^k\ (\mathrm{pmmse}_n(X,1)-\mathrm{mmse}(X,1))=0. \tag{66}$$

Finally, note that the premises of the theorem are also satisfied by $\sqrt{t}X$ for any $t>0$, so we have

$$\lim_{n\to\infty}n^k\ \left(\mathrm{pmmse}_n(\sqrt{t}X,1)-\mathrm{mmse}(\sqrt{t}X,1)\right)=0. \tag{67}$$

Also, one straightforwardly obtains from Lemma 2 that

$$\begin{aligned}&\mathrm{pmmse}_n(X,t)-\mathrm{mmse}(X,t)\\&=\frac{1}{t}\left(\mathrm{pmmse}_n(\sqrt{t}X,1)-\mathrm{mmse}(\sqrt{t}X,1)\right).\end{aligned} \tag{68}$$

Thus, we conclude from (67) the desired asymptotic result that $n^k\left(\mathrm{pmmse}_n(X,t)-\mathrm{mmse}(X,t)\right)\to 0$ as $n\to\infty$ for any fixed reals $k,t\ge 0$ (note that the limit trivially holds for $t=0$ since then both the PMMSE and the MMSE are equal to σ_X^2). □

Remark 11. The convergence rate proved in Theorem 3 is an asymptotic one, and obtaining a finitary version hinges on having explicit characterization of the constants $D(W,r,s)$ in Theorem 7. However, no explicit formula for $D(W,r,s)$ exists in the literature, to the best of our knowledge. To give more details, note that we show in (63) a bound for finite n. Namely, for $k\ge 0$, $r=\lceil k+1\rceil$, and $n\ge\max(r-1,1)$ we have the bound

$$\|E_n[X\mid X+N]-\mathbb{E}[X\mid X+N]\|_2\le\frac{D'(p_X,k)}{n^{r/2}}, \tag{69}$$

where the constant $D'(p_X,k)$ can be chosen as, e.g., with $\mathrm{supp}(p_X)\subset[-M,M]$,

$$\begin{aligned}D'(p_X,k)=D(\sqrt{p_{X+N}},r,2)\cdot\left(2\left(\sqrt{2}M+1\right)\right)^r\\\cdot 2^r C_r\min(\gamma_r,M^r).\end{aligned} \tag{70}$$

Thus, to make explicit the constant of interest to us, $D'(p_X,k)$, it suffices to have an explicit bound on $D(\sqrt{p_{X+N}},r,2)$. However, this latter result, to the best of our knowledge, does not exist in the literature; further, distilling an explicit form for $D(W,r,s)$ from existing proofs is a nontrivial matter. The constants $D(W,r,s)$ carry over from [31, Corollary 3.6], a result that was first proved in [50] (specifically, it is the combination of Theorem 1.2 and Corollary 1.8 in [50]). The constant $D(W,r,s)$ is a universal constant in the sense that Theorem 7 is a Jackson-type theorem, i.e., it gives a polynomial-approximation bound that holds uniformly for all admissible functions u that are to be approximated (although the weight W is fixed). Thus, making $D(W,r,s)$ explicit is in fact a significant improvement on the general approximation-theoretic problem. Note that we do not need to utilize this universality for our PMMSE convergence-rate analysis, since we only need to apply the bound in Theorem 7 for the specific choice of u being the conditional expectation function. This in particular implies the potential of the constant $D(\sqrt{p_{X+N}},r,2)$ being improved for our purposes. Yet, we note that the closely related Jackson-type theorem shown in [51, Theorem 4.1.1] can potentially lead to explicit constants more easily; this result derives inequality (57) in Theorem 7, but with the MRS number a_n replaced with the Freud number q_n (the positive solution to $q_nQ'(q_n)=n$), and it is also premised on a few assumptions on Q''. Finally, since we are interested in guaranteeing convergence in n, the derivation in Theorem 3 is sufficient for our PMMSE analysis. See Remark 13 for further discussion.

Remark 12. Examining the proof of the asymptotic convergence rate in Theorem 3 reveals that it is possible to show that the same convergence rate holds beyond Gaussian channels. Specifically, the following is a blueprint for showing that

$$\begin{aligned}\lim_{n\to\infty}n^k\left(\mathrm{pmmse}_n(X\mid\sqrt{t}X+Z)-\mathrm{mmse}(X\mid\sqrt{t}X+Z)\right)\\=0\end{aligned} \tag{71}$$

for every $k,t\ge 0$, where Z a (non-necessarily Gaussian) continuous noise that is independent of X:

1) Suppose that the random variable $Y=\sqrt{t}X+Z$ is such that the conditional PDFs $p_{Y|X=x}$ form an exponential family. From [52, Proposition 3], the higher-derivative formulas $\frac{d^{r-1}}{dy^{r-1}}\mathbb{E}[X\mid Y=y]=\kappa_r(X\mid Y=y)$ (as in Proposition 3) carries over to this case.
2) The proof of Lemma 9 carries over verbatim to obtain a bound $\left\|\frac{d^{r-1}}{dy^{r-1}}\mathbb{E}[X\mid Y=y]\right\|_2\le 2^rC_r\|X\|_{2rq_r}^r$.
3) Assume that p_Z is a Freud weight, say $p_Z=e^{-Q}$ for $Q(z)\sim z^\ell$ as $z\to\infty$ for some fixed $\ell>1$. Then, the proof of Lemma 8 can be adapted to show that (if, e.g., $p_X\in\mathscr{D}$, where $\mathscr{D}$ is as defined in the beginning of this subsection) the PDF p_Y is also a Freud weight with MRS number of order $n^{1/\ell}$.
4) Applying the Bernstein approximation result stated in Theorem 7, we obtain an upper bound on the approximation error $\mathrm{pmmse}_n(X\mid Y)-\mathrm{mmse}(X\mid Y)$ of order $n^{-k(1-1/\ell)}$ as $n\to\infty$. As this is true for every $k\ge 0$, we conclude the asymptotic rate of convergence $n^k\cdot(\mathrm{pmmse}_n(X\mid Y)-\mathrm{mmse}(X\mid Y))\to 0$ for every $k\ge 0$ and every $t\ge 0$.

IV. New Formulas for Information Measures in Terms of Moments

We apply the derived PMMSE results in the I-MMSE relation to express the differential entropy and mutual information in terms of moments. For example, combining Theorems 2

and 4 shows that for any continuous random variable X that has a MGF, we may express differential entropy as (see (8))

$$h(X) = \frac{1}{2}\lim_{n\to\infty}\int_0^\infty -\frac{1}{2\pi e+t} + \frac{\sigma_X^2 G(n+2)+\cdots+(\det \boldsymbol{M}_{X,n})t^{d_n-1}}{G(n+2)+(\sigma_X^2 G(n+2)d_n)\,t+\cdots+(\det \boldsymbol{M}_{X,n})t^{d_n}}\,dt, \tag{72}$$

where the coefficients of the integrand are all multivariate polynomials in the moments of X. The starting point in deriving this formula is the I-MMSE relation, which we briefly review first.

Theorem 8 (I-MMSE relation, [1])**.** *For any square-integrable random variable X, an independent $N\sim\mathcal{N}(0,1)$, and $\gamma>0$, we have that*

$$I(X;\sqrt{\gamma}X+N) = \frac{1}{2}\int_0^\gamma \operatorname{mmse}(X,t)\,dt. \tag{73}$$

The I-MMSE relation directly yields the following formula for differential entropy: for a square-integrable continuous random variable X we have that [1]

$$h(X) = \frac{1}{2}\log\left(2\pi e\sigma_X^2\right) - \frac{1}{2}\int_0^\infty \frac{\sigma_X^2}{1+\sigma_X^2 t} - \operatorname{mmse}(X,t)\,dt. \tag{74}$$

Since $\int_0^\infty \frac{a}{1+at}-\frac{b}{1+bt}\,dt = \log\frac{a}{b}$ for any $a,b>0$, we may simplify (74) to become

$$h(X) = \frac{1}{2}\int_0^\infty \operatorname{mmse}(X,t) - \frac{1}{2\pi e+t}\,dt. \tag{75}$$

We further extend the representation in (75) to higher dimensions.

Lemma 10. *If the m-dimensional continuous random vector $\boldsymbol{X}$ has a finite covariance matrix, then*

$$h(\boldsymbol{X}) = \frac{1}{2}\int_0^\infty \operatorname{mmse}(\boldsymbol{X},t) - \frac{m}{2\pi e+t}\,dt. \tag{76}$$

Proof. See Appendix E-A. □

The MMSE term in the expression for $h(\boldsymbol{X})$ given in Lemma 10 can be approximated by the PMMSE, resulting in an expression for differential entropy as a function of moments of X. From (74) and (75), and since $\operatorname{mmse}(X,t)\le\operatorname{lmmse}(X,t)$, replacing the MMSE with the LMMSE gives the upper bound on differential entropy $h(X)$

$$h(X)\le h_1(X) := \frac{1}{2}\int_0^\infty \operatorname{lmmse}(X,t) - \frac{1}{2\pi e+t}\,dt \tag{77}$$

$$= \frac{1}{2}\log\left(2\pi e\sigma_X^2\right) = h(\mathcal{N}(0,\sigma_X^2)), \tag{78}$$

which is the maximum possible differential entropy for a continuous random variable with a prescribed variance of σ_X^2. We take this a step further and introduce for each integer $n\ge 1$ (assuming only $\mathbb{E}[X^{2n}]<\infty$) the functional

$$h_n(X) := \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(X,t) - \frac{1}{2\pi e+t}\,dt. \tag{79}$$

By the monotonicity $\operatorname{pmmse}_1(X,t)\ge\operatorname{pmmse}_2(X,t)\ge\cdots\ge\operatorname{mmse}(X,t)$, we also have a monotone sequence $h_1(X)\ge h_2(X)\ge\cdots\ge h(X)$ for a random variable X having moments of all orders. As stated in Theorem 4, which we prove next in the 1-dimensional case, if X also has a MGF then $h_n(X)\searrow h(X)$. The proof for arbitrary dimensions requires extending our PMMSE results to higher dimensions (which we give in Appendix D), hence we relegate it to Appendix E-B.

Proof of Theorem 4 (for the 1-dimensional case). The functions $g_n(t) := \operatorname{lmmse}(X,t)-\operatorname{pmmse}_n(X,t)$ are nonnegative and nondecreasing. By Theorem 3, $g_n\nearrow g$ pointwise, where $g(t) := \operatorname{lmmse}(X,t)-\operatorname{mmse}(t)$. Therefore, by the monotone convergence theorem, $\int_0^\infty g_n(t)\,dt\nearrow\int_0^\infty g(t)\,dt$. Adding and subtracting $1/(2\pi e+t)$ to each integrand, and noting that $t\mapsto\operatorname{lmmse}(X,t)-1/(2\pi e+t)$ is absolutely integrable, we conclude that $h_n(X)\searrow h(X)$. □

Remark 13. It remains a topic of ongoing investigation to derive the convergence rate of the limit $h_n(X)\searrow h(X)$ shown in Theorem 4. Note that we may write the convergence error as

$$h_n(X)-h(X) = \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(X,t)-\operatorname{mmse}(X,t)\,dt. \tag{80}$$

Hence, the convergence rate of $h_n(X)\searrow h(X)$ can be shown if one has the convergence rate of $\operatorname{pmmse}_n(X,t)\searrow\operatorname{mmse}(X,t)$ as a function of t. However, the asymptotic convergence rate bound we show in Theorem 3 does not depend on the parameter t. As discussed in Remark 11, finer characterization of the PMMSE convergence rate hinges on having explicit bounds on the constant $D(W,r,s)$ (see the statement of Theorem 7). This constant is only given implicitly in [50], which is likely due to the universality it enjoys, i.e., the approximation error in Theorem 7 is controlled by $D(W,r,s)$ for a fixed W and *every* function u that is to be approximated by polynomials. In our case, however, we need another type of universality. Precisely, we need to control the best-polynomial error when approximating the class of functions $u_t(y) := \mathbb{E}[X\mid\sqrt{t}X+N=y]$ in their respective weighted Hilbert spaces with weights $W_t := \sqrt{p_{\sqrt{t}X+N}}$ for *every* $t\ge 0$. To the best of our knowledge, no such universality result where the weight can vary parametrically exists in the literature.

The behavior of the PMMSE under affine transformations shown in Lemma 2 implies that each approximation h_n behaves under (injective) affine transformations exactly as differential entropy does.

Corollary 2. *If X is a random variable satisfying $\mathbb{E}[X^{2n}]<\infty$, and $(\alpha,\beta)\in\mathbb{R}^2$ with $\alpha\neq 0$, then we have*

$$h_n(\alpha X+\beta) = h_n(X)+\log|\alpha|. \tag{81}$$

In addition, if X and Y are independent with finite $2n$-th moments, then $h_n(X,Y)=h_n(X)+h_n(Y)$.

The moments-based formula for differential entropy shown in Theorem 4 yields moments-based formulas for mutual information in view of the expansions $I(X;Y)=h(Y)-h(Y\mid X)$ in the discrete-continuous case and $h(X,Y)=h(X)+h(Y)-$

$h(X,Y)$ in the purely continuous case. The proof of these formulas, stated in Theorem 5, is given in Appendix E-C. We discuss here a few implications. If X is discrete and Y is continuous, and if they satisfy the assumptions in the first case of Theorem 5, then we denote the functionals

$$I_n(X;Y) := \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(Y,t) - \mathbb{E}_X\left[\operatorname{pmmse}_n(Y^{(X)},t)\right] dt. \tag{82}$$

Recall that we denote by $Y^{(x)}$ the random variable obtained from Y by conditioning on $\{X = x\}$. If X and Y are continuous satisfying the premises of the second case of Theorem 5, then we denote the functional

$$I_n(X;Y) := \frac{1}{2}\int_0^\infty \operatorname{pmmse}_n(X,t) + \operatorname{pmmse}_n(Y,t) - \operatorname{pmmse}_n((X,Y),t)\, dt. \tag{83}$$

The statement of Theorem 5 is that $I_n(X;Y) \to I(X;Y)$ as $n \to \infty$.

The functionals I_n enjoy properties that resemble those for the mutual information. First, the behavior of the PMMSE under affine transformations exhibited in Lemma 2 implies that $I_n(X;Y)$ is invariant under injective affine transformations of Y. Indeed, this can be seen immediately from the behavior of h_n in Corollary 2. Also, the approximations $I_n(X;Y)$ detect independence exactly.

Corollary 3. *Suppose X and Y are random variables satisfying the premises of Theorem 5 (in either case 1 or case 2). For any constants $(\alpha,\beta) \in \mathbb{R}^2$ with $\alpha \neq 0$, and for any $n \in \mathbb{N}$, we have*

$$I_n(X;\alpha Y+\beta) = I_n(X;Y). \tag{84}$$

In addition, if X and Y are independent, then $I_n(X;Y) = 0$ for every n.

We give full expressions for the first two approximants of mutual information that are generated by the LMMSE and quadratic MMSE, in the discrete-continuous case.

Example 4. When $n = 1$, we obtain

$$I_1(X;Y) = \log \sigma_Y - \mathbb{E}_X\left[\log \sigma_{Y^{(X)}}\right], \tag{85}$$

which is the exact formula for $I(X;Y)$ when both Y is Gaussian and each $Y^{(x)}$ (for $x \in \operatorname{supp}(X)$) is Gaussian; indeed, in such a case, the MMSE is just the LMMSE. □

Example 5. For $n = 2$, we obtain the formula

$$I_2(X;Y) = \frac{1}{6}\log \frac{b_Y^{2,3}}{\prod_{x\in\operatorname{supp}(X)}\left(b_{Y^{(x)}}^{2,3}\right)^{P_X(x)}} + \frac{1}{2}\int_0^\infty \frac{a_Y^{2,1}\, t}{2 + b_Y^{2,1}\, t + b_Y^{2,2}\, t^2 + b_Y^{2,3}\, t^3} - \mathbb{E}_X\left[\frac{a_{Y^{(X)}}^{2,1}\, t}{2 + b_{Y^{(X)}}^{2,1}\, t + b_{Y^{(X)}}^{2,2}\, t^2 + b_{Y^{(X)}}^{2,3}\, t^3}\right] dt \tag{86}$$

where we may compute for any $R \in L^4(P)$

$$b_R^{2,3} := \begin{vmatrix} 1 & \mathbb{E}[R] & \mathbb{E}[R^2] \\ \mathbb{E}[R] & \mathbb{E}[R^2] & \mathbb{E}[R^3] \\ \mathbb{E}[R^2] & \mathbb{E}[R^3] & \mathbb{E}[R^4] \end{vmatrix} \tag{87}$$

$$= \sigma_R^2\mathbb{E}[R^4] + 2\mathbb{E}[R]\mathbb{E}[R^2]\mathbb{E}[R^3] - \mathbb{E}[R^2]^3 - \mathbb{E}[R^3]^2, \tag{88}$$

which is strictly positive when $|\operatorname{supp}(R)| > 2$, and

$$b_R^{2,2} = -4\mathbb{E}[R]\mathbb{E}[R^3] + 3\mathbb{E}[R^2]^2 + \mathbb{E}[R^4] \tag{89}$$

$$b_R^{2,1} = 6\sigma_R^2 \tag{90}$$

$$a_R^{2,1} = 4\mathbb{E}[R]^4 - 8\mathbb{E}[R]^2\mathbb{E}[R^2] + \frac{8}{3}\mathbb{E}[R]\mathbb{E}[R^3] + 2\mathbb{E}[R^2]^2 - \frac{2}{3}\mathbb{E}[R^4]. \tag{91}$$

□

V. Application: Estimation of Information Measures from Data

The approximations introduced in the previous sections naturally motivate estimators for information measures. These estimators are based on (i) approximating moments with sample moments, then (ii) plugging the sample moments into the formulas we have developed for information measures. Since the formulas for information measures depend continuously on the underlying moments, the resulting estimators are asymptotically consistent. Moreover, the estimators also behave as the target information measure under affine transformations, being inherently robust to, for example, rescaling of the samples.

We estimate $h(X)$ from i.i.d. samples $X_1,\cdots,X_m$ as $h_n(U)$ for $U \sim \operatorname{Unif}(\{X_1,\cdots,X_m\})$. More precisely, we introduce the following estimator of differential entropy.

Definition 4. Let $X, X_1,\cdots,X_m$ be i.i.d. continuous random variables, and denote $\mathcal{S} = \{X_j\}_{j=1}^m$. We define the n-th estimate $\widehat{h}_n(\mathcal{S})$ of the differential entropy $h(X)$ as the functional that takes the value $h_n(X)$ if the first $2n$ moments of X are replaced by their respective sample moments. In other words, with $U \sim \operatorname{Unif}(\mathcal{S})$, we set $\widehat{h}_n(\mathcal{S}) := h_n(U)$.

The estimator of mutual information $I(X;Y)$ between a discrete X and a continuous Y is defined next. We utilize Theorem 5. We will need to invert the Hankel matrices of moments $(\mathbb{E}[V^{i+j} \mid U = u])_{i,j\in[n]}$ for each $u \in \operatorname{supp}(U)$, where (U,V) is uniformly distributed over the samples $\mathcal{S} = \{(X_j,Y_j)\}_{j=1}^m$. These Hankel matrices are invertible if and only if for each $u \in \{X_j\}_{j=1}^m$ there are more than n distinct samples (X_j,Y_j) for which $X_j = u$; equivalently, the size of the support set of the random variable V conditioned on $U = u$ exceeds n. Thus, we remove all values u that appear at most n times in the samples $\mathcal{S}$. In other words, we replace $\mathcal{S}$ with the subset

$$\mathcal{S}^{(n)} := \{(X',Y') \in \mathcal{S}\ ;\ |\{1 \le i \le m\ ;\ X_i = X'\}| > n\}. \tag{92}$$

Definition 5. Let $(X,Y),(X_1,Y_1),\cdots,(X_m,Y_m)$ be i.i.d. 2-dimensional random vectors such that X is discrete with finite

support and Y is continuous, and denote $\mathcal{S} = \{(X_j, Y_j)\}_{j=1}^m$. Define $\mathcal{S}^{(1)} \supseteq \mathcal{S}^{(2)} \supseteq \cdots$ by

$$\mathcal{S}^{(n)} := \{(X', Y') \in \mathcal{S} \,;\, |\{1 \le i \le m \,;\, X_i = X'\}| > n\}. \tag{93}$$

For each $n \ge 1$ such that $\mathcal{S}^{(n)}$ is nonempty, let $(U^{(n)}, V^{(n)}) \sim \mathrm{Unif}(\mathcal{S}^{(n)})$. We define the n-th estimate $\widehat{I}_n(\mathcal{S})$ of the mutual information $I(X;Y)$ by $\widehat{I}_n(\mathcal{S}) := I_n(U^{(n)}; V^{(n)})$.

We show in this Appendix F how to implement the proposed estimators numerically. In this section, we prove that the estimators are consistent, and discuss their sample complexity. We end the section by empirically comparing the estimators' performance with other estimators from the literature.

A. Consistency

As sample moments converge almost surely to the moments, and as our expressions for differential entropy and mutual information depend continuously on the moments, the continuous mapping theorem yields that the estimators of differential entropy and mutual information introduced in the beginning of this section are consistent.

Theorem 9. *Let X be a continuous random variable that has a MGF. Let $\{X_j\}_{j=1}^\infty$ be i.i.d. samples drawn according to P_X. Then, for every $n \in \mathbb{N}$, we have the almost-sure convergence*

$$\lim_{m\to\infty} \widehat{h}_n\left(\{X_j\}_{j=1}^m\right) = h_n(X). \tag{94}$$

Furthermore, we have that

$$h(X) = \lim_{n\to\infty} \lim_{m\to\infty} \widehat{h}_n\left(\{X_j\}_{j=1}^m\right) \tag{95}$$

where the convergence in m is almost-sure convergence.

Proof. See Appendix G-A. □

Corollary 4. *Let X be discrete random variable with finite support, and Y be a continuous random variable with a MGF and satisfying $h(Y) > -\infty$. Let $\{(X_j, Y_j)\}_{j=1}^\infty$ be i.i.d. samples drawn according to $P_{X,Y}$. For every $n \in \mathbb{N}$, we have the almost-sure convergence*

$$\lim_{m\to\infty} \widehat{I}_n\left(\{(X_j, Y_j)\}_{j=1}^m\right) = I_n(X;Y). \tag{96}$$

Furthermore,

$$I(X;Y) = \lim_{n\to\infty} \lim_{m\to\infty} \widehat{I}_n\left(\{(X_j, Y_j)\}_{j=1}^m\right) \tag{97}$$

where the convergence in m is almost-sure convergence.

Proof. See Appendix G-B. □

B. Sample Complexity

When X is a continuous random variable of bounded support, we may derive the following sample complexity of the estimator of differential entropy in Definition 4 from Hoeffding's inequality.

Proposition 4. *Fix a bounded-support continuous random variable $X \in L^{2n}(P)$. There is a constant $C = C(X, n)$ such that, for all small enough $\varepsilon, \delta > 0$, any collection $\mathcal{S}$ of i.i.d. samples drawn according to P_X of size*

$$|\mathcal{S}| > \frac{C}{\varepsilon^2} \log \frac{1}{\delta} \tag{98}$$

must satisfy

$$\Pr\left\{\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| < \varepsilon\right\} \ge 1 - \delta. \tag{99}$$

Proof. See Appendix H. □

Remark 14. The sample complexity bound may be rearranged as follows. With $m = |\mathcal{S}|$ denoting the sample size, we have that

$$\Pr\left\{\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \ge \frac{C_1\sqrt{\log(1/\delta)}}{\sqrt{m}}\right\} \le \delta, \tag{100}$$

where C_1 is a constant depending only on p_X and n. There are existing results on the sample complexity rates for estimators that are minimax optimal (see the analysis on the modified Kernel Density Estimator, KDE, in [53]) or near-optimal (see the analysis of the fixed k-nearest neighbor, k-NN, estimator in [54]). These analyses show an upper bound on the root mean-square error $\mathbb{E}\left[\left(\widehat{h}(\mathcal{S}) - h(\boldsymbol{X})\right)^2\right]^{1/2}$ that is roughly of the order $(m \log m)^{-s/(s+d)} + m^{-1/2}$ or $m^{-s/(s+d)} \log m + m^{-1/2}$; here, $\boldsymbol{X}$ is a d-dimensional random vector satisfying certain regularity assumptions that are controlled by the smoothness parameter $s \in (0, 2]$, $\mathcal{S}$ is a set of m i.i.d. samples drawn according to $P_{\boldsymbol{X}}$, and $\widehat{h}$ is the modified KDE or k-NN estimator. When $d = 1$ and $s < 1$ (roughly, X is compactly supported and either does not vanish, or does not vanish smoothly, at the boundary), then the first terms in either of these bounds dominates the $m^{-1/2}$ term. Our bound in (100) contains the relevant asymptotic term $m^{-1/2}$, but it is given instead in terms of probability. Nevertheless, it may be converted to a root mean-square bound of order $\sqrt{(\log m)/m}$ (by choosing $\delta = 1/m$) under the assumption that the probability that the samples $\mathcal{S}$ are well-spaced is not too small, since then one may bound $\widehat{h}_n(\mathcal{S})$ almost surely and apply the reverse Markov inequality. It is worth noting that the sample complexity bound we give in Proposition 4 and (100) holds for all (compactly-supported) PDFs without any regularity assumptions of any kind. However, we also note that the constant in this bound is PDF-dependent.

From Proposition 4, we may also obtain a sample complexity result for the estimate $\widehat{I}_n$ in Definition 5.

Proposition 5. *Fix a finitely-supported discrete random variable X and a bounded-support continuous random variable $Y \in L^{2n}(P)$. There is a constant $C = C(X, Y, n)$ such that, for all small enough $\varepsilon, \delta > 0$, any collection $\mathcal{S}$ of i.i.d. samples drawn according to $P_{X,Y}$ of size*

$$|\mathcal{S}| > \frac{C}{\varepsilon^2} \log \frac{1}{\delta} \tag{101}$$

must satisfy

$$\Pr\left\{\left|\widehat{I}_n(\mathcal{S}) - I_n(X;Y)\right| < \varepsilon\right\} \ge 1 - \delta. \tag{102}$$

Proof. See Appendix H-D. □

C. Numerical Results

We compare via synthetic experiments the performance of our estimators[11] against some of the estimators in the literature.

Our proposed estimator for differential entropy is $\widehat{h}_{10}$, i.e., given samples $\mathcal{S}$ of X we estimate $h(X)$ by $\widehat{h}_{10}(\mathcal{S})$ as given by Definition 4, for a large sample size (e.g., $|\mathcal{S}| > 600$), and it is $\widehat{h}_5$ for a smaller sample size (e.g., $|\mathcal{S}| \leq 600$). We compare this estimator with two estimation methods: k-Nearest-Neighbors (k-NN), and Kernel Density Estimation (KDE). The k-NN-based method we compare against is as provided by the Python package 'entropy_estimators' [56], which we will refer to in this section as KSG. The kernel used for the KDE method is Gaussian, and it is obtained by computing from a set of samples $\{X_j\}_{j=1}^m$ a kernel Φ via the Python function 'scipy.stats.gaussian_kde' [57]; then, the estimate for differential entropy will be $\frac{-1}{m}\sum_{j=1}^m \log \Phi(X_j)$. The parameters for the KSG and the KDE estimators are the default parameters, namely, $k = 3$ for the KSG estimator, and the bandwidth for the KDE estimator is chosen according to Scott's rule (i.e., $m^{-1/(d+4)}$ for a set of m samples of a d-dimensional random vector). We note that a more recent iteration of KDE has been proposed by Han *et al.* in [53], which improves the estimation for the non-smooth part of a PDF.

The mutual information is estimated using $\widehat{I}_5$, i.e., given samples $\mathcal{S}$ of (X, Y) our estimate for $I(X;Y)$ will be $\widehat{I}_5(\mathcal{S})$ as given by Definition 5. This estimator is compared against the partitioning estimator and the Mixed KSG estimator [46] (which is a k-NN-based estimator); we utilize the implementation in [46] for both estimators. In particular, the parameters are fixed throughout, namely, we utilize the parameters used in [46] ($k = 5$ for the Mixed KSG, and 8 bins per dimension for the partitioning estimator).

We perform 250 independent trials for each experiment and each fixed sample size, then plot the absolute error as a percentage of the true value (except for the last experiment, where the ground truth is 0, so we plot the absolute error) against the sample size. The sample sizes chosen for our experiments parallel those in [46], namely, $\{800, 1600, 2400, 3200, 4000\}$. To illustrate the smaller sample size regime, we repeat our Experiment 1 (estimating the differential entropy of Wigner's semicircle law) for sample sizes among $\{100, 200, 400, 600\}$. Since the PMMSE theory we developed in this paper applies only to light-tailed distributions (e.g., those with MGFs), we restrict our experiments to such distributions.

We note that we also performed the mutual information experiments for the Noisy KSG estimator based on the estimator in [42] (with noise strength $\sigma = 0.01$ as in [46]), but its performance was much worse than the other estimators, so we do not include it in the plots.

Remark 15. There is a trade-off between the approximation error $h_n(X) - h(X)$ and the estimation error $|\widehat{h}_n(\mathcal{S}) - h_n(X)|$ as the choice of the polynomial degree n varies. Indeed, as n increases, the approximation error vanishes, since we know that $h_n(X) \searrow h(X)$ by Theorem 4. On the other hand, the estimation error is expected to increase for large n, since the quality of estimating moments via sample moments deteriorates for higher moments and a fixed sample size. Evidently, similar trade-offs can be observed for other estimators in the literature, e.g., for the k-NN estimator one has bias-variance trade-off as k varies. Proposition 4 gives a characterization of the estimation error. To fully understand the best choice of n, one would need both a finer characterization of the constant $C(X, n)$ in Proposition 4 (namely, its dependence on n), and also a convergence rate refinement for $h_n(X) \searrow h(X)$ in Theorem 4 (see Remark 13). Note that the approximation error can be efficiently numerically computed for a given X and n (see Figure 2), and we report this value for the experiments we perform in this section. These experiments show that $n = 5$ gives a favorable estimation error compared to state-of-the-art estimators for moderate sample sizes ($m \leq 600$) and similarly $n = 10$ for larger support sizes ($m > 600$). We note that the compute time it takes to estimate $h(X)$ by $\widehat{h}_5(\mathcal{S})$ is comparable to that of both the k-NN and KDE estimators (in the order of seconds on a commercial laptop), and the compute time for $\widehat{h}_{10}(\mathcal{S})$ is in the order two minutes.

Experiment 1. We estimate the differential entropy of a random variable X distributed according to Wigner's semicircle distribution, i.e.,

$$p_X(x) := \frac{2}{\pi}\sqrt{1-x^2} \cdot 1_{[-1,1]}(x). \tag{103}$$

The ground truth is $h(X) \approx 0.64473$ nats. We generate a set $\mathcal{S}$ of i.i.d. samples distributed according to P_X. The size of $\mathcal{S}$ ranges from 800 to 4000 in increments of 800, and for each fixed sample size we independently generate 250 such sets $\mathcal{S}$ (so we generate a total of 1250 sets of samples). The differential entropy $h(X)$ is estimated by three methods: the moments-based estimator that we propose $\widehat{h}_{10}$, the k-NN-based estimator implemented in [56] (which we refer to as the KSG estimator), and the Gaussian KDE estimator. For the proposed estimator, we use $\widehat{h}_{10}(\mathcal{S})$ as an estimate for $h(X)$. For the KSG estimator, we use the default setting, for which $k = 3$. We also use the default setting for the Gaussian KDE estimator; in particular, the bandwidth is chosen according to Scott's Rule as $m^{-1/(d+4)}$ where $m = |\mathcal{S}|$ and $d = 1$ is the dimensionality of X. The percentage relative absolute error in the estimation (e.g., $100 \cdot |\widehat{h}_{10}(\mathcal{S})/h(X) - 1|$, in %) is plotted against the sample size for the three estimators in Figure 3. The solid lines in Figure 3 are the means of the errors, i.e., the mean in the 250 independent trials of the percentage relative absolute error for each fixed sample size in $\{800, 1600, 2400, 3200, 4000\}$. Via bootstrapping, we infer confidence intervals, which are indicated by the shaded areas around the solid lines in Figure 3. We see that the proposed estimator outperforms the KSG estimator and the KDE estimator for this experiment. We note that we have the true value of the functional $h_{10}(X) \approx 0.64632$ nats (i.e., this is the value if we use the true first 20 moments of X instead of the corresponding sample moments obtained from i.i.d. samples). Hence, the approximation error is $h_{10}(X) - h(X) \approx 0.00159$ nats, i.e., $h_{10}(X)$ is approximately

[11] A Python code can be found at [55].

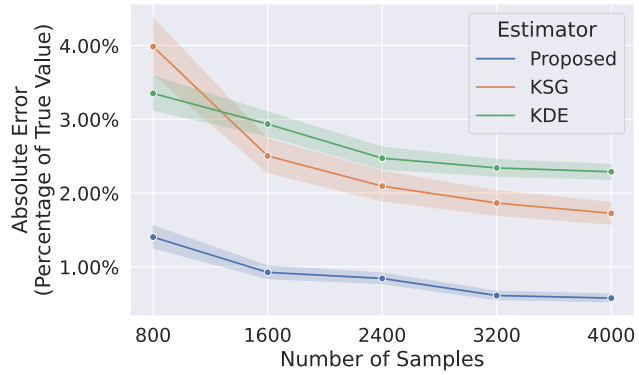

Figure 3. Estimation of differential entropy for a semicircle distribution as in Experiment 1. The vertical axis shows the percentage relative absolute error in the estimation, e.g., for the proposed estimator it is $100 \cdot |\widehat{h}_{10}(\mathcal{S})/h(X) - 1|$ (%) where $\mathcal{S}$ is the set of samples and $h(X) \approx 0.64473$ nats is the ground truth. The horizontal axis shows $|\mathcal{S}|$, the sample size. The proposed estimator $\widehat{h}_{10}$ outperforms the k-NN-based estimator (denoted KSG) and the Gaussian KDE estimator for this experiment.

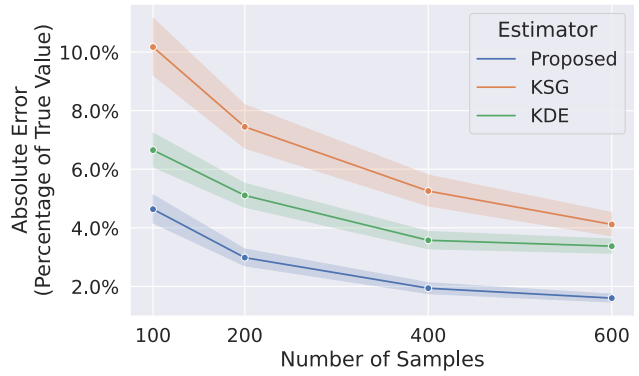

Figure 4. Estimation of differential entropy for a semicircle distribution as in Experiment 1 for the small sample size regime ($100 \leq m \leq 600$). In this regime, the plotted proposed estimator curve refers to the estimation of differential entropy using $\widehat{h}_5$, i.e., $n = 5$. The proposed estimator outperforms both the KSG and the KDE estimators for this experiment in the small sample size regime too.

99.75% accurate when approximating the ground truth $h(X)$ (so $100 - 100 \cdot (h_{10}(X) - h(X))/h(X) \approx 99.75$). For the sake of illustrating the case of smaller sample sizes, we further carry out this experiment with sample sizes in the set $\{100, 200, 400, 600\}$. In this regime, we choose $n = 5$, i.e., our estimator is $\widehat{h}_5$. The results are illustrated in Figure 4. We also notice that the proposed estimator outperforms both the KSG and KDE estimators in this regime. In this case, $h_5(X) \approx 0.6509$ nats, so $h_5(X) - h(X) \approx 0.00617$ nats, giving $h_5(X)$ a 99.04% accuracy as an approximation for $h(X)$.

Experiment 2. We estimate the differential entropy $h(\boldsymbol{X})$ of a random vector $\boldsymbol{X} = (X_1, X_2)^T$ where X_1 and X_2 are i.i.d. distributed according to Wigner's semicircle distribution,

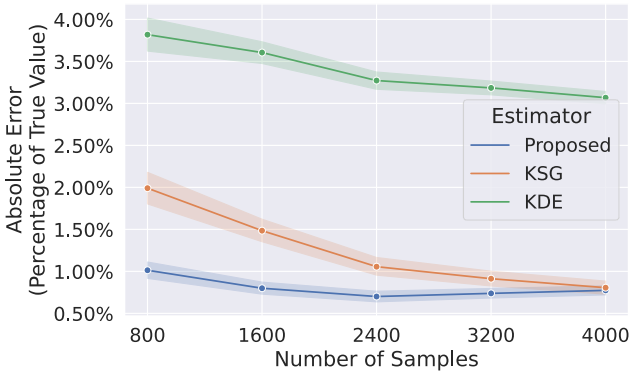

Figure 5. Estimation of differential entropy for a 2-dimensional semicircle distribution as in Experiment 2. The proposed estimator $\widehat{h}_{10}$ outperforms both the KSG and the KDE estimators for this experiment.

namely, $\boldsymbol{X}$ has the PDF

$$p_{\boldsymbol{X}}(x, y) = \frac{4}{\pi^2}\sqrt{(1-x^2)(1-y^2)} \cdot 1_{[-1,1]\times[-1,1]}(x, y). \tag{104}$$

The ground truth is $h(\boldsymbol{X}) \approx 1.28946$ nats. The same numerical setup as in Experiment 1 is performed here. The results are plotted in Figure 5, where we see a similar behavior to the comparison in the 1-dimensional case; in particular, the proposed estimator outperforms the KSG estimator and the KDE estimator for this experiment. By independence of X_1 and X_2, we know that $h(\boldsymbol{X}) = 2h(X_1)$ and $h_{10}(\boldsymbol{X}) = 2h_{10}(X_1)$. Thus, we get the same relative approximation errors as in Experiment 1, namely, $h_{10}(\boldsymbol{X}) - h(\boldsymbol{X}) \approx 0.00318$ nats so $h_{10}(\boldsymbol{X})$ is approximately 99.75% accurate in approximating $h(\boldsymbol{X})$.

Experiment 3. We estimate the differential entropy $h(X)$ of a Gaussian mixture X whose PDF is given by

$$p_X(x) = \sum_{i=1}^{4} \frac{p_i}{\sqrt{2\pi\sigma_i^2}} e^{-(x-\mu_i)^2/(2\sigma_i^2)}, \tag{105}$$

where

$$\boldsymbol{p} = (0.1, 0.2, 0.3, 0.4) \tag{106}$$

$$\boldsymbol{\mu} = (-2, 0, 1, 5) \tag{107}$$

$$\boldsymbol{\sigma} = (1.5, 1, 2, 1). \tag{108}$$

The ground truth is $h(X) \approx 2.34249$ nats. The same numerical setup in Experiments 1 and 2 is used here. The results are plotted in Figure 6. For this experiment, the proposed estimator outperforms the KSG estimator, and it is essentially indistinguishable from the KDE estimator. Note that it is expected that the KDE estimator performs well in this Gaussian mixture experiment, since it is designed specifically to approximate densities by Gaussian mixtures. We have the true value $h_{10}(X) \approx 2.34817$ nats, so the approximation error is $h_{10}(X) - h(X) \approx 0.00568$ nats, making $h_{10}(X)$ approximately 99.76% accurate in approximating the true differential entropy $h(X)$.

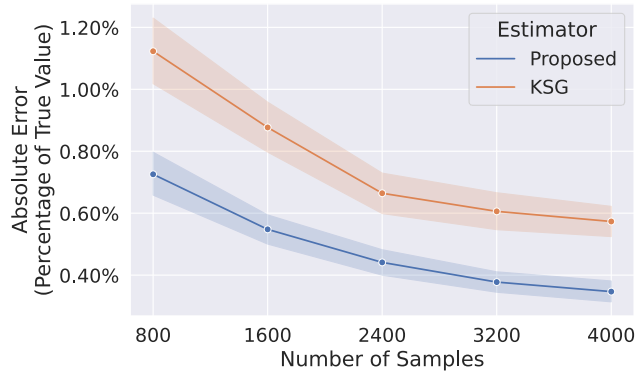

Figure 6. Estimation of differential entropy for a Gaussian mixture as in Experiment 3. The proposed estimator $\widehat{h}_{10}$ outperforms both the KSG and KDE estimators for this experiment. The plot of the KDE estimator's performance is omitted to avoid cluttering, as it lies just above the line for the proposed estimator but overlaps significantly with its uncertainty region.

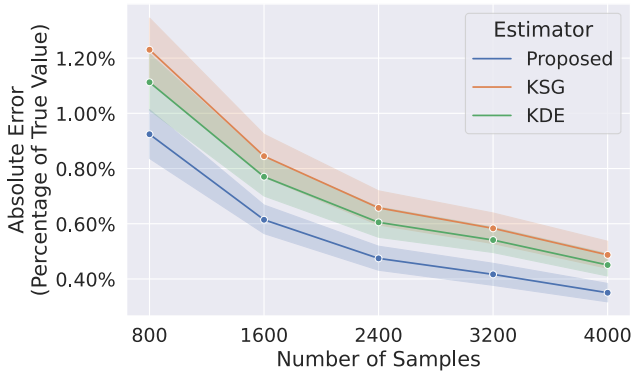

Figure 7. Estimation of differential entropy for a vector Gaussian mixture as in Experiment 4. The proposed estimator $\widehat{h}_{10}$ outperforms both the KSG and KDE estimators for this experiment.

Experiment 4. We estimate the differential entropy $h(\boldsymbol{X})$ of a random vector $\boldsymbol{X}$ that is a mixture of two Gaussians, namely, $\boldsymbol{X}$ has the PDF

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{4\pi\sqrt{\det(\boldsymbol{A})}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2} + \frac{1}{4\pi\sqrt{\det(\boldsymbol{B})}} e^{-(\boldsymbol{x}-\boldsymbol{\nu})^T \boldsymbol{B}^{-1}(\boldsymbol{x}-\boldsymbol{\nu})/2}, \tag{109}$$

where we have the means $\boldsymbol{\mu} = (-1,-1)^T$ and $\boldsymbol{\nu} = (1,1)^T$, and the covariance matrices

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \tag{110}$$

and $\boldsymbol{B} = \boldsymbol{I}_2$. The ground truth is $h(\boldsymbol{X}) \approx 3.22406$ nats. The same numerical setup as in Experiments 1–3 is performed here. The results are plotted in Figure 7. As in the 1-dimensional case in Experiment 3, the proposed estimator outperforms the KSG estimator for this experiment. Further, the proposed estimator also outperforms the KDE estimator in this 2-dimensional setting. We have the true value $h_5(\boldsymbol{X}) \approx 3.22846$ nats, so the approximation error is $h_5(\boldsymbol{X}) - h(\boldsymbol{X}) \approx 0.0044$ nats, making $h_5(\boldsymbol{X})$ approximately 99.86% accurate in approximating the true differential entropy $h(\boldsymbol{X})$.

Experiment 5. We replicate the mixture-distribution part of the zero-inflated Poissonization experiment of [46]. In detail, we let $Y \sim \mathrm{Exp}(1)$, and let $X = 0$ with probability 0.15 and $X \sim \mathrm{Pois}(y)$ given that $Y = y$ with probability 0.85. The quantity to be estimated is the mutual information $I(X;Y)$, and the ground truth is $I(X;Y) \approx 0.25606$ nats. We generate a set of i.i.d. samples $\mathcal{S}$ according to the distribution $P_{X,Y}$, where $\mathcal{S}$ has size in $\{800, 1600, 2400, 3200\}$. We estimate $I(X;Y)$ via the proposed estimator by $\widehat{I}_5(\mathcal{S})$, and we also consider the estimates given by the Mixed KSG estimator and the partitioning estimator, both as implemented in [46] (including the parameters used therein). This estimation process is repeated independently 250 times. The comparison of estimators' performance is plotted in Figure 8. The solid lines indicate the mean percentage relative absolute error, and the shaded areas indicate confidence intervals obtained via bootstrapping. We see in Figure 8 that the proposed estimator outperforms the other considered estimators for this experiment. We note that we have the true value $I_5(X;Y) \approx 0.24677$ nats, which gives an approximation error $|I_5(X;Y) - I(X;Y)| \approx 0.00929$ nats, i.e., $I_5(X;Y)$ is approximately 96.37% accurate in approximating $I(X;Y)$. We also test the affine-transformation invariance property of the proposed estimator. In particular, we consider estimating the mutual information from the scaled samples $\mathcal{S}'$ obtained from $\mathcal{S}$ via scaling the Y samples by 10^4, i.e.,

$$\mathcal{S}' := \{(A, 10^4 B) \ ; \ (A,B) \in \mathcal{S}\}. \tag{111}$$

Plotted in Figure 9 is a comparison of the same estimators using the same samples as those used to generate Figure 8, but where Y is processed through this affine transformation. The ground truth stays unchanged, and so do our estimator and the partitioning estimator, but the Mixed KSG estimates change. This experiment illustrates the resiliency of the proposed estimator to affine transformations. In fact, the computed numerical values in the modified setting by the proposed estimator differ by no more than 10^{-15} nats from those numerically computed in the original setting for each of the 1000 different sets of samples $\mathcal{S}$; in theory, these pairs of values are identical, and the less than 10^{-15} discrepancy is an artifact of the computer implementation. Finally, we note that although the setup is more general than the assumptions we prove our results under in this paper (as X here is not finitely supported), the proposed estimator outperformed the other estimators.

Experiment 6. We test for independence under the following settings. We consider independent $X \sim \mathrm{Bernoulli}(0.5)$ and $Y \sim \mathrm{Unif}([0,2])$. We estimate $I(X;Y)$, whose true value is $I(X;Y) = 0$. We employ the same estimation procedure as in Experiment 5. The results are plotted in Figure 10, which shows that the proposed estimator predicted independence more accurately than the other estimators for the same sample

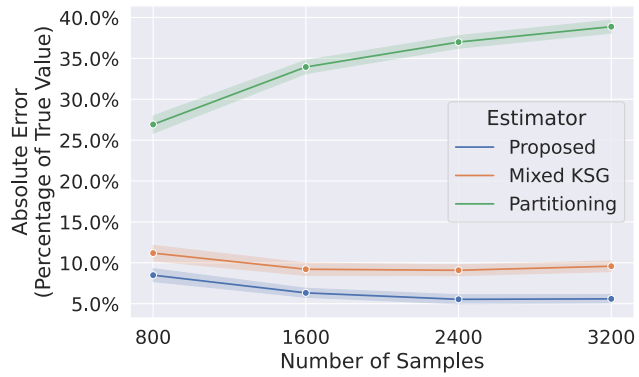

Figure 8. Percentage relative absolute error vs. sample size for unscaled zero-inflated poissonization in Experiment 5. The proposed estimator $\widehat{I}_5$ outperforms both the k-NN-based estimator (denoted Mixed KSG) and the partitioning estimator.

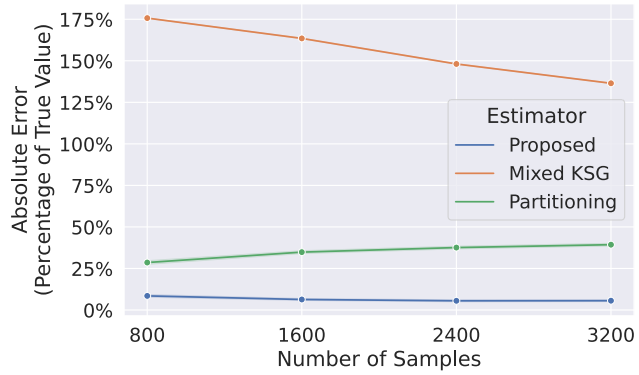

Figure 9. Percentage relative absolute error vs. sample size for the scaled zero-inflated poissonization in Experiment 5. To generate these plots, we use the same samples that yield the plots in Figure 8, but we process them through an affine transformation. Specifically, each sample (A, B) is replaced with $(A, 10^4 B)$. Then the samples are passed to the three estimators. We see that the proposed estimator $\widehat{I}_5$ is resilient to scaling, i.e., the same performance line in Figure 8 is observed here too. This is in contrast to the performance of the Mixed KSG estimator. The partitioning estimator is resilient to scaling, but its performance is not favorable in this experiment (with above 25% relative absolute error).

size. Note that in this case the plot shows the absolute error (in nats) rather than the relative absolute error, as the ground truth is zero. In this case, the true value of $I_5(X;Y)$ is exactly equal to $I(X;Y)$, i.e., $I_5(X;Y)$ is 100% accurate in approximating $I(X;Y)$.

VI. Conclusion

We investigate in this work the interplay between information measures and moments. Via developing the PMMSE, we give polynomial approximations of the conditional expectation. The PMMSE in turn yields new formulas for the differential entropy and mutual information in terms of the underlying moments. These formulas gave rise to a new estimator from data, where simply the moments are estimated from sample moments. The estimator is illustrated in several experiments that indicate a favorable performance as compared

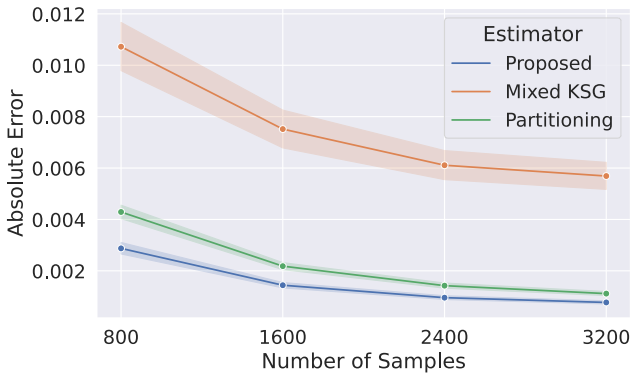

Figure 10. Absolute error (in nats) vs. sample size for the independence testing in Experiment 6. The proposed estimator $\widehat{I}_5$ outperforms the Mixed KSG and the partitioning estimators in this experiment.

to the Gaussian KDE and k-NN estimators. For future work, it is worth investigating the finitary version of the convergence rate of the PMMSE to the MMSE, which would naturally yield convergence rates for the functionals h_n and I_n to the differential entropy and mutual information, and these in turn would tighten the sample complexity analysis. The proposed estimator's performance could also be compared with more recently developed estimators. It is interesting also to apply the PMMSE to the problem of estimating Fisher information, which is tightly related to the MMSE via Brown's identity [58]. Finally, the I-MMSE relation has been extended beyond Gaussian channels (e.g., Poisson channels [59]), and it remains to be seen how the framework we develop in this paper can shed light on those channels.

Appendix A
Proofs of Section II

A. PMMSE Formula: Proof of Lemma 1

The matrix $\boldsymbol{M}_{Y,n}$ is symmetric. We show that it is positive-semidefinite, and that it is positive-definite if and only if $|\mathrm{supp}(Y)| > n$. For any $\boldsymbol{d} \in \mathbb{R}^{n+1}$, we have the inequality

$$\boldsymbol{d}^T \boldsymbol{M}_{Y,n} \boldsymbol{d} = \boldsymbol{d}^T \, \mathbb{E}\left[\boldsymbol{Y}^{(n)} \left(\boldsymbol{Y}^{(n)}\right)^T\right] \boldsymbol{d} \tag{112}$$

$$= \mathbb{E}\left[\boldsymbol{d}^T \boldsymbol{Y}^{(n)} \left(\boldsymbol{Y}^{(n)}\right)^T \boldsymbol{d}\right] \tag{113}$$

$$= \mathbb{E}\left[\left|\boldsymbol{d}^T \boldsymbol{Y}^{(n)}\right|^2\right] \geq 0, \tag{114}$$

so $\boldsymbol{M}_{Y,n}$ is positive-semidefinite. Furthermore, the equality case $\mathbb{E}\left[\left|\boldsymbol{d}^T \boldsymbol{Y}^{(n)}\right|^2\right] = 0$ holds if and only if $\left|\boldsymbol{d}^T \boldsymbol{Y}^{(n)}\right|^2 = 0$, and this latter relation holds if and only if $\boldsymbol{d}^T \boldsymbol{Y}^{(n)} = 0$. Therefore, $\boldsymbol{M}_{Y,n}$ is positive-definite if and only if $\boldsymbol{d}^T \boldsymbol{Y}^{(n)} = 0$ implies $\boldsymbol{d} = \boldsymbol{0}$, i.e., $\boldsymbol{Y}^{(n)}$ does not lie almost surely in a hyperplane in $\mathbb{R}^{n+1}$. Finally, $\boldsymbol{Y}^{(n)}$ lies almost surely in a hyperplane in $\mathbb{R}^{n+1}$ if and only if $|\mathrm{supp}(Y)| \leq n$. Therefore, the desired result that $\boldsymbol{M}_{Y,n}$ is invertible if and only if $|\mathrm{supp}(Y)| > n$ follows.

Next, assume that $|\mathrm{supp}(Y)| > n$, so by what we have shown above, $\boldsymbol{M}_{Y,n}$ is invertible. Let $\boldsymbol{M}_{Y,n}^{1/2}$ denote the lower-triangular matrix in the Cholesky decomposition of $\boldsymbol{M}_{Y,n}$, i.e., $\boldsymbol{M}_{Y,n}^{1/2}$ is the unique lower-triangular matrix with positive diagonal entries that satisfies $\boldsymbol{M}_{Y,n}^{1/2}\left(\boldsymbol{M}_{Y,n}^{1/2}\right)^T = \boldsymbol{M}_{Y,n}$, and denote $\boldsymbol{M}_{Y,n}^{-1/2} := \left(\boldsymbol{M}_{Y,n}^{1/2}\right)^{-1}$. We show that the entries of the vector $\boldsymbol{V} = \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}$ comprise an orthonormal basis for $\mathscr{P}_n(Y)$. We have that

$$\mathbb{E}\left[\boldsymbol{V}\boldsymbol{V}^T\right] = \mathbb{E}\left[\boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}\left(\boldsymbol{Y}^{(n)}\right)^T\left(\boldsymbol{M}_{Y,n}^{-1/2}\right)^T\right] \quad (115)$$

$$= \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{M}_{Y,n}\left(\boldsymbol{M}_{Y,n}^{-1/2}\right)^T = \boldsymbol{I}_{n+1}. \quad (116)$$

Hence, the entries of the vector $\boldsymbol{V}$ form an orthonormal subset of $\mathscr{P}_n(Y)$. Since $\{1, Y, \cdots, Y^n\}$ spans $\mathscr{P}_n(Y)$, and $\boldsymbol{M}_{Y,n}^{-1/2}$ is invertible, we conclude that the entries of $\boldsymbol{V}$ also span $\mathscr{P}_n(Y)$. Hence, the entries of $\boldsymbol{V}$ form an orthonormal basis of $\mathscr{P}_n(Y)$.

Then, the general expansion of orthogonal projections yields the formula $E_n[X \mid Y] = \mathbb{E}\left[X\boldsymbol{V}^T\right]\boldsymbol{V}$. Substituting $\boldsymbol{V} = \boldsymbol{M}_{Y,n}^{-1/2}\boldsymbol{Y}^{(n)}$ we obtain (28). Then, expanding the PMMSE formula $\mathrm{pmmse}_n(X \mid Y) = \mathbb{E}[(X - E_n[X \mid Y])^2]$, we obtain (29). The proof of the lemma is thus complete.

We note that an alternative proof of this lemma, once one obtains the invertibility of $\boldsymbol{M}_{Y,n}$, is via differentiation under the integral sign with respect to the polynomial coefficients in $E_n[X \mid Y]$ in the same way as the LMMSE is usually derived.

B. PMMSE for Symmetric random variables: Proof of Lemma 4

We may assume that X and Z are symmetric around 0, since $E_m[X + a \mid X + Z + b] = a + E_m[X \mid X + Z]$ for every $m \in \mathbb{N}$ and $a, b \in \mathbb{R}$. Then, $\mathbb{E}[X^j] = \mathbb{E}[Z^j] = 0$ for every odd $j \in \mathbb{N}$. Set $Y = X + Z$ and $n = 2k$. Then, $\mathbb{E}[Y^j] = 0$ for every odd $j \in \mathbb{N}$, and $\mathbb{E}[XY^\ell] = 0$ for every even $\ell \in \mathbb{N}$. Then, the coefficient of Y^n in $E_n[X \mid Y]$ is

$$\frac{1}{\det \boldsymbol{M}_{Y,n}} \sum_{\substack{\ell \in [n] \\ \ell \text{ odd}}} \mathbb{E}\left[XY^\ell\right]\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n}, \quad (117)$$

where $\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n}$ denotes the (ℓ, n)-th entry of $\boldsymbol{M}_{Y,n}^{-1}$. Fix an odd $\ell \in [n]$. Let $T_n^{(\ell,n)}$ denote the set of permutations of $[n]$ that send ℓ to n. We have that

$$\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n} = -\sum_{\pi \in T_n^{(\ell,n)}} \mathrm{sgn}(\pi) \prod_{r \in [n] \setminus \{\ell\}} \mathbb{E}\left[Y^{r+\pi(r)}\right]. \quad (118)$$

We have that, for every $\pi \in T_n^{(\ell,n)}$, $\sum_{r \in [n]\setminus\{\ell\}} r + \pi(r) = n(n+1) - \ell - n$, which is odd. Therefore, for at least one $r \in [n] \setminus \{\ell\}$, the integer $r + \pi(r)$ is odd. Hence, $\mathbb{E}[Y^{r+\pi(r)}] = 0$, implying that $\left[\boldsymbol{M}_{Y,n}^{-1}\right]_{\ell,n} = 0$. As this is true for every odd $\ell \in [n]$, we conclude that the coefficient of Y^n in $E_n[X \mid Y]$ is 0. In other words, we have $E_{2k}[X \mid X + Z] = E_{2k-1}[X \mid X + Z]$, and the proof is complete.

C. PMMSE Convergence Theorems: Proof of Lemma 5

Note that in (i) the sequences $\{X_k Y^j\}_{k\in\mathbb{N}}$, for each fixed $j \in [n]$, are monotone almost surely. Also, X_0 is integrable, as we are assuming that $X_0 \in L^2(\mathcal{F})$. Note also that in (ii) each sequence $\{X_k Y^j\}_{k\in\mathbb{N}}$, for $j \in [n]$, is dominated by $M|Y|^j$, which is integrable since both M and Y^j are square-integrable. Thus, monotone convergence and dominated convergence both hold in $L^1(\mathcal{F})$ for each of the sequences $\{X_k Y^j\}_{k\in\mathbb{N}}$, where $j \in [n]$ is fixed. In addition, the formula

$$E_n\left[X_k \mid Y = y\right] = \mathbb{E}\left[X_k \boldsymbol{Y}^{(n)}\right]^T \boldsymbol{M}_{Y,n}^{-1}\left(1, y, \cdots, y^n\right)^T$$

$$= \sum_{j=0}^{n} c_j \mathbb{E}\left[X_k Y^j\right] \quad (119)$$

expresses $E_n\left[X_k \mid Y = y\right]$ as an $\mathbb{R}$-linear combination of $\{X_k Y^j\}_{j\in[n]}$ (where the c_j do not depend on k). Thus, the convergence theorems in (i) and (ii) also hold.

Remark 16. A version of Fatou's lemma that holds for a subset of values of y is also derivable. Namely, suppose that there is a random variable $M \in L^1(\mathcal{F})$ such that $X_k Y^j \geq -M$ for every $(k, j) \in \mathbb{N} \times [n]$, and that $\liminf_{k\to\infty} X_k$ is square-integrable. Then, the same argument in the proof of Lemma 5 shows that

$$\liminf_{k\to\infty} E_n[X_k \mid Y = y] \geq E_n\left[\liminf_{k\to\infty} X_k \,\middle|\, Y = y\right] \quad (120)$$

for every $y \in \mathbb{R}$ such that $\boldsymbol{M}_{Y,n}^{-1}(1, y, \cdots, y^n)^T$ consists of non-negative entries. For example, when $n = 1$, Fatou's lemma holds for $y \geq \mathbb{E}[Y]$ if $\mathbb{E}[Y] \leq 0$, and it holds for $y \in [\mathbb{E}[Y], \mathbb{E}[Y^2]/\mathbb{E}[Y]]$ if $\mathbb{E}[Y] > 0$.

Appendix B
Rationality of the PMMSE (Theorem 2): Proofs of Section III-A

A. Proof of Lemma 6

We introduce the following functions. Recall that we denote $\mathcal{X}_k = \mathbb{E}[X^k]$. For $k \in [n]$, we define the function $v_{X,k} : [0, \infty) \to \mathbb{R}$ at each $t \geq 0$ by

$$v_{X,k}(t) := \mathbb{E}\left[X\left(\sqrt{t}X + N\right)^k\right]. \quad (121)$$

For example, $v_{X,0}(t) = \mathcal{X}_1$, $v_{X,1}(t) = \sqrt{t}\mathcal{X}_2$, and $v_{X,2}(t) = t\mathcal{X}_3 + \mathcal{X}_1$ if $X \in L^3(P)$. Define the vector-valued function $\boldsymbol{v}_{X,n} : [0, \infty) \to \mathbb{R}^{n+1}$ via

$$\boldsymbol{v}_{X,n} := (v_{X,0}, \cdots, v_{X,n})^T. \quad (122)$$

In view of Lemma 1, we may represent the PMMSE as

$$\mathrm{pmmse}_n(X, t) = \mathbb{E}\left[X^2\right] - \boldsymbol{v}_{X,n}(t)^T \boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} \boldsymbol{v}_{X,n}(t). \quad (123)$$

Therefore, defining $F_{X,n} : [0, \infty) \to [0, \infty)$ by

$$F_{X,n}(t) := \boldsymbol{v}_{X,n}(t)^T \boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} \boldsymbol{v}_{X,n}(t), \quad (124)$$

we have the equation

$$\mathrm{pmmse}_n(X, t) = \mathbb{E}\left[X^2\right] - F_{X,n}(t). \quad (125)$$

The functions $F_{X,n}$ are non-negative because the matrices $\boldsymbol{M}_{\sqrt{t}X+N,n}$ are positive-definite (see Lemma 1). In view of (125), PMMSE is fully characterized by $F_{X,n}$, so we focus on this function.

We introduce the following auxiliary polynomials, where R is a random variable independent of N. For ℓ even, we set

$$e_{R,X,\ell}(t) := \mathbb{E}\left[R\left(\sqrt{t}X+N\right)^{\ell}\right], \tag{126}$$

and for ℓ odd we set (for $t>0$)

$$o_{R,X,\ell}(t) := t^{-1/2}\mathbb{E}\left[R\left(\sqrt{t}X+N\right)^{\ell}\right]. \tag{127}$$

That $e_{R,X,\ell}$ and $o_{R,X,\ell}$ are polynomials in t can be seen as follows. Recall that $\mathbb{E}[N^r]=0$ for odd $r\in\mathbb{N}$. If ℓ is even then expanding the right hand side of (126) yields

$$e_{R,X,\ell}(t) = \sum_{\substack{k\in[\ell]\\ k \text{ even}}} \binom{\ell}{k} t^{k/2}\mathbb{E}\left[RX^k\right]\mathbb{E}\left[N^{\ell-k}\right], \tag{128}$$

whereas if ℓ is odd then the right hand side of (127) yields

$$o_{R,X,\ell}(t) = \sum_{\substack{k\in[\ell]\\ k \text{ odd}}} \binom{\ell}{k} t^{(k-1)/2}\mathbb{E}\left[RX^k\right]\mathbb{E}\left[N^{\ell-k}\right]. \tag{129}$$

Both expressions on the right hand sides of (128) and (129) are polynomials of degree at most $\lfloor \ell/2\rfloor$. Further, the coefficient of $t^{\lfloor \ell/2\rfloor}$ in either polynomial is $\mathbb{E}\left[RX^{\ell}\right]$.

Let $\mathcal{S}_{[n]}$ denote the symmetric group of permutations on the $n+1$ elements of $[n]$. We utilize the following auxiliary result on the parity of $i+\pi(i)$ for a permutation $\pi\in\mathcal{S}_{[n]}$.

Lemma 11. *For any permutation $\pi\in\mathcal{S}_{[n]}$, there is an even number of elements $i\in[n]$ such that $i+\pi(i)$ is odd, i.e., the following is an even integer*

$$\delta(\pi) := \left|\{i\in[n]\ ;\ i+\pi(i)\ \text{ is odd}\}\right|. \tag{130}$$

Proof. The integer $i+\pi(i)$ is odd if and only if i and $\pi(i)$ have opposite parities. Thus, the desired result follows from the following more general characterization. For any partition $[n]=A\cup B$, the cardinality of the set

$$I := \{i\in[n]\ ;\ (i,\pi(i))\in(A\times B)\cup(B\times A)\} \tag{131}$$

is even. The desired result follows by letting A and B be even and odd integers, respectively, in $[n]$. Now, we show that the general characterization holds.

Let $A_\pi\subset A$ denote the subset of elements of A that get mapped by π into B, i.e.,

$$A_\pi := \{i\in A\ ;\ \pi(i)\in B\}, \tag{132}$$

and define B_π similarly. Then, $I=A_\pi\cup B_\pi$ is a partition. As $|A_\pi|=|B_\pi|$, we get that $|I|=2|A_\pi|$, and the desired result that $|I|$ is even follows. □

We show first that the function $t\mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t, and show that the coefficient of t^{d_n} in it is $\det \boldsymbol{M}_{X,n}$. By Leibniz's formula,

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi\in\mathcal{S}_{[n]}} \operatorname{sgn}(\pi)\prod_{r\in[n]}\mathbb{E}\left[\left(\sqrt{t}X+N\right)^{r+\pi(r)}\right]. \tag{133}$$

With the auxiliary polynomials $e_{1,X,\ell}$ and $o_{1,X,\ell}$ as defined in (126) and (127) (i.e., with $R=1$), and δ as defined in (130), we may write

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi\in\mathcal{S}_{[n]}} \operatorname{sgn}(\pi) t^{\delta(\pi)/2} \prod_{\substack{i\in[n]\\ i+\pi(i)\text{ odd}}} o_{1,X,i+\pi(i)}(t) \prod_{\substack{j\in[n]\\ j+\pi(j)\text{ even}}} e_{1,X,j+\pi(j)}(t), \tag{134}$$

thereby showing that $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t by evenness of each $\delta(\pi)$ (Lemma 11). Furthermore, for each permutation $\pi\in\mathcal{S}_{[n]}$,

$$\begin{aligned}
&\deg\left(t^{\delta(\pi)/2}\prod_{i+\pi(i)\text{ odd}} o_{1,X,i+\pi(i)}(t)\prod_{j+\pi(j)\text{ even}} e_{1,X,j+\pi(j)}(t)\right)\\
&\le \frac{\delta(\pi)}{2} + \sum_{i+\pi(i)\text{ odd}}\frac{i+\pi(i)-1}{2} + \sum_{j+\pi(j)\text{ even}}\frac{j+\pi(j)}{2} \qquad (135)\\
&= \frac{1}{2}\sum_{k=0}^{n} k+\pi(k) = \frac{n(n+1)}{2} = d_n. \qquad (136)
\end{aligned}$$

Therefore, we also have $\deg\left(\det \boldsymbol{M}_{\sqrt{t}X+N,n}\right)\le d_n$. Finally, taking the terms of highest degrees in $\sqrt{t}$ in (133), we obtain that the coefficient of t^{d_n} in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is

$$\sum_{\pi\in\mathcal{S}_{[n]}}\operatorname{sgn}(\pi)\prod_{r\in[n]}\mathcal{X}_{r+\pi(r)}, \tag{137}$$

which is equal to $\det \boldsymbol{M}_{X,n}$ by the Leibniz determinant formula. This coefficient is non-negative because $\boldsymbol{M}_{X,n}$ is positive-semidefinite, and it is nonzero if and only if $|\operatorname{supp}(X)|>n$ by Lemma 1.

The same approach can be used to show that the mapping $t\mapsto F_{X,n}(t)\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t and to characterize its leading coefficient. In this case, we utilize $e_{X,X,\ell}$ and $o_{X,X,\ell}$ (i.e., $R=X$).

For each $(i,j)\in[n]^2$ let the subset $T_n^{(i,j)}\subset\mathcal{S}_{[n]}$ denote the collection of permutations sending i to j, i.e.,

$$T_n^{(i,j)} := \left\{\pi\in\mathcal{S}_{[n]}\ ;\ \pi(i)=j\right\}. \tag{138}$$

We define, for each $(i,j)\in[n]^2$, the cofactor functions $c_{X,n}^{(i,j)}:[0,\infty)\to\mathbb{R}$ and the products $d_{X,n}^{(i,j)}:[0,\infty)\to\mathbb{R}$ by

$$c_{X,n}^{(i,j)}(t) := \sum_{\pi\in T_n^{(i,j)}}\operatorname{sgn}(\pi)\prod_{\substack{k\in[n]\\ k\neq i}}\left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{k,\pi(k)}, \tag{139}$$

$$d_{X,n}^{(i,j)}(t) := v_{X,i}(t)\ c_{X,n}^{(i,j)}(t)\ v_{X,j}(t). \tag{140}$$

Here, $\left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{a,b}$ is the (a,b)-th entry of $\boldsymbol{M}_{\sqrt{t}X+N,n}$, i.e.,

$$\left(\boldsymbol{M}_{\sqrt{t}X+N,n}\right)_{a,b} = \mathbb{E}\left[\left(\sqrt{t}X+N\right)^{a+b}\right]. \tag{141}$$

Note that $c_{X,n}^{(i,j)}(t)$ is the (i,j)-th cofactor of $\boldsymbol{M}_{\sqrt{t}X+N,n}$. The cofactor matrix $\boldsymbol{C}_{X,n}:[0,\infty)\to\mathbb{R}^{(n+1)\times(n+1)}$ of $t\mapsto\boldsymbol{M}_{\sqrt{t}X+N,n}$ is given by

$$\boldsymbol{C}_{X,n} := \left(c_{X,n}^{(i,j)}\right)_{(i,j)\in[n]^2}. \tag{142}$$

We define the function $D_{X,t} : [0,\infty) \to \mathbb{R}$ by

$$D_{X,n} := \boldsymbol{v}_{X,n}^T \boldsymbol{C}_{X,n} \boldsymbol{v}_{X,n}. \tag{143}$$

We have the following two relations. First, $D_{X,n}$ is the sum of the $d_{X,n}^{(i,j)}$

$$D_{X,n}(t) = \sum_{(i,j)\in[n]^2} d_{X,n}^{(i,j)}(t). \tag{144}$$

Second, by Cramer's rule, and because symmetry of the matrix $\boldsymbol{M}_{\sqrt{t}X+N,n}$ implies that its cofactor is equal to its adjugate, we have the formula

$$\boldsymbol{M}_{\sqrt{t}X+N,n}^{-1} = \frac{1}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} \boldsymbol{C}_{X,n}. \tag{145}$$

Therefore, we obtain

$$F_{X,n}(t) = \frac{D_{X,n}(t)}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} = \frac{\sum_{(i,j)\in[n]^2} d_{X,n}^{(i,j)}(t)}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}}. \tag{146}$$

Hence, it suffices to study the $d_{X,n}^{(i,j)}$.

We start with a characterization of the cofactors $c_{X,n}^{(i,j)}$. Namely, we show that if $i+j$ is even then $c_{X,n}^{(i,j)}(t)$ is a polynomial in t, and if $i+j$ is odd then $\sqrt{t}c_{X,n}^{(i,j)}(t)$ is a polynomial in t. If $i+j$ is even, then

$$\begin{aligned} &c_{X,n}^{(i,j)}(t) \\ &= \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) t^{\delta(\pi)/2} \prod_{\substack{k\in[n]\\ k+\pi(k) \text{ odd}}} o_{1,X,k+\pi(k)}(t) \prod_{\substack{r\in[n],\ r\neq i\\ r+\pi(r) \text{ even}}} e_{1,X,r+\pi(r)}(t), \end{aligned} \tag{147}$$

whereas if $i+j$ is odd then

$$\begin{aligned} &c_{X,n}^{(i,j)}(t) \\ &= \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) t^{\frac{\delta(\pi)-1}{2}} \prod_{\substack{k\in[n],\ k\neq i\\ k+\pi(k) \text{ odd}}} o_{1,X,k+\pi(k)}(t) \prod_{\substack{r\in[n]\\ r+\pi(r) \text{ even}}} e_{1,X,r+\pi(r)}(t). \end{aligned} \tag{148}$$

Thus, evenness of $\delta(\pi)$ for each $\pi \in \mathcal{S}_{[n]}$ implies that each $c_{X,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is even and that each $\sqrt{t}c_{X,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is odd. Further, the degree of $c_{X,n}^{(i,j)}$ for even $i+j$ is upper bounded by

$$\begin{aligned} &\frac{\delta(\pi)}{2} + \sum_{k+\pi(k) \text{ odd}} \frac{k+\pi(k)-1}{2} + \sum_{r+\pi(r) \text{ even}\,;\, r\neq i} \frac{r+\pi(r)}{2} \\ &\quad = \frac{n(n+1)}{2} - \frac{i+j}{2}, \end{aligned} \tag{149}$$

whereas the degree of $\sqrt{t}c_{X,n}^{(i,j)}$ and for odd $i+j$ is upper bounded by

$$\begin{aligned} &\frac{\delta(\pi)}{2} + \sum_{k+\pi(k) \text{ odd}\,;\, k\neq i} \frac{k+\pi(k)-1}{2} + \sum_{r+\pi(r) \text{ even}} \frac{r+\pi(r)}{2} \\ &\quad = \frac{n(n+1)}{2} - \frac{i+j-1}{2}. \end{aligned} \tag{150}$$

We note that both upper bounds are equal to

$$\frac{n(n+1)}{2} - \left\lfloor \frac{i+j}{2} \right\rfloor. \tag{151}$$

Finally, considering the terms of highest order, we see that the term

$$\sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) \prod_{k\in[n]\setminus\{i\}} \mathcal{X}_{k+\pi(k)} \tag{152}$$

is the coefficient of $t^{\frac{n(n+1)}{2} - \left\lfloor \frac{i+j}{2} \right\rfloor}$ in $c_{X,n}^{(i,j)}$ when $i+j$ is even and in $\sqrt{t}c_{X,n}^{(i,j)}$ when $i+j$ is odd.

Now, to show that $D_{X,n}$ is a polynomial, it suffices to check that each $d_{X,n}^{(i,j)}$ is. We consider separately the parity of $i+j$ and build upon the characterization of $c_{X,n}^{(i,j)}$. If $i+j$ is even, so i and j have the same parity, then

$$\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^i\right]\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^j\right]$$

is a polynomial in t of degree at most $(i+j)/2$ with the coefficient of $t^{(i+j)/2}$ being $\mathcal{X}_{i+1}\mathcal{X}_{j+1}$. If $i+j$ is odd, so i and j have different parities, then

$$t^{-1/2}\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^i\right]\mathbb{E}\left[X\left(\sqrt{t}X+N\right)^j\right]$$

is a polynomial in t of degree at most $(i+j-1)/2$ with the coefficient of $t^{(i+j-1)/2}$ being $\mathcal{X}_{i+1}\mathcal{X}_{j+1}$.

Thus, from the characterization of $c_{X,n}^{(i,j)}$, regardless of the parity of $i+j$ we obtain that $d_{X,n}^{(i,j)}(t)$ is a polynomial in t of degree at most $n(n+1)/2 = d_n$. Thus, from (146), the function $t \mapsto F_{X,n}(t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most d_n. Further, note that $\operatorname{pmmse}_n(X,t) \le \operatorname{lmmse}(X,t) \to 0$ as $t\to\infty$. Thus, writing

$$\operatorname{pmmse}_n(X,t) = \frac{(\mathcal{X}_2 - F_{X,n}(t)) \det \boldsymbol{M}_{\sqrt{t}X+N,n}}{\det \boldsymbol{M}_{\sqrt{t}X+N,n}} \tag{153}$$

and recalling that we have shown that $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t of degree at most d_n, we conclude that the numerator $t \mapsto \operatorname{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial of degree at most $d_n - 1$.

Next, we derive the constant terms. Denote by $a_X^{n,0}$ and $b_X^{n,0}$ the constant terms of the polynomials $t \mapsto \operatorname{pmmse}_n(X,t) \det \boldsymbol{M}_{\sqrt{t}X+N,n}$ and $t \mapsto \det \boldsymbol{M}_{\sqrt{t}X+N,n}$, respectively. The formulas for $a_X^{n,0}$ and $b_X^{n,0}$ follow simply by setting $t=0$. Indeed, if $N \sim \mathcal{N}(0,1)$ is independent of X, then

$$F_{X,n}(0) = \mathcal{X}_1^2\ \mathbb{E}\left[\boldsymbol{N}^{(n)}\right]^T \boldsymbol{M}_{N,n}^{-1}\ \mathbb{E}\left[\boldsymbol{N}^{(n)}\right] = \mathcal{X}_1^2 \tag{154}$$

because $\mathbb{E}\left[\boldsymbol{N}^{(n)}\right]$ is the leftmost column of $\boldsymbol{M}_{N,n}$. Therefore,

$$a_X^{n,0} = \sigma_X^2 \det \boldsymbol{M}_{N,n} = \sigma_X^2 b_X^{n,0}. \tag{155}$$

Further, by direct computation or using the connection between Hankel matrices and orthogonal polynomials [60, Appendix A] along with the fact that the probabilist's Hermite polynomials q_k satisfy the recurrence $xq_k(x) = q_{k+1}(x) + kq_{k-1}(x)$, it follows that $\det \boldsymbol{M}_{N,n} = \prod_{k=1}^n k! = G(n+2)$ where G is the Barnes G-function.

Next, we show the last statement in the lemma, namely, that each coefficient in either of the two considered polynomials stays unchanged if X is shifted by a constant. This property will allow us to prove the claim that the coefficient of t in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is $\sigma_X^2 G(n+2)d_n$. By what we have shown thus far, we may define constants $a_X^{n,j}$ and $b_X^{n,j}$ by the polynomial identities

$$\mathrm{pmmse}_n(X,t)\ \det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n-1]} a_X^{n,j}\ t^j, \quad (156)$$

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j}\ t^j. \quad (157)$$

Fix $s\in\mathbb{R}$. For any i.i.d. random variables $Z, Z_0, \cdots, Z_n$, we have that (see, e.g., [60, Appendix A])

$$\det \boldsymbol{M}_{Z,n} = \frac{1}{(n+1)!}\,\mathbb{E}\left[\prod_{0\le i<j\le n}(Z_i-Z_j)^2\right]. \quad (158)$$

From equation (158), since $(Z_i+s)-(Z_j+s)=Z_i-Z_j$, we obtain that

$$\det \boldsymbol{M}_{Z+s,n} = \det \boldsymbol{M}_{Z,n}. \quad (159)$$

Let $N\sim\mathcal{N}(0,1)$ be independent of X. Then, for every $t\in[0,\infty)$, considering $Z=\sqrt{t}X+N$ in (159), we obtain

$$\det \boldsymbol{M}_{\sqrt{t}(X+s)+N,n} = \det \boldsymbol{M}_{\sqrt{t}X+N,n}. \quad (160)$$

As both sides of (160) are polynomials in t, we obtain that $b_{X+s}^{n,j}=b_X^{n,j}$ for every $j\in[d_n]$. Since we also have $\mathrm{pmmse}_n(X+s,t)=\mathrm{pmmse}_n(X,t)$, it follows that

$$t\mapsto \sum_{j\in[d_n-1]} a_X^{n,j}t^j = \mathrm{pmmse}_n(X,t)\sum_{j\in[d_n]} b_X^{n,j}t^j \quad (161)$$

is also invariant under shifting X, so we also obtain $a_{X+s}^{n,j}=a_X^{n,j}$.

By the shift-invariance of $b_X^{n,1}$, we may assume that $\mathcal{X}_1=0$ (so $\mathcal{X}_2=\sigma_X^2$). Now, as each entry in $\boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in $\sqrt{t}$, we see that we may drop any term of order $(\sqrt{t})^3$ or above for the sake of finding $b_X^{n,1}$ (which is the coefficient of t in $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$). In other words,

$$b_X^{n,1} = \det\left(\binom{i+j}{2}\sigma_X^2\mathbb{E}\left[N^{i+j-2}\right]t+\mathbb{E}\left[N^{i+j}\right]\right)_{(i,j)\in[n]^2}. \quad (162)$$

By Leibniz's formula, we conclude

$$b_X^{n,1} = \sigma_X^2\sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k\in[n]}}\mathrm{sgn}(\pi)\binom{k+\pi(k)}{2}\mathbb{E}\left[N^{k+\pi(k)-2}\right]\times \prod_{i\in[n]\setminus\{k\}}\mathbb{E}\left[N^{i+\pi(i)}\right]. \quad (163)$$

But, for any non-negative integer m

$$\binom{m}{2}\mathbb{E}\left[N^{m-2}\right]=\frac{m}{2}\mathbb{E}\left[N^m\right]. \quad (164)$$

Therefore, (163) simplifies to

$$b_X^{n,1} = \frac{\sigma_X^2}{2}\sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k\in[n]}}\mathrm{sgn}(\pi)(k+\pi(k))\prod_{i\in[n]}\mathbb{E}\left[N^{i+\pi(i)}\right]. \quad (165)$$

Evaluating the summation over k for each fixed π, we obtain that

$$b_X^{n,1} = \binom{n+1}{2}\sigma_X^2\sum_{\pi\in\mathcal{S}_{[n]}}\mathrm{sgn}(\pi)\prod_{i\in[n]}\mathbb{E}\left[N^{i+\pi(i)}\right]. \quad (166)$$

Finally, by Leibniz's formula for $\det \boldsymbol{M}_{N,n}$, we obtain that

$$b_X^{n,1} = \binom{n+1}{2}\sigma_X^2\det \boldsymbol{M}_{N,n}, \quad (167)$$

as desired. This completes the proof of Lemma 6.

B. Expanded Formulas for the Coefficients in (48)

As stated in Remark 8, we give here fully-expanded formulas for the coefficients $a_X^{n,j}$ and $b_X^{n,j}$, which will yield further restrictions on which moments can appear in any of these coefficients. Recall that we set $\mathcal{X}_k=\mathbb{E}[X^k]$.

We have the expansion (see (133))

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\pi\in\mathcal{S}_{[n]}}\mathrm{sgn}(\pi)\prod_{r\in[n]}\mathbb{E}\left[\left(\sqrt{t}X+N\right)^{r+\pi(r)}\right] \quad (168)$$

by the Leibniz formula. In the expressions that follow, we denote the tuple $\boldsymbol{k}=(k_0,\cdots,k_n)$. Expanding the powers inside the expectation and computing the expectation, we get a formula of the form

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k_r\in[r+\pi(r)],\ \forall r\in[n]}} t^{(k_0+\cdots+k_n)/2}\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}\beta_{\pi;\boldsymbol{k}}, \quad (169)$$

where the $\beta_{\pi;\boldsymbol{k}}$ are integers given by[12]

$$\beta_{\pi,\boldsymbol{k}} := \mathrm{sgn}(\pi)\prod_{r\in[n]}\binom{r+\pi(r)}{k_r}\mathbb{E}[N^{r+\pi(r)-k_r}]. \quad (170)$$

By Lemma 6, only the summands for which the integer $k_0+\cdots+k_n$ is even can be non-trivial, because $\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial in t. Thus, we have

$$\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} t^j \sum_{\substack{\pi\in\mathcal{S}_{[n]}\\ k_r\in[r+\pi(r)],\ \forall r\in[n]\\ k_0+\cdots+k_n=2j}}\beta_{\pi;\boldsymbol{k}}\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}. \quad (171)$$

[12]From this formula, one may deduce an alternative proof that $t\mapsto\det \boldsymbol{M}_{\sqrt{t}X+N,n}$ is a polynomial. The term $\beta_{\pi;\boldsymbol{k}}$ is nonzero if and only if all the differences $r+\pi(r)-k_r$ are even. Suppose, for the sake of contradiction, that this is true for some fixed permutation $\pi\in\mathcal{S}_{[n]}$ and naturals $k_0,\cdots,k_n$ for which $k_0+\cdots+k_n$ is odd. Then, there is an odd number of odd numbers k_r. But, by Lemma 11, there is an even number of odd numbers $r+\pi(r)$. Therefore, there is an $r\in[n]$ for which $r+\pi(r)$ and k_r have different parities, contradicting evenness of $r+\pi(r)-k_r$.

Because the coefficients $b_X^{n,j}$ were defined by equality of polynomials $\det \boldsymbol{M}_{\sqrt{t}X+N,n} = \sum_{j\in[d_n]} b_X^{n,j} t^j$ (see (47)), we obtain that for each $j \in [d_n]$

$$b_X^{n,j} = \sum_{\substack{\pi\in\mathcal{S}_{[n]} \\ k_r\in[r+\pi(r)],\ \forall r\in[n] \\ k_0+\cdots+k_n=2j}} \beta_{\pi;\boldsymbol{k}} \mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n}. \tag{172}$$

The coefficient $a_X^{n,j}$ may be expanded similarly to obtain the following formula. Define the integers

$$\gamma_{i,\pi,\boldsymbol{k},w,z} = (-1)^{i+\pi(i)}\operatorname{sgn}(\pi)\binom{i}{w}\binom{\pi(i)}{z}\mathbb{E}[N^{i-w}]\times \mathbb{E}[N^{\pi(i)-z}]\prod_{r\in[n]\setminus\{i\}}\binom{r+\pi(r)}{k_r}\mathbb{E}[N^{r+\pi(r)-k_r}], \tag{173}$$

and the restricted sums

$$s_i(\boldsymbol{k}) = \sum_{r\in[n]\setminus\{i\}} k_r. \tag{174}$$

Then, we have the formula

$$a_X^{n,j} = \sum_{\substack{\pi\in\mathcal{S}_{[n]} \\ k_r\in[r+\pi(r)],\ \forall r\in[n] \\ k_0+\cdots+k_n=2j}} \beta_{\pi;k_0,\cdots,k_n}\mathcal{X}_2\mathcal{X}_{k_0}\cdots\mathcal{X}_{k_n} - \sum_{\substack{(i,\pi)\in[n]\times\mathcal{S}_{[n]} \\ (w,z)\in[i]\times[\pi(i)] \\ k_r\in[r+\pi(r)],\ \forall r\in[n]\setminus\{i\} \\ w+z+s_i(\boldsymbol{k})=2j}} \gamma_{i,\pi,\boldsymbol{k},w,z}\mathcal{X}_{w+1}\mathcal{X}_{z+1}\prod_{r\in[n]\setminus\{i\}}\mathcal{X}_{k_r}. \tag{175}$$

From the formulas for $a_X^{n,j}$ and $b_X^{n,j}$ in (175) and (172), we obtain the following restrictions on how they can contain any of the moments of X. We need to define the following set of homogeneous polynomials in the moments of X. We use the notation $\boldsymbol{\lambda} = (\lambda_1,\cdots,\lambda_m)^T \in \mathbb{N}^m$.

Definition 6. For $(\ell,m,k) \in \mathbb{N}^3$, let $\Pi_{\ell,m,k}$ denote the set of unordered partitions of ℓ into at most m parts each of which not exceeding k, i.e., $\Pi_{\ell,m,k} := \{\boldsymbol{\lambda}\in\mathbb{N}^m\ ;\ k\ge\lambda_1\ge\cdots\ge\lambda_m,\ \boldsymbol{\lambda}^T\mathbf{1}=\ell\}$. We define the set of homogeneous integer-coefficient polynomials with weighted-degree ℓ and width at most m in the first k moments $\mathcal{X}_1,\cdots,\mathcal{X}_k$ of X as

$$H_{\ell,m,k}(X) := \left\{ \sum_{\boldsymbol{\lambda}\in\Pi_{\ell,m,k}} c_{\boldsymbol{\lambda}} \prod_{i=1}^m \mathcal{X}_{\lambda_i}\ ;\ c_{\boldsymbol{\lambda}}\in\mathbb{Z}\right\}. \tag{176}$$

If $\Pi_{\ell,m,k}=\emptyset$, we set $H_{\ell,m,k}(X)=\mathbb{Z}$.

Remark 17. An element $q(X) \in H_{\ell,m,k}(X)$ will be an integer linear combination of terms $\prod_{i=1}^m \mathcal{X}_{\lambda_i}$. Each of these terms is a product of at most m of the moments of X (hence the terminology *width*). The highest moment that can appear is $\mathcal{X}_k$, because $\boldsymbol{\lambda}\in\Pi_{\ell,m,k}$. Suppose $\Pi_{\ell,m,k}\neq\emptyset$. Then, each summand shares the property that $\sum_{i=1}^m \lambda_i = \ell$. Further, looking at each $\mathcal{X}_j$ as an indeterminate of "degree" j, we may view $q(X)$ as a "homogeneous" polynomial in the moments of X of "degree" ℓ. In other words, for any constant c, $q(cX) = c^\ell q(X)$; in fact, this homogeneity holds for each term in the sum defining q, $\prod_{i=1}^m \mathbb{E}\left[(cX)^{\lambda_i}\right] = c^\ell \prod_{i=1}^m \mathbb{E}\left[X^{\lambda_i}\right]$.

Example 6. The partitions of the integer 6 into at most 3 parts each of which not exceeding 4 are given by $\Pi_{6,3,4} = \{(4,2,0),(4,1,1),(3,3,0),(3,2,1),(2,2,2)\}$. Note the resemblance between the elements of $\Pi_{6,3,4}$ and the terms appearing in the expression for $\det \boldsymbol{M}_{X,2}$, namely, (see (37))

$$\det \boldsymbol{M}_{X,2} = \mathcal{X}_4\mathcal{X}_2 - \mathcal{X}_4\mathcal{X}_1^2 - \mathcal{X}_3^2 + 2\mathcal{X}_3\mathcal{X}_2\mathcal{X}_1 - \mathcal{X}_2^3. \tag{177}$$

A term $\prod_{i=1}^3 \mathcal{X}_{\lambda_i}$ with $\lambda_1\ge\lambda_2\ge\lambda_3$ appears in $\det \boldsymbol{M}_{X,2}$ if and only if $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\lambda_3)$ is in $\Pi_{6,3,4}$. In particular, $\det \boldsymbol{M}_{X,2} \in H_{6,3,4}(X)$. Leibniz's formula for the determinant can be used to show that, in general, $\det \boldsymbol{M}_{X,n} \in H_{n(n+1),n+1,2n}(X)$.

From (175) and (172), we have that the constants $a_X^{n,j}$ and $b_X^{n,j}$ satisfy

$$a_X^{n,j} \in H_{2j+2,\,\min(n,2j)+2,\,\tau_n(j)}(X) \tag{178}$$

$$b_X^{n,j} \in H_{2j,\,\min(n+1,2j),\,2\min(n,j)}(X), \tag{179}$$

with $H_{\ell,m,k}(X)$ as given in Definition 6 and $\tau_n(j) \le 2\min(n,j+1)$ is defined by

$$\tau_n(j) = \begin{cases} 2 & \text{if } j=0, \\ 2j+1 & \text{if } 1\le j\le \frac{n}{2}, \\ 2j & \text{if } \frac{n+1}{2}\le j\le n, \\ 2n & \text{if } j>n. \end{cases} \tag{180}$$

C. Proof of Proposition 2

We proceed by induction on m. The case $m=1$ follows because then by assumption on p we have that $p(k)=0$ for every positive integer k as can be seen by taking $X\sim\mathcal{N}(k,1)$, but the only polynomial with infinitely many zeros is the zero polynomial. Now, assume that the statement of the proposition holds for every polynomial in $m-1$ variables, where $m\ge 2$.

Fix a polynomial p in m variables, and assume that $p|_{\mathcal{C}^m}=0$. Regarding p as a polynomial in one of the variables with coefficients being polynomials in the remaining $m-1$ variables, we may write

$$p(u_1,\cdots,u_m) = \sum_{j\in[d]} p_j(u_1,\cdots,u_{m-1})u_m^j, \tag{181}$$

for some polynomials $p_0,\cdots,p_d$ in $m-1$ variables, where d is the total degree of p. We show that $p=0$ identically by showing that each p_j vanishes on $\mathcal{C}_{m-1}$ and using the induction hypothesis to obtain $p_j=0$ identically.

Fix $\boldsymbol{\mu} = (\mu_1,\cdots,\mu_{m-1}) \in \mathcal{C}_{m-1}$. Let μ_m be a variable, and set $\ell = \lfloor m/2\rfloor$. We have that $\ell = (m-1)/2$ if m is odd, and $\ell = m/2$ if m is even. Set $\boldsymbol{H} = (\mu_{i+j})_{(i,j)\in[\ell]^2}$. If m is even, then $\det \boldsymbol{H} = \alpha\mu_m + \beta$ for some constants $\alpha,\beta\in\mathbb{R}$ determined by $\boldsymbol{\mu}$, with $\alpha = \det(\mu_{i+j})_{(i,j)\in[\ell-1]^2} > 0$. In the case m is even, we set $t = -\beta/\alpha$, and in the case m is odd, we set $t=0$. Then, $\boldsymbol{H}$ is positive definite whenever $\mu_m > t$.

For each integer $k\ge 1$ and real $\varepsilon>0$, Lemma 7 yields a random variable $X_{k,\varepsilon}\in\mathcal{R}_m$ satisfying

$$\delta_{k,\ell}(\varepsilon) := \mathbb{E}[X_{k,\varepsilon}^\ell] - \mu_\ell \in (-\varepsilon,\varepsilon) \tag{182}$$

for each $\ell \in \{1, \cdots, m-1\}$ and

$$\delta_{k,m}(\varepsilon) := \mathbb{E}[X_{k,\varepsilon}^m] - (t+k) \in (-\varepsilon, \varepsilon). \tag{183}$$

Then, by assumption on p, for every $\varepsilon > 0$ and $k \in \mathbb{N}_{\geq 1}$,

$$\sum_{j\in[d]} p_j \left(\boldsymbol{\mu} + (\delta_{k,\ell}(\varepsilon))_{1\leq \ell \leq m-1}\right)(t+k+\delta_{k,m}(\varepsilon))^j = 0. \tag{184}$$

Taking the limit $\varepsilon \to 0^+$, we deduce that

$$\sum_{j\in[d]} p_j(\mu_1, \cdots, \mu_{m-1})(t+k)^j = 0. \tag{185}$$

Considering the left-hand side in (185) as a univariate polynomial in k, and noting that the vanishing in (185) holds at infinitely many values of k, we deduce that

$$p_j(\mu_1, \cdots, \mu_{m-1}) = 0 \tag{186}$$

for every $j \in [d]$. This holds for every $(\mu_1, \cdots, \mu_{m-1}) \in \mathcal{C}_{m-1}$, i.e., the premise of the proposition applies to each p_j (namely, for every $X \in \mathcal{R}_{m-1}$ we have $p_j(\mathbb{E}[X], \cdots, \mathbb{E}[X^{m-1}]) = 0$). By the induction hypothesis, we obtain $p_j = 0$, as polynomials, for every $j \in [d]$. Therefore, $p = 0$, and the proof is complete.

Appendix C
Convergence of the PMMSE to the MMSE in Gaussian Channels (Theorem 3): Proofs of Section III-B

We derive in Appendix C-A the uniform convergence $\sup_{t\geq 0}\ \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t) \searrow 0$ stated in equation (13). Lemma 8 regarding Freud weights is derived in Appendix C-B, and the bound on the higher-order derivatives of the conditional expectation given in Lemma 9 is shown in Appendix C-C.

A. Uniform Convergence of PMMSE to MMSE (13)

We start the proof by obtaining from Theorem 1 pointwise convergence. Let $N \sim \mathcal{N}(0,1)$ be independent of X. The MGF of $\sqrt{t}X + N$ exists (it is the product of the MGFs of $\sqrt{t}X$ and N) and this implies that $\sqrt{t}X + N$ satisfies Carleman's condition [4, Sec. 4.2]. Hence, by Theorem 1, we have $\lim_{n\to\infty} \mathrm{pmmse}_n(X,t) = \mathrm{mmse}(X,t)$ pointwise for each fixed $t \geq 0$. Now, we show that the convergence is in fact uniform in t.

For each $n \in \mathbb{N}$ and $t \in [0,\infty)$, write $g_n(t) := \mathrm{pmmse}_n(X,t) - \mathrm{mmse}(X,t)$. We will show that

$$\lim_{n\to\infty} \sup_{t\in[0,\infty)} g_n(t) = 0. \tag{187}$$

The limit $\mathrm{pmmse}_n(X,t) \searrow \mathrm{mmse}(X,t)$ as $n \to \infty$ says that $g_n(t) \searrow 0$ as $n \to \infty$ for every fixed $t \geq 0$. In addition, the asymptotics given in Corollary 1 imply that for each fixed $n \in \mathbb{N}$, $g_n(t) \to 0$ as $t \to \infty$. Note that $\{g_n\}_{n\in\mathbb{N}}$ is a pointwise decreasing sequence of nonnegative functions. We finish the proof via Cantor's intersection theorem.

Fix $\varepsilon > 0$. For each $n \in \mathbb{N}$, let $C_{\varepsilon,n} = g_n^{-1}([\varepsilon,\infty))$, where g_n^{-1} denotes the set-theoretic inverse. As $\{g_n\}_{n\in\mathbb{N}}$ is decreasing, $C_{\varepsilon,1} \supseteq C_{\varepsilon,2} \supseteq \cdots$ is decreasing too. As each g_n is continuous, each $C_{\varepsilon,n}$ is closed. Further, $\lim_{t\to\infty} g_1(t) = 0$ implies that $C_{\varepsilon,1}$ is bounded, hence each $C_{\varepsilon,n}$ is bounded. Thus, each $C_{\varepsilon,n}$ is compact. But, the intersection $\bigcap_{n\in\mathbb{N}} C_{\varepsilon,n}$ is empty, for if t_0 were in the intersection then $\liminf_{n\to\infty} g_n(t_0) \geq \varepsilon$ violating that $\lim_{n\to\infty} g_n(t_0) = 0$. Hence, by Cantor's intersection theorem, it must be that the $C_{\varepsilon,n}$ are eventually empty, so there is an $m \in \mathbb{N}$ such that $\sup_{t\in[0,\infty)} g_n(t) \leq \varepsilon$ for every $n > m$. This is precisely the uniform convergence in (187), and the proof is complete.

B. Proof of Lemma 8

Write $Y = X + N$ and $p_Y = e^{-Q}$. To see that p_Y^s is a Freud weight, it suffices to show that p_Y is a Freud weight, since it can be easily seen that the conditions in Definition 2 hold for p_Y^s if they hold for p_Y. First, we note that $Q'(y)$ is equal to $\mathbb{E}[N \mid Y = y]$.

Lemma 12. *Fix a random variable X and let $Y = X + N$ where $N \sim \mathcal{N}(0,1)$ is independent of X. Writing $p_Y = e^{-Q}$, we have that $Q'(y) = \mathbb{E}[N \mid Y = y]$.*

Proof. We have that $p_Y(y) = \mathbb{E}[e^{-(y-X)^2/2}]/\sqrt{2\pi}$. Differentiating this equation, we obtain that $p_Y'(y) = \mathbb{E}[(X-y)e^{-(y-X)^2/2}]/\sqrt{2\pi}$, where the exchange of differentiation and integration is warranted since $t \mapsto te^{-t^2/2}$ is bounded. Now, $Q = -\log p_Y$, so $Q' = -p_Y'/p_Y$, i.e.,

$$Q'(y) = y - \frac{\mathbb{E}[Xe^{-(y-X)^2/2}]}{\mathbb{E}[e^{-(y-X)^2/2}]} = y - \mathbb{E}\left[X \mid Y = y\right]. \tag{188}$$

The proof is completed by substituting $X = Y - N$. □

In view of Lemma 12, that p is even and non-increasing over $[0,\infty) \cap \mathrm{supp}(p)$ imply that Q satisfies conditions (1)–(4) of Definition 2. It remains to show that property (5) holds. To this end, we show that if $\mathrm{supp}(p) \subset [-M, M]$ and $\lambda = M + 2$, then for every $y > M + 4$ we have that

$$1 < \frac{M^2+5M+8}{2(M+2)} \leq \frac{Q'(\lambda y)}{Q'(y)} \leq \frac{M^2+7M+8}{4}. \tag{189}$$

First, since $Q'(y) = y - \mathbb{E}[X \mid Y = y]$ (see (188)), we have the bounds $y - M \leq Q'(y) \leq y + M$ for every $y \in \mathbb{R}$. Therefore, $y > M$ and $\lambda > 1$ imply that

$$\frac{\lambda y - M}{y + M} \leq \frac{Q'(\lambda y)}{Q'(y)} \leq \frac{\lambda y + M}{y - M}. \tag{190}$$

Further, since $y > M + 4$ and $\lambda = M + 2$, we have

$$\frac{M^2+5M+8}{2(M+2)} < \lambda - \frac{M(M+3)}{y+M} = \frac{\lambda y - M}{y+M} \tag{191}$$

and

$$\frac{\lambda y + M}{y - M} = \lambda + \frac{M(M+3)}{y-M} \leq \frac{M^2+7M+8}{4}. \tag{192}$$

The fact that $1 < \frac{M^2+5M+8}{2(M+2)}$ follows since the discriminant of $M^2 + 3M + 4$ is $-7 < 0$. Therefore, p_Y is a Freud weight.

Next, we derive the bound on $a_n(sQ)$ stated in (58). By definition of a_n, we have that $a_n(sQ) = a_{n/s}(Q)$. Thus, it suffices to show $a_n(Q) \le (2M+\sqrt{2})\sqrt{n}$. By Lemma 12,

$$Q'(y) = \mathbb{E}[N \mid Y = y] = y - \mathbb{E}[X \mid Y = y]. \tag{193}$$

Therefore $X \le M$ implies that, for any constant $z \ge 0$, we have

$$\int_0^1 \frac{ztQ'(zt)}{\sqrt{1-t^2}}\,dt = \frac{\pi}{4}z^2 - z\int_0^1 \frac{t}{\sqrt{1-t^2}} \frac{\mathbb{E}\left[Xe^{-(X-zt)^2/2}\right]}{\mathbb{E}\left[e^{-(X-zt)^2/2}\right]}dt \tag{194}$$

$$\ge \frac{\pi}{4}z^2 - Mz. \tag{195}$$

We have $\pi z^2/4 - Mz > n$ for $z = (2M+\sqrt{2})\sqrt{n}$. Since $y \mapsto yQ'(y)$ is strictly increasing over $(0,\infty)$ (condition (3) of Definition 2), we conclude that $a_n(Q) \le (2M+\sqrt{2})\sqrt{n}$. This completes the proof of Lemma 8.

C. Proof of Lemma 9

We use the formula of the conditional expectation derivative given in Proposition 3, with the conditional cumulant being expanded in terms of conditional moments using Bell polynomials, then apply Hölder's inequality to each ensuing summand. We use the following notation. The set of all finite-length tuples of non-negative integers is denoted by $\mathbb{N}^*$. For every integer $r \ge 2$, let Π_r be the set of unordered integer partitions $r = r_1 + \cdots + r_k$ of r into integers $r_j \ge 2$. We encode Π_r via a list of the multiplicities of the parts as

$$\Pi_r := \{(\lambda_2, \cdots, \lambda_\ell) \in \mathbb{N}^* \ ; \ 2\lambda_2 + \cdots + \ell\lambda_\ell = r\}. \tag{196}$$

In (196), $\ell \ge 2$ is free, and trailing zeros are ignored (i.e., $\lambda_\ell > 0$). For a partition $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$ having $m = \lambda_2 + \cdots + \lambda_\ell$ parts, we denote

$$c_{\boldsymbol{\lambda}} := \frac{1}{m}\binom{m}{\lambda_2, \cdots, \lambda_\ell}\binom{r}{\underbrace{2, \cdots, 2}_{\lambda_2}; \cdots ; \underbrace{\ell, \cdots, \ell}_{\lambda_\ell}} \tag{197}$$

and $e_{\boldsymbol{\lambda}} := (-1)^{m-1}c_{\boldsymbol{\lambda}}$. Set $C'_r := \sum_{\boldsymbol{\lambda}\in\Pi_r} c_{\boldsymbol{\lambda}}$. For each $(y,k) \in \mathbb{R}\times\mathbb{N}$, denote $f(y) := \mathbb{E}[X \mid Y = y]$ and

$$g_k(y) := \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^k \mid Y = y\right]. \tag{198}$$

For $\ell \ge 2$ and $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \mathbb{N}^{\ell-1}$, denote $\boldsymbol{g}^{\boldsymbol{\lambda}} := \prod_{i=2}^{\ell} g_i^{\lambda_i}$, with the understanding that $g_i^0 = 1$. Using Proposition 3, and expanding $\kappa_r(X \mid Y = y)$ in terms of the conditional moments $\mathbb{E}[X^k \mid Y = y]$, we obtain (see [2, Proposition 1])

$$f^{(r-1)} = \sum_{\boldsymbol{\lambda}\in\Pi_r} e_{\boldsymbol{\lambda}}\boldsymbol{g}^{\boldsymbol{\lambda}}. \tag{199}$$

Fix $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$. By the generalization of Hölder's inequality stating $\|\psi_1 \cdots \psi_k\|_1 \le \prod_{i=1}^k \|\psi_i\|_k$, we have that

$$\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2^2 = \left\|\prod_{\lambda_i\neq 0} g_i^{2\lambda_i}(Y)\right\|_1 \le \prod_{\lambda_i\neq 0}\left\|g_i^{2\lambda_i}(Y)\right\|_s \tag{200}$$

where s is the number of nonzero entries in $\boldsymbol{\lambda}$. By Jensen's inequality for conditional expectation, for each i such that $\lambda_i \neq 0$, we have that

$$\left\|g_i^{2\lambda_i}(Y)\right\|_s \le \|X - \mathbb{E}[X \mid Y]\|_{2i\lambda_i}^{2i\lambda_i}. \tag{201}$$

Now, $r = \sum_{i=2}^{\ell} i\lambda_i \ge \sum_{i=2}^{s+1} i = \frac{(s+1)(s+2)}{2} - 1$, so we have that $s^2 + 3s - 2r \le 0$, i.e., $s \le q_r$. Further, $i\lambda_i \le r$ for each i. Hence, monotonicity of norms and inequalities (200) and (201) imply the uniform (in $\boldsymbol{\lambda}$) bound

$$\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2 \le \|X - \mathbb{E}[X \mid Y]\|_{2rq_r}^r. \tag{202}$$

Observe that $\|X - \mathbb{E}[X \mid Y]\|_k \le 2\min\left((k!)^{1/(2k)}, \|X\|_k\right)$ (see [12]). Therefore, applying the triangle inequality in (199) we obtain

$$\left\|f^{(r-1)}(Y)\right\|_2 \le \sum_{\boldsymbol{\lambda}\in\Pi_r} c_{\boldsymbol{\lambda}}\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2 \tag{203}$$

$$\le 2^r C'_r \ \min\left(\gamma_r, \|X\|_{2rq_r}^r\right), \tag{204}$$

where $\gamma_r = (2rq_r)!^{1/(4q_r)}$.

It only remains then to note that $C'_r = C_r$. The integer $c_{\boldsymbol{\lambda}}$ (as defined in (197)) can be easily seen to be equal to the number of cyclically-invariant ordered set-partitions of an r-element set into $m = \lambda_2 + \cdots + \lambda_\ell$ subsets where, for each $k \in \{2, \cdots, \ell\}$, exactly λ_k parts have size k. Hence, the integer C'_r equals the total number of cyclically-invariant ordered set-partitions of an r-element set into subsets of sizes at least 2, which is given by sequence A032181 at [49]. The formula for C'_r stated in [49] coincides with our definition of C_r in (60) in the statement of the lemma, from which we obtain $C'_r = C_r$. Finally, since the formula in [49] is stated without proof, we provide a proof here for completeness. Using the notation of [61], we have that

$$C'_r = \sum_{k=1}^{r}(k-1)!\left\{{r \atop k}\right\}_{\ge 2} \tag{205}$$

where $\left\{{r \atop k}\right\}_{\ge 2}$ denotes the number of partitions of an r-element set into k subsets each of which contains at least 2 elements (note that there are $(k-1)!$ cyclically-invariant arrangements of k parts). The exponential generating function of the sequence $r \mapsto \left\{{r \atop k}\right\}_{\ge 2}$ is $(e^x - 1 - x)^k/k!$. Now, we may write

$$(e^x - 1 - x)^k = \sum_{a+b\le k}\binom{k}{a,b}(-1)^{k-a}x^b\sum_{t\in\mathbb{N}}\frac{(ax)^t}{t!}. \tag{206}$$

Therefore, the coefficient of x^r in $(e^x - 1 - x)^k/k!$ is

$$\frac{1}{r!}\left\{{r \atop k}\right\}_{\ge 2} = \sum_{a+b\le k}\frac{(-1)^{k-a}a^{r-b}}{a!b!(k-a-b)!(r-b)!} \tag{207}$$

$$= \frac{1}{r!}\sum_{b=0}^{k}\binom{r}{b}\sum_{a=0}^{k-b}(-1)^{k-a}\frac{a^{r-b}}{a!(k-a-b)!} \tag{208}$$

$$= \frac{1}{r!}\sum_{b=0}^{k}\binom{r}{b}\left\{{r-b \atop k-b}\right\}(-1)^b, \tag{209}$$

which when combined with (205) gives $C'_r = C_r$ in view of (60). This completes the proof of the lemma.

Remark 18. A closer analysis reveals that $i\lambda_i s$ in (201) cannot exceed $\beta_r := t_r^2(t_r+1/2)$ where $t_r := (\sqrt{6r+7}-1)/3$. For $r \to \infty$, we have $rq_r/\beta_r \sim 3^{3/2}/2 \approx 2.6$. The reduction when, e.g., $r=7$, is from $rq_r = 14$ to $\beta_r = 10$.

APPENDIX D
GENERALIZATIONS TO ARBITRARY BASES AND MULTIPLE DIMENSIONS

We extend our approximation results for the conditional expectation from the polynomial-basis setting to arbitrary bases, and from conditioning on random variables to conditioning on arbitrary σ-algebras. An extension to the multidimensional case is also presented, which straightforwardly yields an approximation theorem for differential entropy of random vectors. Another byproduct of the multidimensional generalization is the expression for mutual information between two continuous random variables completely in terms of moments, as given in Theorem 5.

A. Arbitrary Bases and σ-Algebras

Up to here, our exposition dealt with the polynomial basis of $L^2(P_Y)$. However, our results can be extended to a more general setup. Recall that we have defined

$$\boldsymbol{M}_{Y,n} = \mathbb{E}\left[\boldsymbol{Y}^{(n)}\left(\boldsymbol{Y}^{(n)}\right)^T\right], \tag{210}$$

and derived

$$\mathbb{E}[X \mid Y] = \lim_{n\to\infty} \mathbb{E}\left[X\boldsymbol{Y}^{(n)}\right] \boldsymbol{M}_{Y,n}^{-1}\boldsymbol{Y}^{(n)} \tag{211}$$

in Theorem 1 under two requirements: Y satisfies Carleman's condition, and $|\mathrm{supp}(Y)| = \infty$. Along similar lines, we derive a generalization where the set of polynomials of Y is replaced with any linearly-independent subset of $L^2(\Sigma)$ having a dense span, where $\Sigma \subset \mathcal{F}$ is any σ-algebra, and $L^2(\Sigma)$ denote the subset of $L^2(P)$ consisting of Σ-measurable random variables. Denseness replaces Carleman's condition, while linear independence replaces the infinite-support requirement.

Theorem 10. *Fix a σ-algebra $\Sigma \subset \mathcal{F}$ and a set $\{\psi_k\}_{k\in\mathbb{N}} = \mathcal{S} \subset L^2(\Sigma)$. For each $n \in \mathbb{N}$, define the random vector $\boldsymbol{\psi}^{(n)} = (\psi_0, \cdots, \psi_n)^T$ and the matrix of inner products*

$$\boldsymbol{M}_{\mathcal{S},n} := \mathbb{E}\left[\boldsymbol{\psi}^{(n)}\left(\boldsymbol{\psi}^{(n)}\right)^T\right]. \tag{212}$$

If $\mathcal{S}$ is linearly independent and $\mathrm{span}(\mathcal{S})$ is dense in $L^2(\Sigma)$, then

$$\mathbb{E}[X \mid \Sigma] = \lim_{n\to\infty} \mathbb{E}\left[X\boldsymbol{\psi}^{(n)}\right]^T \boldsymbol{M}_{\mathcal{S},n}^{-1}\,\boldsymbol{\psi}^{(n)} \tag{213}$$

in $L^2(\Sigma)$ for any random variable $X \in L^2(P)$.

For the proof of Theorem 10, we will need the following formula for the closest element in a finite-dimensional subspace of $L^2(P)$ to a random variable $X \in L^2(P)$, which will also be used for the extension of our results to random vectors later in this appendix. The following formula is simply an instantiation of the fact that, in a separable Hilbert space, the orthogonal projection onto a closed subspace is the unique closest element.

Lemma 13. *For any fixed finite-dimensional subspace $\mathcal{V} \subset L^2(P)$ having a basis $\{V_0, V_1, \cdots, V_n\}$, denoting $\boldsymbol{V} = (V_0, V_1, \cdots, V_n)^T$, we have that for every $X \in L^2(P)$*

$$\mathbb{E}\left[X\boldsymbol{V}\right]^T \mathbb{E}\left[\boldsymbol{V}\boldsymbol{V}^T\right]^{-1}\boldsymbol{V} = \operatorname*{argmin}_{V\in\mathcal{V}} \|X - V\|_2 . \tag{214}$$

In view of Lemma 13, we introduce the following notation.

Definition 7. Fix a random variable $X \in L^2(P)$, a σ-algebra $\Sigma \subset \mathcal{F}$, and a linearly-independent set $\{\theta_j\}_{j\in\mathbb{N}} = \Theta \subset L^2(\Sigma)$. Write $\boldsymbol{\theta}^{(n)} = (\theta_0, \cdots, \theta_n)^T$ for each $n \in \mathbb{N}$. We define the n-th approximation of $\mathbb{E}[X \mid \Sigma]$ with respect to Θ by

$$E_{n,\Theta}\left[X \mid \Sigma\right] := \mathbb{E}\left[X\boldsymbol{\theta}^{(n)}\right] \mathbb{E}\left[\boldsymbol{\theta}^{(n)}\left(\boldsymbol{\theta}^{(n)}\right)^T\right]^{-1}\boldsymbol{\theta}^{(n)}. \tag{215}$$

Note that $E_{n,\Theta}[X \mid \Sigma]$ belongs to $\mathrm{span}(\{\theta_j\}_{j\in[n]})$. Further, according to Lemma 13, $E_{n,\Theta}[X \mid \Sigma]$ is the unique closest element in $\mathrm{span}(\{\theta_j\}_{j\in[n]})$ to X,

$$E_{n,\Theta}[X \mid \Sigma] = \operatorname*{argmin}_{V \,\in\, \mathrm{span}(\{\theta_j\}_{j\in[n]})} \|X - V\|_2 . \tag{216}$$

If $Y \in L^{2n}(P)$, $\Theta = \{Y^j\}_{j\in\mathbb{N}}$, and $\Sigma = \sigma(Y)$, then the estimate reduces to $E_{n,\Theta}[X \mid \Sigma] = E_n[X \mid Y]$.

The central claim in Theorem 10 is that if $\mathrm{span}(\Theta)$ is dense in $L^2(\Sigma)$ then we have the limit

$$\mathbb{E}[X \mid \Sigma] = \lim_{n\to\infty} E_{n,\Theta}[X \mid \Sigma]. \tag{217}$$

The proof of Theorem 1 can be adapted *mutatis mutandis* to derive the above limit, so we omit the details.

B. The Multidimensional PMMSE

We extend our results on the PMMSE of random variables to random vectors. We begin with some notation. The Hilbert space of q-integrable m-dimensional random vectors is denoted by $L^q(\mathbb{R}^m, P)$, with norm also denoted by $\|\cdot\|_q$. The subspace of Σ-measurable random vectors is denoted by $L^q(\mathbb{R}^m, \Sigma)$. We keep the notations $L^q(\mathbb{R}, \Sigma) = L^q(\Sigma)$ and $L^q(\mathbb{R}, P_Y) = L^q(P_Y)$. By a generalization of Hölder's inequality, for any $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T \in L^\beta(\mathbb{R}^m, P)$, we also have that $Y_1^{\alpha_1}\cdots Y_m^{\alpha_m} \in L^1(P)$ for any constants $\alpha_1, \cdots, \alpha_m \geq 0$ such that $\alpha_1 + \cdots + \alpha_m \leq \beta$.

We extend the notation $\boldsymbol{Y}^{(n)}$ to random vectors as follows. For an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$, we let $\boldsymbol{Y}^{(n,m)}$ denote the random vector whose entries are monomials in the Y_j of total degree at most n, ordered first by total degree then reverse-lexicographically in the exponents. For example, if $m = 3$ so $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$, then for $n = 2$

$$\boldsymbol{Y}^{(2,3)} = (1, Y_1, Y_2, Y_3, Y_1^2, Y_1Y_2, Y_1Y_3, Y_2^2, Y_2Y_3, Y_3^2)^T \tag{218}$$

because we are taking the totally ordered set of exponents $(\ \{\boldsymbol{v} \in \mathbb{N}^3 \mid \mathbf{1}^T\boldsymbol{v} \leq 2\}\ ,\ <\ )$ to have the order[13]

$$(0,0,0) < (1,0,0) < (0,1,0) < (0,0,1) < (2,0,0) < (1,1,0) < (1,0,1) < (0,2,0) < (0,1,1) < (0,0,2).$$

[13]Note that this ordering is not the same as the degree reverse lexicographical order nor its reverse.

A straightforward stars-and-bars counting argument reveals that the length of $\boldsymbol{Y}^{(n,m)}$ is $\binom{n+m}{m}$.

Let $\mathscr{P}_{n,m}$ denote the set of polynomials in m variables with real coefficients of total degree at most n. For a fixed m-dimensional random vector $\boldsymbol{Y}$, denote $\mathscr{P}_{n,m}(\boldsymbol{Y}) := \{p(\boldsymbol{Y}) \; ; \; p \in \mathscr{P}_{n,m}\}$. Note that $\mathscr{P}_{n,1} = \mathscr{P}_n$. Also, the notation $\boldsymbol{Y}^{(n,1)}$, while avoided, is disambiguated by interpreting it as $\boldsymbol{Y}^{(n)}$, i.e., $\boldsymbol{Y}^{(n,1)} = (1, Y, \cdots, Y^n)^T$ where the subscript on Y_1 is dropped. We denote the product sets of $\mathscr{P}_{n,m}(\boldsymbol{Y})$ by $\mathscr{P}^{\ell}_{n,m}(\boldsymbol{Y})$, and consider their elements as vectors rather than tuples. In other words, we denote the set of length-ℓ vectors whose coordinates are multivariate polynomial expressions of an m-dimensional random vector $\boldsymbol{Y}$ with total degree at most n by

$$\mathscr{P}^{\ell}_{n,m}(\boldsymbol{Y}) = \left\{(p_1(\boldsymbol{Y}), \cdots, p_\ell(\boldsymbol{Y}))^T \; ; \; p_1, \cdots, p_\ell \in \mathscr{P}_{n,m}\right\}. \tag{219}$$

The multivariate generalization of the PMMSE is defined as follows.

Definition 8 (Multivariate Polynomial MMSE)**.** Fix positive integer ℓ, m, and n. Fix an ℓ-dimensional random vector $\boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$ and an m-dimensional random vector $\boldsymbol{Y} \in L^{2n}(\mathbb{R}^m, P)$, and set $k = \binom{n+m}{m}$. We define the n-th order PMMSE for estimating $\boldsymbol{X}$ given $\boldsymbol{Y}$ by

$$\operatorname{pmmse}_n(\boldsymbol{X} \mid \boldsymbol{Y}) := \min_{\boldsymbol{C} \in \mathbb{R}^{\ell \times k}} \left\| \boldsymbol{X} - \boldsymbol{C}\boldsymbol{Y}^{(n,m)} \right\|_2^2, \tag{220}$$

and the n-th order PMMSE estimate of $\boldsymbol{X}$ given $\boldsymbol{Y}$ by

$$E_n[\boldsymbol{X} \mid \boldsymbol{Y}] := \boldsymbol{C}\boldsymbol{Y}^{(n,m)} \in \mathscr{P}^{\ell}_{n,m}(\boldsymbol{Y}) \tag{221}$$

for any minimizing matrix $\boldsymbol{C} \in \mathbb{R}^{\ell \times k}$ in (220).

Remark 19. For any minimizer $\boldsymbol{C}$ in (220), the ℓ-dimensional random vector $\boldsymbol{C}\boldsymbol{Y}^{(n,m)}$ is the unique orthogonal projection of $\boldsymbol{X}$ onto $\mathscr{P}^{\ell}_{n,m}(\boldsymbol{Y})$; in particular, $E_n[\boldsymbol{X} \mid \boldsymbol{Y}]$ is well-defined by (221).

Denote, for $\boldsymbol{Y} \in L^{2n}(\mathbb{R}^m, P)$,

$$\boldsymbol{M}_{\boldsymbol{Y},n} := \mathbb{E}\left[\boldsymbol{Y}^{(n,m)} \left(\boldsymbol{Y}^{(n,m)}\right)^T\right]. \tag{222}$$

For $n \in \mathbb{N}$ and an ℓ-dimensional random vector $(X_1, \cdots, X_\ell)^T = \boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$, if $\boldsymbol{M}_{\boldsymbol{Y},n}$ is invertible, Lemma 13 yields that

$$E_n[\boldsymbol{X} \mid \boldsymbol{Y}] = \begin{pmatrix} E_n[X_1 \mid \boldsymbol{Y}] \\ \vdots \\ E_n[X_\ell \mid \boldsymbol{Y}] \end{pmatrix} \tag{223}$$

$$= \begin{pmatrix} \mathbb{E}\left[X_1 \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \\ \vdots \\ \mathbb{E}\left[X_\ell \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \end{pmatrix}. \tag{224}$$

We say that the Y_j do not satisfy a polynomial relation if the monomials $\prod_{j=1}^m Y_j^{\alpha_j}$, for $\alpha_1, \cdots, \alpha_m \in \mathbb{N}$, are linearly independent, i.e., if the mapping

$$\varphi : \bigcup_{n \in \mathbb{N}} \mathscr{P}_{n,m} \to \bigcup_{n \in \mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y}), \qquad \varphi(p) = p(\boldsymbol{Y}) \tag{225}$$

is an isomorphism of vector spaces.

Generalizing our results on random variables to random vectors can be done in view of the following polynomial denseness result.

Theorem 11 ([62])**.** *For any m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$ and $q > 1$, if we have the denseness $\bigcup_{n \in \mathbb{N}} \mathscr{P}_n(Y_j) = L^q(P_{Y_j})$ for each $j \in \{1, \cdots, m\}$, then we have the denseness $\overline{\bigcup_{n \in \mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y})} = L^r(P_{\boldsymbol{Y}})$ for every $r \in [1, q)$.*

Since $\|\boldsymbol{Z}\|_r^r = \sum_j \|Z_j\|_r^r$, this inferred denseness in Theorem 11 over $L^r(P_{\boldsymbol{Y}})$ may be extended to denseness over $L^r(\mathbb{R}^m, P_{\boldsymbol{Y}})$, i.e., we have the following immediate corollary.

Corollary 5. *Fix an integer $m \geq 1$ and an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$. If each of the random variables $Y_1, \cdots, Y_m$ satisfies Carleman's condition, then the set of vectors of polynomials $\bigcup_{n \in \mathbb{N}} \mathscr{P}^m_{n,m}(\boldsymbol{Y})$ is dense in $L^q(\mathbb{R}^m, P_{\boldsymbol{Y}})$ for any $q \geq 1$.*

We deduce the following result on the convergence of the multivariate PMMSE to the MMSE.

Theorem 12. *Fix an m-dimensional random vector $\boldsymbol{Y} = (Y_1, \cdots, Y_m)^T$ and an ℓ-dimensional random vector $\boldsymbol{X} \in L^2(\mathbb{R}^\ell, P)$. If each Y_j satisfies Carleman's condition, and if the Y_j do not satisfy a polynomial relation, then we have the $L^2(\mathbb{R}^\ell, P_{\boldsymbol{Y}})$-limit*

$$\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}] = \lim_{n \to \infty} E_n[\boldsymbol{X} \mid \boldsymbol{Y}]. \tag{226}$$

Proof. Since the Y_j do not satisfy a polynomial relation, the matrix $\boldsymbol{M}_{\boldsymbol{Y},n}$ is invertible for each $n \in \mathbb{N}$. Further, the entries of $\boldsymbol{Y}^{(n,m)}$ are linearly independent for each n. Then, by Lemma 13, equation (224) follows, i.e., the PMMSE estimate $E_n[\boldsymbol{X} \mid \boldsymbol{Y}]$ is the ℓ-dimensional random vector whose k-th entry is $\mathbb{E}\left[X_k \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)}$. By Corollary 5, since each Y_j satisfies Carleman's condition, the set of vectors of polynomials $\bigcup_{n \in \mathbb{N}} \mathscr{P}^m_{n,m}(\boldsymbol{Y})$ is dense in $L^2(\mathbb{R}^m, P_{\boldsymbol{Y}})$. In particular, $\bigcup_{n \in \mathbb{N}} \mathscr{P}_{n,m}(\boldsymbol{Y})$ is dense in $L^2(P_{\boldsymbol{Y}})$. By Theorem 10, we have the $L^2(P_{\boldsymbol{Y}})$ limits

$$\mathbb{E}[X_k \mid \boldsymbol{Y}] = \lim_{n \to \infty} \mathbb{E}\left[X_k \boldsymbol{Y}^{(n,m)}\right]^T \boldsymbol{M}_{\boldsymbol{Y},n}^{-1} \boldsymbol{Y}^{(n,m)} \tag{227}$$

for each $k \in \{1, \cdots, \ell\}$. We conclude that $E_n[\boldsymbol{X} \mid \boldsymbol{Y}] \to \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}]$ in $L^2(\mathbb{R}^\ell, P_{\boldsymbol{Y}})$, as desired. □

The approach for showing the rationality of $t \mapsto \operatorname{pmmse}_n(X, t)$ for a random variable $X \in L^{2n}(P)$ in Theorem 2 may be generalized to deduce rationality of $t \mapsto \operatorname{pmmse}_n(\boldsymbol{X}, t)$ for an m-dimensional random vector $\boldsymbol{X} \in L^{2n}(\mathbb{R}^m, P)$. Here, we are denoting $\operatorname{pmmse}_n(\boldsymbol{X}, t) := \operatorname{pmmse}_n(\boldsymbol{X} \mid \sqrt{t}\boldsymbol{X} + \boldsymbol{N})$, where $\boldsymbol{N} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ is independent of $\boldsymbol{X}$. For brevity, we give a blueprint of how this generalization of rationality can be obtained. First, Lemma 6

may be generalized to yield that $\det \boldsymbol{M}_{\sqrt{t}\boldsymbol{X}+\boldsymbol{N}}$ is a polynomial in t of degree at most $d_{n,m}$ which is given by

$$d_{n,m} := \sum_{k\in[n]} k \cdot |\{(\lambda_1,\cdots,\lambda_m)\in\mathbb{N}^m\,;\,\lambda_1+\cdots+\lambda_m=k\}| \tag{228}$$

$$= \sum_{k\in[n]} k\binom{k+m-1}{m-1} \tag{229}$$

$$= \sum_{k\in[n]} m\binom{k+m-1}{m} = m\binom{n+m}{m+1}. \tag{230}$$

Further, the coefficient of $t^{d_{n,m}}$ in $\det \boldsymbol{M}_{\sqrt{t}\boldsymbol{X}+\boldsymbol{N}}$ is $\det \boldsymbol{M}_{\boldsymbol{X}}$. Note that $d_{n,1} = d_n$. Then, the PMMSE expression in Theorem 2 may be generalized to give

$$\mathrm{pmmse}_n(\boldsymbol{X},t) = \frac{(\operatorname{tr}\boldsymbol{\Sigma}_{\boldsymbol{X}})\det\boldsymbol{M}_{\boldsymbol{N},n}+\cdots+(\operatorname{tr}\boldsymbol{\Sigma}_{\boldsymbol{N}})(\det\boldsymbol{M}_{\boldsymbol{X},n})\ t^{d_{n,m}-1}}{\det\boldsymbol{M}_{\boldsymbol{N},n}+\cdots+(\det\boldsymbol{M}_{\boldsymbol{X},n})\ t^{d_{n,m}}}. \tag{231}$$

To deduce (231), the multidimensional MMSE dimension result in Theorem 6 is used, as follows. Note that $\operatorname{tr}\boldsymbol{\Sigma}_{\boldsymbol{N}} = m$ for $\boldsymbol{N}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)$. By Theorem 6, we have that $\mathrm{mmse}(\boldsymbol{X},t)\sim m/t$. It is also true that $\mathrm{lmmse}(\boldsymbol{X},t)\sim m/t$. Therefore, $\mathrm{pmmse}_n(\boldsymbol{X},t)\sim m/t$ for every integer $n\geq 1$. Note that $\mathrm{pmmse}_n(\boldsymbol{X},0) = \operatorname{tr}\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Expression (231) follows via the same proof technique for Theorem 2.

With the definition of the multivariate PMMSE at hand, we show that the PMMSE estimate satisfies a tower property similar to the conditional expectation.

Proposition 6 (Tower Property)**.** *Fix $n\in\mathbb{N}$ and three random variables $X\in L^2(P)$ and $Y_1,Y_2\in L^{2n}(P)$. Suppose that $|\mathrm{supp}(Y_1)|,|\mathrm{supp}(Y_2)|>n$. Then*

$$E_n\left[E_n[X\mid Y_1]\mid Y_1,Y_2\right] = E_n[X\mid Y_1], \tag{232}$$

and

$$E_n\left[E_n[X\mid Y_1,Y_2]\mid Y_2\right] = E_n[X\mid Y_2]. \tag{233}$$

Proof. Set $\boldsymbol{Y} = (Y_1,Y_2)^T$. Equation (232) is straightforward: since $E_n[X\mid Y_1]\in\mathscr{P}_n(Y_1)\subset\mathscr{P}_{n,2}(\boldsymbol{Y})$, the projection of $E_n[X\mid Y_1]$ onto $\mathscr{P}_{n,2}(\boldsymbol{Y})$ is $E_n[X\mid Y_1]$ again. Equation (233) also follows by an orthogonal projection argument. There is a unique representation $X = p_{1,2}+p_{1,2}^{\perp}$ for $(p_{1,2},p_{1,2}^{\perp})\in\mathscr{P}_{n,2}(\boldsymbol{Y})\times\mathscr{P}_{n,2}(\boldsymbol{Y})^{\perp}$. There is also a unique representation $p_{1,2} = q_2+q_2^{\perp}$ for $(q_2,q_2^{\perp})\in\mathscr{P}_n(Y_2)\times\mathscr{P}_n(Y_2)^{\perp}$. The projection of X onto $\mathscr{P}_{n,2}(\boldsymbol{Y})$ is $p_{1,2}$, whose projection onto $\mathscr{P}_n(Y_2)$ is q_2, i.e.,

$$E_n\left[E_n[X\mid Y_1,Y_2]\mid Y_2\right] = q_2. \tag{234}$$

Furthermore, we have the representation $X = q_2+(q_2^{\perp}+p_{1,2}^{\perp})$, for which $(q_2,q_2^{\perp}+p_{1,2}^{\perp})\in\mathscr{P}_n(Y_2)\times\mathscr{P}_n(Y_2)^{\perp}$. Hence, the projection of X onto $\mathscr{P}_n(Y_2)$ is q_2 too, i.e.,

$$E_n[X\mid Y_2] = q_2. \tag{235}$$

From (234) and (235) we get (233). Equation (233) can also be deduced from the formula of $W := \mathbb{E}[X\mid\boldsymbol{Y}]$. Denote $\boldsymbol{Y}_2^{(n)} = (1,Y_2,\cdots,Y_2^n)^T$. We have that

$$W = \mathbb{E}\left[X\boldsymbol{Y}^{(n,2)}\right]^T\boldsymbol{M}_{\boldsymbol{Y},n}^{-1}\boldsymbol{Y}^{(n,2)} \tag{236}$$

and

$$E_n[W\mid Y_2] = \mathbb{E}\left[W\boldsymbol{Y}_2^{(n)}\right]^T\boldsymbol{M}_{Y_2,n}^{-1}\boldsymbol{Y}_2^{(n)}. \tag{237}$$

For $k\in[n]$, let $\delta(k)\in\left[\binom{n+2}{2}-1\right]$ be the index of the entry in $\boldsymbol{Y}^{(n,2)}$ that equals Y_2^k. Then,

$$\mathbb{E}\left[Y_2^k\boldsymbol{Y}^{(n,2)}\right] = \boldsymbol{M}_{\boldsymbol{Y},n}\boldsymbol{e}_{\delta(k)}, \tag{238}$$

where $\boldsymbol{e}_0,\cdots,\boldsymbol{e}_{\binom{n+2}{2}-1}$ are the standard basis vectors of $\mathbb{R}^{\binom{n+2}{2}}$. Therefore, plugging (236) into (237), we obtain

$$E_n[W\mid Y_2] = \mathbb{E}\left[X\boldsymbol{Y}_2^{(n)}\right]^T\boldsymbol{M}_{Y_2,n}^{-1}\boldsymbol{Y}_2^{(n)}, \tag{239}$$

which is just $E_n[X\mid Y_2]$, as desired. □

Appendix E
Information Measures in Terms of Moments: Proofs of Section IV

A. Proof of Lemma 10

By finiteness of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, we get that $h(\boldsymbol{X})$ is well defined and less that ∞, but it could be $-\infty$. First, the case that $\det\boldsymbol{\Sigma}_{\boldsymbol{X}} = 0$ follows since both sides of (76) would then equal $-\infty$, which can be seen as follows. That $h(\boldsymbol{X}) = -\infty$ follows by a limiting argument starting from $0\leq D_{\mathrm{kl}}\left(P_{\boldsymbol{X}}\|\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}_{\boldsymbol{X}}+\varepsilon\boldsymbol{I}_m)\right)$, and inferring that $h(\boldsymbol{X})\leq\frac{1}{2}\log\left((2\pi)^m\det\left(\boldsymbol{\Sigma}_{\boldsymbol{X}}+\varepsilon\boldsymbol{I}_m\right)\right)+\frac{1}{2}\mathrm{rank}(\boldsymbol{\Sigma}_{\boldsymbol{X}})$ for all $\varepsilon>0$, then taking $\varepsilon\to 0^+$. That the right-hand side of (76) equals $-\infty$ follows from $\mathrm{mmse}(\boldsymbol{X},t)\leq\mathrm{lmmse}(\boldsymbol{X},t)$ and $\mathrm{lmmse}(\boldsymbol{X},t)\sim\frac{\mathrm{rank}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}{t}$. So, we may assume $\det\boldsymbol{\Sigma}_{\boldsymbol{X}}\neq 0$.

In the same way that (74) is derived in [1] (see Lemma 7 and Theorem 14 therein), one may obtain

$$h(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^m\det\boldsymbol{\Sigma}_{\boldsymbol{X}}\right) - \frac{1}{2}\lim_{\gamma\to\infty}\left[\log\left(\det\left(\gamma\boldsymbol{\Sigma}_{\boldsymbol{X}}+\boldsymbol{I}_m\right)\right) - \int_0^{\gamma}\mathrm{mmse}(\boldsymbol{X},t)\,dt\right]. \tag{240}$$

Building on (240), we infer via the monotone convergence theorem that, with the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ denoted by $\lambda_1,\cdots,\lambda_m$,

$$h(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^m\prod_{i=1}^m\lambda_i\right) + \frac{1}{2}\int_0^{\infty}\mathrm{mmse}(\boldsymbol{X},t) - \sum_{i=1}^m\frac{\lambda_i}{1+\lambda_i t}\,dt. \tag{241}$$

This equation yields the desired result $h(\boldsymbol{X}) = \frac{1}{2}\int_0^{\infty}\mathrm{mmse}(\boldsymbol{X},t) - \frac{m}{2\pi e+t}\,dt$ since $\log\left((2\pi e)^m\prod_{i=1}^m\lambda_i\right) = \int_0^{\infty}\sum_{i=1}^m\frac{\lambda_i}{1+\lambda_i t} - \frac{m}{2\pi e+t}\,dt$, completing the proof of the lemma.

B. Proof of Theorem 4

We derive the multidimensional version of Theorem 4 here. Fix an m-dimensional random vector $\boldsymbol{V}$. We may assume $\det \boldsymbol{\Sigma}_{\boldsymbol{V}} \neq 0$, for otherwise the result follows immediately from $h_n(\boldsymbol{V}) = h(\boldsymbol{V}) = -\infty$ for all n. In view of monotonicity of $\mathrm{pmmse}_n(\boldsymbol{V}, t)$ in n, and since $h_1(\boldsymbol{V})$ is finite, it suffices by the monotone convergence theorem and the equation

$$h(\boldsymbol{V}) = \frac{1}{2}\int_0^\infty \mathrm{mmse}(\boldsymbol{V}, t) - \frac{m}{2\pi e + t}\, dt \tag{242}$$

to show that $\mathrm{pmmse}_n(\boldsymbol{V}, t) \to \mathrm{mmse}(\boldsymbol{V}, t)$ as $n \to \infty$. Let $\boldsymbol{N} \sim \mathcal{N}(0, \boldsymbol{I}_m)$ be independent of $\boldsymbol{V}$. A simple application of the triangle inequality yields that it suffices to prove the convergence

$$E_n\left[\boldsymbol{V} \mid \sqrt{t}\boldsymbol{V} + \boldsymbol{N}\right] \to \mathbb{E}\left[\boldsymbol{V} \mid \sqrt{t}\boldsymbol{V} + \boldsymbol{N}\right]. \tag{243}$$

We deduce (243) from Theorem 12, as follows.

Denote $\boldsymbol{Z}^{(t)} := \sqrt{t}\boldsymbol{V} + \boldsymbol{N}$, and let $Z_j^{(t)}$ be the j-th entry of $\boldsymbol{Z}^{(t)}$. Fix $t \geq 0$. To apply Theorem 12, we only need to show that the $Z_j^{(t)}$ do not satisfy a nontrivial polynomial relation. We show this by induction on m. The case $m = 1$ follows since $Z_1^{(t)}$ is continuous. Assume that we have shown that $Z_1^{(t)}, \cdots, Z_{m-1}^{(t)}$ do not satisfy a nontrivial polynomial relation, and that $m \geq 2$. Suppose, for the sake of contradiction, that q is a polynomial in m variables such that $q(\boldsymbol{Z}^{(t)}) = 0$. Write $q(u_1, \cdots, u_m) = \sum_{k\in[d]} q_k(u_1, \cdots, u_{m-1})u_m^k$ for some polynomials q_k in $m-1$ variables such that $q_d \neq 0$. Squaring $q(\boldsymbol{Z}^{(t)}) = 0$ and taking the conditional expectation with respect to N_m we obtain

$$0 = \mathbb{E}\left[q\left(\boldsymbol{Z}^{(t)}\right)^2 \middle| N_m\right] = \sum_{k\in[2d]} \beta_k N_m^k \tag{244}$$

for some real constants β_k with the leading constant $\beta_{2d} := \|q_d(Z_1^{(t)}, \cdots, Z_{m-1}^{(t)})\|_2^2$. Since N_m is continuous, equation (244) cannot be a nontrivial polynomial relation for N_m. Thus, we must have $\beta_{2d} = 0$, i.e., $q_d(Z_1^{(t)}, \cdots, Z_{m-1}^{(t)}) = 0$. By the induction hypothesis, $q_d = 0$ identically, a contradiction. Therefore, no nontrivial polynomial relation $q(\boldsymbol{Z}^{(t)}) = 0$ can hold, and the inductive proof is complete. Finally, applying Theorem 12, we deduce the limit in (243), thereby completing the proof of the theorem.

C. Proof of Theorem 5

Consider the first case, namely, X is discrete with finite support and Y is continuous whose MGF exists and for which $h(Y) > -\infty$. The existence of the MGF of Y implies the existence of the MGFs of $Y^{(x)}$ for each $x \in \mathrm{supp}(X)$. Since $\sigma_Y^2 < \infty$, we have that $h(Y)$ is finite. In addition, for each $x \in \mathrm{supp}(X)$, we infer from $\sigma_{Y^{(x)}}^2 < \infty$ the existence of the differential entropy $h(Y \mid X = x)$ and that $h(Y \mid X = x) < \infty$. If $\min_{x\in\mathrm{supp}(X)} h(Y \mid X = x) > -\infty$, then $I(X;Y) = h(Y) - h(Y \mid X)$; this latter equation also holds if $h(Y \mid X = x) = -\infty$ for some $x \in \mathrm{supp}(X)$. Therefore, Theorem 4 implies (23).

Now, consider the second case instead, so both X and Y are continuous random variables whose MGFs exist and that satisfy $h(X), h(Y) > -\infty$. We also assume that $I(X;Y) < \infty$ or else (X, Y) is not continuous. From these assumptions, we conclude that both $h(X)$ and $h(Y)$ are finite and $h(X, Y)$ exists. Thus, we obtain $I(X;Y) = h(X) + h(Y) - h(X, Y)$. By Theorem 4, we have that $h_n(X) \to h(X)$ and $h_n(Y) \to h(Y)$ as $n \to \infty$. Finally, note that the MGF of (X, Y) exists by the assumption that the MGFs of X and Y exist. Thus, by Theorem 4, we have that $h_n(X, Y) \to h(X, Y)$ too. The desired result (24) follows.

Appendix F
Estimator Implementation

We show in this appendix how to implement the proposed estimators numerically. Note that $\mathrm{pmmse}_n(X, t)$ contains roughly n^2 terms, and that numerically integrating this rational function can be done efficiently using built-in quadrature methods. Precomputing the function $t \mapsto \mathrm{pmmse}_{10}(X, t)$ takes a couple of minutes on a commercial laptop, whereas querying this rational function can be done in constant time. However, we need to develop the expressions of our approximations of differential entropy further to avoid possible issues that could arise from numerically computing the improper integral over $[0, \infty)$. To illustrate this issue, consider the expression for $h_2(X)$. For convenience, define the function $\delta_{X,n} : (0, \infty) \to [0, \infty)$ by

$$\delta_{X,n}(t) := \det \boldsymbol{M}_{\sqrt{t}X+N,n} \tag{245}$$

for a $2n$-times integrable random variable X. Recall that $\delta_{X,n}$ is the denominator of $\mathrm{pmmse}_n(X, \cdot)$. Recall from (10) that a zero-mean unit-variance random variable X satisfies

$$\mathrm{pmmse}_2(X, t) = \frac{2 + 4t + (\mathcal{X}_4 - \mathcal{X}_3^2 - 1)t^2}{2 + 6t + (\mathcal{X}_4 + 3)t^2 + (\mathcal{X}_4 - \mathcal{X}_3^2 - 1)t^3}. \tag{246}$$

For example, when $X \sim \mathrm{Unif}([-\sqrt{3}, \sqrt{3}])$, so

$$(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4) = \left(0, 1, 0, \frac{9}{5}\right), \tag{247}$$

we obtain

$$\mathrm{pmmse}_2(X, t) = \frac{5 + 10t + 2t^2}{5 + 15t + 12t^2 + 2t^3}. \tag{248}$$

Now, consider the expression for $h_2(X)$ in (79), namely,

$$h_2(X) = \frac{1}{2}\int_0^\infty \frac{5 + 10t + 2t^2}{5 + 15t + 12t^2 + 2t^3} - \frac{1}{2\pi e + t}\, dt. \tag{249}$$

The integral in (249) converges, but a numerical computation might not be able to capture this convergence as the expression for the integrand is a difference of non-integrable functions that both decay as $1/t$. To avoid this possible issue, we subtract a $1/t$ term from both of these non-integrable functions. More precisely, denoting differentiation with respect to t by a prime, we write

$$\begin{aligned}\mathrm{pmmse}_2(X, t) &= \frac{5 + 10t + 2t^2 - \frac{1}{3}\delta'_{X,2}(t) + \frac{1}{3}\delta'_{X,2}(t)}{\delta_{X,2}(t)} \\ &= \frac{2t}{5 + 15t + 12t^2 + 2t^3} + \frac{1}{3}\frac{d}{dt}\log \delta_{X,2}(t)\end{aligned}$$

and

$$\frac{1}{2\pi e+t}=\frac{d}{dt}\log(2\pi e+t). \tag{250}$$

The integrand $\mathrm{pmmse}_2(X,t)-1/(2\pi e+t)$ now becomes

$$\frac{2t}{5+15t+12t^2+2t^3}+\frac{d}{dt}\log\frac{\delta_{X,2}(t)^{1/3}}{2\pi e+t}. \tag{251}$$

The advantage in having the integrand in this form is that the first term is well-behaved (it decays as $1/t^2$), and the second term's integral can be given in closed form

$$\int_0^\infty\left(\log\frac{\delta_{X,2}(t)^{1/3}}{2\pi e+t}\right)'dt=\log\left(2\pi e\left(\frac{2}{5}\right)^{1/3}\right). \tag{252}$$

Therefore, equation (249) becomes

$$h_2(X)=\frac{1}{2}\log\frac{2\pi e}{(5/2)^{1/3}}+\int_0^\infty\frac{t}{5+15t+12t^2+2t^3}\,dt. \tag{253}$$

We use equation (253) instead of (249) for numerical computation. Note that this resolves the same numerical instability issue when estimating from data: if $\mathcal{S}=\{X_j\}_{j=1}^m$ is a multiset of i.i.d. samples distributed according to P_X, and if $U\sim\mathrm{Unif}(\mathcal{S})$, we compute the estimate $\widehat{h}_2(\mathcal{S})=h_2(U)$ of $h_2(X)$ via an expression analogous to that in (253) where X is replaced with U.

The procedure of obtaining expression (253) from (249) can be carried out for a general X and n such that $\mathbb{E}[X^{2n}]<\infty$ and $|\mathrm{supp}(X)|>n$, as follows. Let $\theta_{X,n}:[0,\infty)\to[0,\infty)$ be the polynomial that is the numerator of $\mathrm{pmmse}_n(X,t)$, i.e., $\theta_{X,n}(t):=\delta_{X,n}(t)\cdot\mathrm{pmmse}_n(X,t)$. Thus, we have that

$$\mathrm{pmmse}_n(X,t)=\frac{\theta_{X,n}(t)}{\delta_{X,n}(t)}. \tag{254}$$

We define the function $\rho_{X,n}:[0,\infty)\to\mathbb{R}$ by

$$\rho_{X,n}(t):=\frac{\theta_{X,n}(t)-d_n^{-1}\delta'_{X,n}(t)}{2\delta_{X,n}(t)}, \tag{255}$$

where $d_n=\binom{n+1}{2}$. By the analysis of the coefficients in $\mathrm{pmmse}_n(X,t)$ proved in Theorem 2, we have that $\rho_{X,n}(0)=0$ and

$$\rho_{X,n}(t)=O\left(t^{-2}\right) \tag{256}$$

as $t\to\infty$. In particular, $\rho_{X,n}$ is integrable over $[0,\infty)$. The following formula for differential entropy directly follows from the definition of h_n in (79).

Lemma 14. *For any random variable X satisfying $\mathbb{E}[X^{2n}]<\infty$ and $|\mathrm{supp}(X)|>n$, we have the formula*

$$h_n(X)=\frac{1}{2}\log\left(2\pi e\left(\frac{\det\boldsymbol{M}_{X,n}}{\det\boldsymbol{M}_{N,n}}\right)^{1/d_n}\right)+\int_0^\infty\rho_{X,n}(t)\,dt, \tag{257}$$

where $d_n=\binom{n+1}{2}$, $N\sim\mathcal{N}(0,1)$, and $\rho_{X,n}$ is as defined in (255).

A similar conclusion holds for mutual information.

Lemma 15. *Fix a discrete random variable X with finite support, and a $2n$-times integrable continuous random variable Y. We have that*

$$I_n(X;Y)=\frac{1}{n(n+1)}\log\frac{\det\boldsymbol{M}_{Y,n}}{\prod_{x\in\mathrm{supp}(X)}\left(\det\boldsymbol{M}_{Y^{(x)},n}\right)^{P_X(x)}}+\int_0^\infty\rho_{Y,n}(t)-\mathbb{E}_X\left[\rho_{Y^{(X)},n}(t)\right]\,dt, \tag{258}$$

where for each $x\in\mathrm{supp}(X)$ we denote by $Y^{(x)}$ the random variable Y conditioned on $\{X=x\}$.

Note that in Lemmas 14 and 15, the determinants $\det\boldsymbol{M}_{A,n}$ and the rational functions $\rho_n(A;t)$, for $A\in\{X,Y\}$ or $A\in\{Y^{(x)}\,;\,x\in\mathrm{supp}(X)\}$, are completely determined by the first $2n$ moments of A. To obtain the estimates $\widehat{h}_n$ and $\widehat{I}_n$ given samples, the moments of A are replaced with their respective sample moments in formulas (257) and (258).

Appendix G
Proofs of Subsection V-A: Consistency

A. Proof of Theorem 9: Consistency of the Differential Entropy Estimator

We use the formula for h_n given in Lemma 14,

$$h_n(X)=\frac{1}{2}\log\left(2\pi e\left(\frac{\det\boldsymbol{M}_{X,n}}{\det\boldsymbol{M}_{N,n}}\right)^{1/d_n}\right)+\int_0^\infty\rho_{X,n}(t)\,dt, \tag{259}$$

where $d_n=\binom{n+1}{2}$ and $N\sim\mathcal{N}(0,1)$. We may assume that N is independent of X and the X_j. For each $m\in\mathbb{N}$, let $\mathcal{S}_m:=\{X_j\}_{j\in[m]}$, and consider the sequence $\{U_m\sim\mathrm{Unif}(\mathcal{S}_m)\}_{m\in\mathbb{N}}$. For each $m\in\mathbb{N}$, let $\mathfrak{E}_m$ be the event that $X_0,\cdots,X_m$ are distinct, and let $\mathfrak{E}$ be the event that the X_j, for $j\in\mathbb{N}$, are all distinct. Whenever $m\geq n$ and $\mathfrak{E}_m$ occurs, we have by Definition 4 of $\widehat{h}_n$ and formula (259) for h_n the following estimate

$$\widehat{h}_n\left(\mathcal{S}_m\right)=\frac{1}{2}\log\left(2\pi e\left(\frac{\det\boldsymbol{M}_{U_m,n}}{\det\boldsymbol{M}_{N,n}}\right)^{1/d_n}\right)+\int_0^\infty\rho_{U_m,n}(t)\,dt. \tag{260}$$

Since X is continuous, we have that $P(\mathfrak{E}_m)=1$ for every $m\in\mathbb{N}$. Further, $\mathfrak{E}_0\supset\mathfrak{E}_1\supset\cdots$ and $\mathfrak{E}=\bigcap_{m\in\mathbb{N}}\mathfrak{E}_m$, hence $P(\mathfrak{E})=1$. Therefore, for the purpose of proving the almost-sure limit $\widehat{h}_n\left(\mathcal{S}_m\right)\to h_n(X)$, we may assume that $\mathfrak{E}$ occurs. We first treat convergence of the integral part. We show that the integral part is a continuous function of the moments, then the continuous mapping theorem yields that

$$\int_0^\infty\rho_{U_m,n}(t)\,dt\to\int_0^\infty\rho_{X,n}(t)\,dt \tag{261}$$

almost surely as $m\to\infty$ because sample moments converge almost surely to the moments. A similar method is then applied to the convergence of the $\log\det\boldsymbol{M}_{X,n}$ part.

We fix $n \in \mathbb{N}_{\geq 1}$, and assume $m \geq n$ throughout the proof. We use the following notation. The $2n$-dimensional random vector $\boldsymbol{\mu}^{(m)}$ consists of the first $2n$ moments of U_m

$$\boldsymbol{\mu}^{(m)} := \left(\frac{\sum_{j=0}^{m} X_j}{m+1}, \cdots, \frac{\sum_{j=0}^{m} X_j^{2n}}{m+1} \right)^T . \tag{262}$$

Let $\mu_k^{(m)}$ be the k-th coordinate of $\boldsymbol{\mu}^{(m)}$, so $\boldsymbol{\mu}^{(m)} = \left(\mu_1^{(m)}, \cdots, \mu_{2n}^{(m)}\right)^T$. We write $\mathcal{X}_k := \mathbb{E}\left[X^k\right]$ for $k \in \mathbb{N}$, and consider the constant vector

$$\boldsymbol{\mathcal{X}} := (\mathcal{X}_k)_{1 \leq k \leq 2n} . \tag{263}$$

By the strong law of large numbers, we have the almost-sure convergence $\mu_k^{(m)} \to \mathcal{X}_k$ for each $1 \leq k \leq 2n$. Then, $\boldsymbol{\mu}^{(m)} \to \boldsymbol{\mathcal{X}}$ almost surely as $m \to \infty$. We show next that the function $\boldsymbol{\mathcal{X}} \mapsto \int_0^\infty \rho_{X,n}(t)\, dt$ is continuous.

By definition of $\rho_{X,n}$ (see (255)), there are polynomials $A_1, \cdots, A_{d_n-2}$ and $B_1, \cdots, B_{d_n}$ in $2n$ variables such that

$$\rho_{X,n}(t) = \frac{\sum_{j=1}^{d_n-2} A_j(\boldsymbol{\mathcal{X}})\ t^j}{c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}})\ t^j} \tag{264}$$

where $c_n := \prod_{k=1}^{n} k!$ (we are subsuming the $1/2$ factor in (255) in the numerator, so we have the equality $\delta_{X,n}(t) = c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}}) t^j$). Being polynomials, each of the A_j and the B_ℓ is continuous over $\mathbb{R}^{2n}$. Then, by the continuous mapping theorem, we have the almost-sure convergences

$$A_j\left(\boldsymbol{\mu}^{(m)}\right) \to A_j(\boldsymbol{\mathcal{X}}) \quad \text{and} \quad B_\ell\left(\boldsymbol{\mu}^{(m)}\right) \to B_\ell(\boldsymbol{\mathcal{X}}) \tag{265}$$

as $m \to \infty$ for each $1 \leq j \leq d_n - 2$ and $1 \leq \ell \leq d_n$. Denote

$$\boldsymbol{A}(\boldsymbol{\mathcal{X}}) := (A_j(\boldsymbol{\mathcal{X}}))_{1 \leq j \leq d_n - 2} , \tag{266}$$

$$\boldsymbol{B}(\boldsymbol{\mathcal{X}}) := (B_j(\boldsymbol{\mathcal{X}}))_{1 \leq j \leq d_n} . \tag{267}$$

We show next that the there is an open set $\mathcal{O} \subset \mathbb{R}^{d_n}$ containing the point $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$ such that the mapping $f : \mathbb{R}^{d_n-2} \times \mathcal{O} \to \mathbb{R}$ defined by

$$f(p_1, \cdots, p_{d_n-2}, q_1, \cdots, q_{d_n}) := \int_0^\infty \frac{\sum_{j=1}^{d_n-2} p_j t^j}{c_n + \sum_{j=1}^{d_n} q_j t^j}\, dt \tag{268}$$

is continuous at the point $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$. To this end, we shall show first that the mapping in (268) is well-defined on an open neighborhood of $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$. In other words, the denominator of the integrand $t \mapsto c_n + \sum_{j=1}^{d_n} q_j t^j$ cannot have a root $t \in [0, \infty)$ for any $\boldsymbol{q} \in \mathcal{O}$, and the rational function integrand has to be integrable. For integrability, we will restrict the set $\mathcal{O}$ to contain only points having $q_{d_n} > 0$, so showing that the integrand's denominator is strictly positive over $t \in [0, \infty)$ will be enough to deduce integrability in (268).

We consider the subset $\mathcal{G} \subset \mathbb{R}^{d_n}$ defined by

$$\mathcal{G} := \left\{ \boldsymbol{g} \in \mathbb{R}^{d_n} \,;\, g_{d_n} > 0, \sum_{\ell=1}^{d_n} g_j t^j > -c_n \ \text{ for all } t \geq 0 \right\} \tag{269}$$

where in this definition and the subsequent argument we set $\boldsymbol{g} = (g_1, \cdots, g_{d_n})^T$. Note that $\boldsymbol{B}(\boldsymbol{\mathcal{X}}) \in \mathcal{G}$. Indeed, since X is continuous, $B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X,n} > 0$; similarly, for every $t \in [0, \infty)$, continuity of $\sqrt{t}X + N$ implies that $\det \boldsymbol{M}_{\sqrt{t}X+N} > 0$ (recall that $c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}}) t^j = \det \boldsymbol{M}_{\sqrt{t}X+N}$). We show that $\mathcal{G}$ is an open set. Fix $\boldsymbol{g} \in \mathcal{G}$ and $\varepsilon_1 \in (0, g_{d_n})$. We have that the polynomial $\sum_{j=1}^{d_n} (g_j - \varepsilon_1) t^j$ is eventually increasing and approaches infinity as $t \to \infty$. Let $t_0 > 1$ be such that for every $t > t_0$ we have

$$\sum_{\ell=1}^{d_n} (g_j - \varepsilon_1) t^j > -c_n. \tag{270}$$

Being continuous, the polynomial $\sum_{j=1}^{d_n} g_j t^j$ attains its minimum over the compact set $[0, t_0]$. Let s denote this minimum, and note that $s > -c_n$. Let $\varepsilon \in (0, 1)$ be defined by

$$\varepsilon := \frac{1}{2} \min \left(\varepsilon_1, \frac{(s + c_n)(t_0 - 1)}{t_0 (t_0^{d_n} - 1)} \right). \tag{271}$$

As $\varepsilon < \varepsilon_1$, inequality (270) yields that for every $t > t_0$

$$\sum_{j=1}^{d_n} (g_j - \varepsilon) t^j > -c_n. \tag{272}$$

In addition, for any $t \in [0, t_0]$,

$$\begin{aligned} \sum_{j=1}^{d_n} (g_j - \varepsilon) t^j &= \sum_{j=1}^{d_n} g_j t^j - \varepsilon \sum_{j=1}^{d_n} t^j \geq s - \varepsilon \sum_{j=1}^{d_n} t_0^j \\ &> s - \frac{(s + c_n)(t_0 - 1)}{t_0 (t_0^{d_n} - 1)} \sum_{j=1}^{d_n} t_0^j \\ &= s - (s + c_n) = -c_n. \end{aligned} \tag{273}$$

Thus, combining (272) and (273) we obtain

$$\sum_{j=1}^{d_n} (g_j - \varepsilon) t^j > -c_n \tag{274}$$

for every $t \in [0, \infty)$. Hence, for any $(\delta_j)_{1 \leq j \leq d_n} =: \boldsymbol{\delta} \in \mathbb{R}^{d_n}$ such that $\|\boldsymbol{\delta}\|_2 < \varepsilon$, we have that for all $t \in [0, \infty)$

$$\sum_{j=1}^{d_n} (g_j - \delta_j) t^j \geq \sum_{j=1}^{d_n} (g_j - \|\boldsymbol{\delta}\|_2) t^j \geq \sum_{j=1}^{d_n} (g_j - \varepsilon) t^j > -c_n. \tag{275}$$

In other words, the open ball $\{\boldsymbol{q} \in \mathbb{R}^{d_n} \,;\, \|\boldsymbol{q} - \boldsymbol{g}\| < \varepsilon\}$ lies within $\mathcal{G}$. This completes the proof that $\mathcal{G}$ is open. Then, the function f given by (268) is well-defined on the open set $\mathbb{R}^{d_n-2} \times \mathcal{G}$. We will replace $\mathcal{G}$ with an open box $\mathcal{O} \subset \mathcal{G}$ to simplify the notation for the proof of continuity of f.

By openness of $\mathcal{G}$, there is an $\eta_1 \in (0, B_{d_n}(\boldsymbol{\mathcal{X}}))$ such that the open box

$$\mathcal{O}_1 := \prod_{j=1}^{d_n} (B_j(\boldsymbol{\mathcal{X}}) - \eta_1, B_j(\boldsymbol{\mathcal{X}}) + \eta_1) \subset \mathcal{G} \tag{276}$$

contains $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$. Since $\mathcal{O}_1 \subset \mathcal{G}$, we have by the definition of $\mathcal{G}$ in (269) that for any $\boldsymbol{g} \in \mathcal{O}_1$ the lower bound

$$c_n + \sum_{\ell=1}^{d_n} g_\ell t^\ell > 0 \tag{277}$$

holds for every $t \geq 0$. In particular, with $\eta := \eta_1/2$, the set

$$\mathcal{O} := \prod_{j=1}^{d_n} (B_j(\boldsymbol{\mathcal{X}}) - \eta, B_j(\boldsymbol{\mathcal{X}}) + \eta) \subset \mathcal{O}_1 \subset \mathcal{G} \tag{278}$$

is an open set containing $\boldsymbol{B}(\boldsymbol{\mathcal{X}})$, and the point $(B_j(\boldsymbol{\mathcal{X}}) - \eta)_{1 \leq j \leq 2n}$ lies inside $\mathcal{G}$. Then, the function $f : \mathbb{R}^{d_n - 2} \times \mathcal{O} \to \mathbb{R}$ given by (268) is well-defined, and for any $\boldsymbol{g} \in \mathcal{O}$ we have the lower bound (over $t \in [0, \infty)$)

$$c_n + \sum_{\ell=1}^{d_n} g_\ell t^\ell \geq c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - \eta) t^\ell > 0. \tag{279}$$

From (279), Lebesgue's dominated convergence shows continuity of f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$, as follows.

Let $\boldsymbol{w} := (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{d_n - 2} \times \mathcal{O}$ be such that $\|\boldsymbol{w}\|_2 < \eta$. The integrand in f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) - (\boldsymbol{u}, \boldsymbol{v})$ may be bounded as

$$\left| \frac{\sum_{j=1}^{d_n - 2} (A_j(\boldsymbol{\mathcal{X}}) - u_j) t^j}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - v_\ell) t^\ell} \right| = \frac{\left| \sum_{j=1}^{d_n - 2} (A_j(\boldsymbol{\mathcal{X}}) - u_j) t^j \right|}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - v_\ell) t^\ell} \tag{280}$$

$$\leq \frac{\sum_{j=1}^{d_n - 2} (|A_j(\boldsymbol{\mathcal{X}})| + \eta) t^j}{c_n + \sum_{\ell=1}^{d_n} (B_\ell(\boldsymbol{\mathcal{X}}) - \eta) t^\ell}. \tag{281}$$

The bound in (281) is uniform in $\boldsymbol{w}$, and the upper bound is integrable over $[0, \infty)$ as the denominator's degree exceeds that of the numerator by at least 2 and the denominator is strictly positive by (279). Hence, by Lebesgue's dominated convergence

$$\lim_{\|\boldsymbol{w}\| \to 0} f\left((\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) - \boldsymbol{w} \right) = f\left(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}) \right), \tag{282}$$

i.e., f is continuous at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$, as desired. Denote

$$\boldsymbol{A}^{(m)} := \left(A_j(\boldsymbol{\mu}^{(m)}) \right)_{1 \leq j \leq d_n - 2}, \tag{283}$$

$$\boldsymbol{B}^{(m)} := \left(B_\ell(\boldsymbol{\mu}^{(m)}) \right)_{1 \leq \ell \leq d_n}. \tag{284}$$

We have the formulas

$$f(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) = \int_0^\infty \rho_{U_m, n}(t)\, dt \tag{285}$$

and

$$f(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}})) = \int_0^\infty \rho_{X,n}(t)\, dt. \tag{286}$$

Since $(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) \to (\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$ almost surely, continuity of f at $(\boldsymbol{A}(\boldsymbol{\mathcal{X}}), \boldsymbol{B}(\boldsymbol{\mathcal{X}}))$ implies by the continuous mapping theorem that

$$f(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}) \to f(\boldsymbol{A}(\boldsymbol{\nu}), \boldsymbol{B}(\boldsymbol{\nu})) \tag{287}$$

almost surely as $m \to \infty$, i.e., (261) holds.

Now, for the convergence of the logarithmic part, recall that we have the almost sure convergence

$$\det \boldsymbol{M}_{U_m, n} = B_{d_n}(\boldsymbol{\mu}^{(m)}) \to B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X,n} \tag{288}$$

as $m \to \infty$. As the mapping $\mathbb{R}_{>0} \to \mathbb{R}$ defined by $q \mapsto \log q$ is continuous, the continuous mapping theorem yields that

$$\log \det \boldsymbol{M}_{U_m, n} \to \log \det \boldsymbol{M}_{X,n} \tag{289}$$

almost surely as $m \to \infty$. Combining (287) and (289), we obtain that

$$\widehat{h}_n\left(\mathcal{S}_m\right) \to h_n(X) \tag{290}$$

almost surely as $m \to \infty$. Finally, (95) follows from (290) by Theorem 4.

B. Proof of Corollary 4: Consistency of the Mutual Information Estimator

Denote $\mathcal{S}_m = \{(X_j, Y_j)\}_{j \in [m]}$, and consider the empirical measure

$$\widehat{P}_m(x) := \sum_{j \in [m]} \frac{\delta_x(X_j)}{m+1}. \tag{291}$$

Let $\mathfrak{D}_m$ be the event that for each $x \in \operatorname{supp}(X)$ there is a subset of indices $J_x \subset [m]$ of size at least $n+1$ such that: **i)** $X_j = x$ for each $j \in J_x$, and **ii)** the Y_j, for $j \in J_x$, are distinct. If $\mathfrak{D}_m$ occurs, then we may write

$$\widehat{I}_n(\mathcal{S}_m) = \widehat{h}_n(\mathcal{A}_m) - \sum_{x \in \operatorname{supp}(X)} \widehat{P}_m(x)\, \widehat{h}_n(\mathcal{B}_{m,x}), \tag{292}$$

where $\mathcal{A}_m := \{Y_j\}_{j \in [m]}$ and $\mathcal{B}_{m,x} := \{Y_j \ ; \ j \in [m], X_j = x\}$. By the assumption of continuity of Y, it holds with probability 1 that the Y_j, for $j \in \mathbb{N}$, are all distinct. In addition, we have that $P_X(x) > 0$ for each $x \in \operatorname{supp}(X)$. Therefore, $P(\mathfrak{D}_m) \to 1$ as $m \to \infty$. Note that $\mathfrak{D}_0 \subset \mathfrak{D}_1 \subset \cdots$.

Let $\mathfrak{C}$ be the event that $\lim_{m \to \infty} \widehat{h}_n(\mathcal{A}_m) = h_n(Y)$ and, for each $x \in \operatorname{supp}(X)$, $\lim_{m \to \infty} \widehat{h}_n(\mathcal{B}_{m,x}) = h_n(Y^{(x)})$. By Theorem 9 and finiteness of $\operatorname{supp}(X)$, for each integer $m' \geq (n+1)|\operatorname{supp}(X)|$, we have that $P(\mathfrak{C} \mid \mathfrak{D}_{m'}) = 1$. Let $\mathfrak{F}$ be the event that the empirical measure $\widehat{P}_m$ converges to P_X, i.e., that for each $x \in \operatorname{supp}(X)$ the limit $\widehat{P}_m(x) \to P_X(x)$ holds as $m \to \infty$. By the strong law of large numbers, $P(\mathfrak{F}) = 1$. Therefore,

$$\begin{aligned} P\left(\lim_{m \to \infty} \widehat{I}_n(\mathcal{S}_m) = I_n(X;Y) \right) &\geq P(\mathfrak{C} \cap \mathfrak{F} \cap \mathfrak{D}_{m'}) \\ &\geq P(\mathfrak{F}) + P(\mathfrak{C} \cap \mathfrak{D}_{m'}) - 1 \\ &= P(\mathfrak{D}_{m'}). \end{aligned} \tag{293}$$

Taking $m' \to \infty$, we deduce that $\widehat{I}_n(\mathcal{S}_m) \to I_n(X;Y)$ almost surely.

Appendix H
Proofs of Subsection V-B: Sample Complexity

A. Proof of Proposition 4: Differential Entropy

Suppose $\operatorname{supp}(X) \subset [p, q] \subset (0, \infty)$, and write $\mathcal{S} = \{X_j\}_{j=1}^m$; note that we may assume, without loss of generality, that X is strictly positive because h_n is shift-invariant. We use the same notation in Appendix G. In particular, $\mathcal{X}_k = \mathbb{E}[X^k]$, and $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \cdots, \mathcal{X}_{2n})^T$. Let $U \sim \operatorname{Unif}(\mathcal{S})$. Let $\mathfrak{E}_m$ be the event that $X_1, \cdots, X_m$ are distinct. From (259)–(260), if $m > n$ and $\mathfrak{E}_m$ holds, then we have that

$$\begin{aligned} \widehat{h}_n(\mathcal{S}) - h_n(X) = \frac{1}{2d_n} \log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \\ + \int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\, dt. \end{aligned} \tag{294}$$

By the assumption of continuity of X, we have that $P(\mathfrak{E}_m) = 1$ for every m. Therefore, for the purpose of proving a sample complexity bound, we may assume that $m > n$ and that $\mathfrak{E}_m$ occurs.

We will consider the determinant part and the integral part in (294) separately, but the proof technique will be the same. Let A_j and B_ℓ be the polynomials as defined by (264) in Appendix G, so

$$\rho_{X,n}(t) = \frac{\sum_{j=1}^{d_n-2} A_j(\boldsymbol{\mathcal{X}})\ t^j}{c_n + \sum_{j=1}^{d_n} B_j(\boldsymbol{\mathcal{X}})\ t^j} \tag{295}$$

where $c_n := \prod_{k=1}^{n} j!$. We split each of the polynomials A_j and B_ℓ into a positive part and a negative part. More precisely, we collect the terms in A_j that have positive coefficients into a polynomial $A_j^{(+)}$, and the terms in A_j with negative coefficients into a polynomial $-A_j^{(-)}$ (so $A_j^{(-)}$ has positive coefficients, and $A_j = A_j^{(+)} - A_j^{(-)}$). Define $B_\ell^{(+)}$ and $B_\ell^{(-)}$ from B_ℓ similarly. By positivity of X, each moment $\mathcal{X}_k$ is (strictly) positive. Then, we may write

$$\rho_{X,n}(t) = \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \tag{296}$$

with the polynomials in t

$$f_X(t) := \sum_{j=1}^{d_n-2} A_j^{(+)}(\boldsymbol{\mathcal{X}}) t^j \tag{297}$$

$$g_X(t) := \sum_{j=1}^{d_n-2} A_j^{(-)}(\boldsymbol{\mathcal{X}}) t^j \tag{298}$$

$$u_X(t) := c_n + \sum_{\ell=1}^{d_n} B_\ell^{(+)}(\boldsymbol{\mathcal{X}}) t^\ell \tag{299}$$

$$v_X(t) := \sum_{\ell=1}^{d_n} B_\ell^{(-)}(\boldsymbol{\mathcal{X}}) t^\ell, \tag{300}$$

having all non-negative coefficients. We note that we have suppressed the dependence on n in the notation used for these polynomials for readability. For $q \in \{f, g, u, v\}$, let q_U be the random variable whose value is what is obtained via q_X when the moments of X are replaced with the sample moments obtained from the samples $\mathcal{S}$, e.g.,

$$f_U(t) := \sum_{j=1}^{d_n-2} A_j^{(+)}\left(\frac{\sum_{i=1}^m X_i}{m}, \cdots, \frac{\sum_{i=1}^m X_i^{2n}}{m}\right) t^j. \tag{301}$$

Note that $u_U(t) - v_U(t) = \det \boldsymbol{M}_{\sqrt{t}U+N,n} > 0$, where $N \sim \mathcal{N}(0,1)$ is independent of $X, X_1, \cdots, X_m$. Then the function

$$\rho_{U,n}(t) = \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} \tag{302}$$

is well-defined over $t \in [0, \infty)$. By the homogeneity properties proved in Theorem 2, we know that the total degree of A_j is at most $2j+2$, and the total degree of B_ℓ is at most 2ℓ. Therefore, for any $\eta \in (0,1)$ and $\boldsymbol{\xi} \in \mathbb{R}^{2n}_{\geq 0}$, we have the inequalities

$$(1-\eta)^{2j+2} A_j^{(\pm)}(\boldsymbol{\xi}) \leq A_j^{(\pm)}((1-\eta)\boldsymbol{\xi}) \tag{303}$$

$$A_j^{(\pm)}((1+\eta)\boldsymbol{\xi}) \leq (1+\eta)^{2j+2} A_j^{(\pm)}(\boldsymbol{\xi}) \tag{304}$$

$$(1-\eta)^{2\ell} B_\ell^{(\pm)}(\boldsymbol{\xi}) \leq B_\ell^{(\pm)}((1-\eta)\boldsymbol{\xi}) \tag{305}$$

$$B_\ell^{(\pm)}((1+\eta)\boldsymbol{\xi}) \leq (1+\eta)^{2\ell} B_\ell^{(\pm)}(\boldsymbol{\xi}) \tag{306}$$

for every $1 \leq j \leq d_n - 2$ and $1 \leq \ell \leq d_n$.

For each $\eta \in (0,1)$, we denote the event

$$\mathfrak{A}_{n,\eta}(\mathcal{S}) := \left\{1 - \eta \leq \frac{\sum_{i=1}^m X_i^k}{m\mathcal{X}_k} \leq 1 + \eta \text{ for all } k \in [2n]\right\}, \tag{307}$$

Hoeffding's inequality yields that, for any $z > 0$ and $1 \leq k \leq 2n$,

$$P\left(\left|\mathcal{X}_k - \frac{1}{m}\sum_{i=1}^m X_i^k\right| \geq z\right) \leq 2e^{-2mz^2/\left(q^k - p^k\right)^2}. \tag{308}$$

Setting $z = \eta\mathcal{X}_k \geq \eta p^k > 0$ for $\eta \in (0,1)$ yields that

$$\begin{aligned} &P\left((1-\eta)\mathcal{X}_k < \frac{1}{m}\sum_{i=1}^m X_i^k < (1+\eta)\mathcal{X}_k\right) \\ &\quad \geq 1 - 2e^{-2m\eta^2/\left((q/p)^k - 1\right)^2}. \end{aligned} \tag{309}$$

Therefore, the union bound yields that

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \geq 1 - 4ne^{-2m\eta^2/\left((q/p)^{2n} - 1\right)^2}. \tag{310}$$

If $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs, we show a bound on the estimation error that is linear in η

$$\widehat{h}_n(\mathcal{S}) - h_n(X) = O_{X,n}(\eta), \tag{311}$$

independent of the number of samples m, for all small enough η. Then, we choose η to be linear in the error ε to conclude the proof.

We may bound $\rho_{U,n}(t)$ (see (302)) via the bounds in (303)–(306) under the assumption that $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs. If $(1-\eta)\mathcal{X}_k \leq \frac{1}{m}\sum_{i=1}^m X_i^m \leq (1+\eta)\mathcal{X}_k$ holds for every $1 \leq k \leq 2n$, then by (303)–(306) we have that for every $t \geq 0$ and $\eta \in (0,1)$

$$\begin{aligned} &\frac{(1-\eta)^2 f_X((1-\eta)^2 t) - (1+\eta)^2 g_X((1+\eta)^2 t)}{u_X((1+\eta)^2 t) - v_X((1-\eta)^2 t)} \\ &\leq \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} = \rho_{U,n}(t). \end{aligned} \tag{312}$$

For an analogous upper bound, we first verify the positivity

$$u_X((1-\eta)^2 t) - v_X((1+\eta)^2 t) > 0 \tag{313}$$

for every small enough η. Let

$$\mu_X := \sup_{t \in [0,\infty)} \frac{v_X(t)}{u_X(t)}. \tag{314}$$

We show that $\mu_X < 1$. We have the limit

$$\xi_X := \lim_{t\to\infty} \frac{v_X(t)}{u_X(t)} = \frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}. \tag{315}$$

Recall that $B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}) = B_{d_n}(\boldsymbol{\mathcal{X}}) = \det \boldsymbol{M}_{X,n} > 0$ and both $B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})$ and $B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})$ are non-negative, hence $B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) > 0$. Then, $\xi_X < 1$. Thus, there is a $t_0 \geq 0$ such that $v_X(t)/u_X(t) < (1+\xi_X)/2 < 1$ whenever $t > t_0$. Further, by the extreme value theorem, there is a $t_1 \in [0, t_0]$ such that $v_X(t)/u_X(t) \leq v_X(t_1)/u_X(t_1) < 1$ for every $t \in [0, t_0]$. Therefore, $\mu_X \leq \max((1+\xi_X)/2, v_X(t_1)/u_X(t_1)) < 1$, as desired. Note that if $\mu_X = 0$ then $v_X \equiv 0$ identically, in which case (313) trivially holds by positivity of u_X. So, for the purpose of showing (313), it suffices to consider the case $\mu_X \in (0,1)$. Denote

$$\nu := \left(\frac{1+\eta}{1-\eta}\right)^2. \tag{316}$$

Now, since v_X is a polynomial of degree at most d_n, we have that $v_X(\alpha\tau) \leq \alpha^{d_n} v_X(\tau)$ for every $\alpha \geq 1$ and $\tau \geq 0$. Therefore, for every $1 \leq \nu < \mu_X^{-1/d_n}$ and $t \geq 0$, we have that

$$\frac{v_X((1+\eta)^2 t)}{u_X((1-\eta)^2 t)} \leq \left(\frac{1+\eta}{1-\eta}\right)^{2d_n} \cdot \frac{v_X((1-\eta)^2 t)}{u_X((1-\eta)^2 t)} \tag{317}$$

$$\leq \nu^{d_n} \mu_X < 1, \tag{318}$$

i.e., inequality (313) holds. Therefore, for every $1 \leq \nu < \mu_X^{-1/d_n}$ (if $\mu_X = 0$, we allow $1 \leq \nu < \infty$), inequalities (303)–(306) imply the bound

$$\rho_{U,n}(t) = \frac{f_U(t) - g_U(t)}{u_U(t) - v_U(t)} \tag{319}$$

$$\leq \frac{(1+\eta)^2 f_X((1+\eta)^2 t) - (1-\eta)^2 g_X((1-\eta)^2 t)}{u_X((1-\eta)^2 t) - v_X((1+\eta)^2 t)}. \tag{320}$$

Combining (312) and (320), then integrating with respect to t over $[0,\infty)$ and performing a change of variables from t to $(1-\eta)^2 t$, we obtain the bounds

$$\int_0^\infty \frac{f_X(t) - \nu g_X(\nu t)}{u_X(\nu t) - v_X(t)}\, dt \leq \int_0^\infty \rho_{U,n}(t)\, dt \tag{321}$$

$$\leq \int_0^\infty \frac{\nu f_X(\nu t) - g_X(t)}{u_X(t) - v_X(\nu t)}\, dt. \tag{322}$$

Next, we further develop these bounds. For any $s \in (0,1)$, denote

$$\nu_{X,n,s} := \left(\frac{1 - s\mu_X}{1-s}\right)^{1/d_n}. \tag{323}$$

Consider the functions

$$\varphi_X(t;\nu) := \frac{u_X(t) - v_X(t)}{u_X(t) - v_X(\nu t)}, \tag{324}$$

$$\psi_X(t;\nu) := \frac{u_X(t) - v_X(t)}{u_X(\nu t) - v_X(t)}. \tag{325}$$

We show in Appendix H-B that, for any constants $s \in (0, (1-\mu_X)/(1+\mu_X))$ and $1 \leq \nu \leq \nu_{X,n,s}$, the uniform bounds

$$1 - s \leq \psi_X(t;\nu) \leq 1 \leq \varphi_X(t;\nu) \leq 1 + s \tag{326}$$

hold over $t \in [0,\infty)$. Fix $s \in (0, (1-\mu_X)/(1+\mu_X))$ and $1 \leq \nu \leq \nu_{X,n,s}$.

Now, the integrand in the upper bound in (322) can be rewritten as

$$\frac{\nu f_X(\nu t) - g_X(t)}{u_X(t) - v_X(\nu t)} = \varphi_X(t;\nu)\left(\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + \frac{\nu f_X(\nu t) - f_X(t)}{u_X(t) - v_X(t)}\right). \tag{327}$$

The integrand in the lower bound in (321) can be rewritten as

$$\frac{f_X(t) - \nu g_X(\nu t)}{u_X(\nu t) - v_X(t)} = \psi_X(t;\nu)\left(\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + \frac{g_X(t) - \nu g_X(\nu t)}{u_X(t) - v_X(t)}\right). \tag{328}$$

By the bounds in (326), we have that for every $t \geq 0$

$$0 \leq \varphi_X(t;\nu) - 1 \leq s. \tag{329}$$

Hence, by non-negativity of f_X and g_X, we deduce

$$(\varphi_X(t;\nu) - 1) \cdot \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \leq s \cdot \frac{f_X(t)}{u_X(t) - v_X(t)}, \tag{330}$$

i.e., the inequality

$$\varphi_X(t;\nu)\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \leq \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} + s\frac{f_X(t)}{u_X(t) - v_X(t)} \tag{331}$$

hold of all $t \geq 0$. In addition, since $f_X(\nu t) \leq \nu^{d_n - 2} f_X(t)$ over $t \in [0,\infty)$, inequality (329) implies that

$$\varphi_X(t;\nu) \cdot \frac{\nu f_X(\nu t) - f_X(t)}{u_X(t) - v_X(t)} \leq \frac{(1+s)(\nu^{d_n - 1} - 1) f_X(t)}{u_X(t) - v_X(t)}. \tag{332}$$

Therefore, applying inequalities (331) and (332) in formula (327), we deduce in view of the upper bound in (322) the inequality

$$\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\, dt \leq \left((1+s)\nu^{d_n - 1} - 1\right) \int_0^\infty \frac{f_X(t)}{u_X(t) - v_X(t)}\, dt. \tag{333}$$

Similarly, we derive a lower bound on (328). By (326), we have that for every $t \geq 0$

$$s \geq 1 - \psi_X(t;\nu) \geq 0. \tag{334}$$

Hence, by non-negativity of f_X and g_X,

$$s \cdot \frac{f_X(t)}{u_X(t) - v_X(t)} \geq (1 - \psi_X(t;\nu)) \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)}, \tag{335}$$

i.e., the inequality

$$\psi_X(t;\nu)\frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} \geq \frac{f_X(t) - g_X(t)}{u_X(t) - v_X(t)} - s\frac{f_X(t)}{u_X(t) - v_X(t)} \tag{336}$$

holds for all $t \geq 0$. In addition, from $\psi_X(t;\nu) \leq 1 \leq \nu$ and $g_X(\nu t) \leq \nu^{d_n - 2} g_X(t)$ for $t \geq 0$, we deduce

$$\psi_X(t;\nu) \cdot \frac{g_X(t) - \nu g_X(\nu t)}{u_X(t) - v_X(t)} \geq \psi_X(t;\nu) \cdot \frac{(1 - \nu^{d_n - 1}) g_X(t)}{u_X(t) - v_X(t)}$$

$$\geq \left(1 - \nu^{d_n - 1}\right) \frac{g_X(t)}{u_X(t) - v_X(t)}. \tag{337}$$

Therefore, applying inequalities (336) and (337) in formula (328), the lower bound in (321) yields the bound

$$\begin{aligned}\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt \\ \geq -s\int_0^\infty \frac{f_X(t)}{u_X(t)-v_X(t)}\,dt \\ -\left(\nu^{d_n-1}-1\right)\int_0^\infty \frac{g_X(t)}{u_X(t)-v_X(t)}\,dt.\end{aligned} \tag{338}$$

In particular, (338) implies that

$$\begin{aligned}\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt \\ \geq -\left(\nu^{d_n-1}-(1-s)\right)\int_0^\infty \frac{f_X(t)+g_X(t)}{u_X(t)-v_X(t)}\,dt.\end{aligned} \tag{339}$$

Now, note that $(1+s)\nu^{d_n-1}-1 \geq \nu^{d_n-1}-(1-s)$. Therefore, combining the upper bound in (333) and the lower bound in (339), we deduce that

$$\begin{aligned}\left|\int_0^\infty \rho_{U,n}(t) - \rho_{X,n}(t)\,dt\right| \\ \leq \left((1+s)\nu^{d_n-1}-1\right)\int_0^\infty \frac{f_X(t)+g_X(t)}{u_X(t)-v_X(t)}\,dt.\end{aligned} \tag{340}$$

The upper bound in (340) may be made as small as needed by choosing a small s then choosing a small ν.

The second part of the proof, given in Appendix H-C, derives the following error bound for estimating $\log\det \boldsymbol{M}_{X,n}$ from samples. If $B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}) > 0$, we denote

$$\tau_{X,n} := \left(\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})/B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})+1}{2}\right)^{1/(n+1)} \in (1,\infty) \tag{341}$$

and

$$\eta_{X,n} := \min\left(\frac{1}{2}, \frac{\tau_{X,n}-1}{\tau_{X,n}+1}\right) \in (0,1/2]. \tag{342}$$

If $B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}}) = 0$, then we set $\tau_{X,n} = \infty$ and $\eta_{X,n} = 1/2$. We show that for all $\eta \in (0,\eta_{X,n})$, if $\mathfrak{A}_{n,\eta}(\mathcal{S})$ holds, then we have the bound

$$\left|\frac{1}{2d_n}\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| \leq \frac{6\eta}{n}\cdot\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})+B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})-B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}. \tag{343}$$

To finish the proof, we choose η so that the desired accuracy is achieved with high probability. Recall from (310) that

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \geq 1 - 4ne^{-m\eta^2\alpha_{X,n}} \tag{344}$$

where we denote the constant

$$\alpha_{X,n} := 2\cdot\left(\left(\frac{q}{p}\right)^{2n}-1\right)^{-2}. \tag{345}$$

In addition, from (340) and (343), we know that if $s \in (0,(1-\mu_X)/(1+\mu_X))$, $\nu \in [1,\nu_{X,n,s}]$, $\eta \in (0,\eta_{X,n})$, and $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs, then

$$\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \leq \eta\cdot\beta_{X,n} + \left((1+s)\nu^{d_n-1}-1\right)\cdot\gamma_{X,n} \tag{346}$$

where we denote the constants

$$\beta_{X,n} := \frac{6}{n}\cdot\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})+B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})-B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}, \tag{347}$$

$$\gamma_{X,n} := \int_0^\infty \frac{f_X(t)+g_X(t)}{u_X(t)-v_X(t)}\,dt. \tag{348}$$

Consider the constant $\varepsilon_{X,n} \in (0, 2\min(\gamma_{X,n},\beta_{X,n})]$ defined by

$$\varepsilon_{X,n} := \min\left(2\gamma_{X,n}\cdot\frac{1-\mu_X}{1+\mu_X},\ 2\beta_{X,n}\right). \tag{349}$$

Fix $\varepsilon \in (0,\varepsilon_{X,n})$, set $s := \varepsilon/(6\gamma_{X,n}) \in (0,1/3]$, denote

$$\kappa_{X,n} := \min\left(3, \tau_{X,n}, \left(\frac{1-s\mu_X}{1-s}\right)^{\frac{1}{2d_n}}, \frac{1+\varepsilon/(2\beta_{X,n})}{1-\varepsilon/(2\beta_{X,n})}\right), \tag{350}$$

and fix $\eta \in (0,(\kappa_{X,n}-1)/(\kappa_{X,n}+1))$. Since $\kappa_{X,n} \leq 3$, we obtain $\eta < 1/2$. In addition, $\kappa_{X,n} \leq \tau_{X,n}$, hence $\eta < (\kappa_{X,n}-1)/(\kappa_{X,n}+1)$ implies that $\eta < \eta_{X,n}$. Note that, for $a \in (0,1)$ and $b > 1$, the inequality $a \leq (b-1)/(b+1)$ is equivalent to $(1+a)/(1-a) \leq b$. By definition,

$$\kappa_{X,n} \leq \left(\frac{1-s\mu_X}{1-s}\right)^{1/(2d_n)}, \tag{351}$$

hence we have

$$(1+s)\nu^{d_n} = (1+s)\left(\frac{1+\eta}{1-\eta}\right)^{2d} < (1+s)\kappa_{X,n}^{2d} \tag{352}$$

$$\leq (1+s)\cdot\frac{1-s\mu_X}{1-s} \leq \frac{1+s}{1-s} \tag{353}$$

$$\leq \frac{1+s+s(1-3s)}{1-s} = 1+3s. \tag{354}$$

In addition, since

$$\kappa_{X,n} \leq \frac{1+\varepsilon/(2\beta_{X,n})}{1-\varepsilon/(2\beta_{X,n})}, \tag{355}$$

and since we assume $\eta < (\kappa_{X,n}-1)/(\kappa_{X,n}+1)$, we deduce the inequality $\eta < \varepsilon/(2\beta_{X,n})$. Applying the two inequalities $\eta < \varepsilon/(2\beta_{X,n})$ and $(1+s)\nu^{d_n} \leq 1+3s$ (see (354)) into inequality (346), we conclude that

$$\begin{aligned}\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| &\leq \eta\cdot\beta_{X,n} + \left((1+s)\nu^{d_n-1}-1\right)\cdot\gamma_{X,n} \\ &\leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon\end{aligned} \tag{356}$$

whenever $\mathfrak{A}_{n,\eta}(\mathcal{S})$ occurs.

Now, fix $\delta \in (0,1/(4n))$. Set

$$\eta := \frac{1}{2}\cdot\frac{\kappa_{X,n}-1}{\kappa_{X,n}+1}. \tag{357}$$

We show that $\eta \geq \varepsilon c_{X,n}$, where we define the constant $c_{X,n}$ by

$$c_{X,n} := \min\left(\frac{1}{8\gamma_{X,n}}, \frac{\tau_{X,n}-1}{4\gamma_{X,n}(\tau_{X,n}+1)}, \frac{1-\mu_X}{72\gamma_{X,n}d_n}, \frac{1}{4\beta_{X,n}}\right). \tag{358}$$

In this definition of $c_{X,n}$, the term involving $\tau_{X,n}$ is removed if $\tau_{X,n} = \infty$. We assume that

$$m \geq \frac{2/(c_{X,n}^2\alpha_{X,n})}{\varepsilon^2}\log\frac{1}{\delta}. \tag{359}$$

From $\eta \geq \varepsilon c_{X,n}$ and (359), it follows that the probability that the event $\mathfrak{A}_{n,\eta}(\mathcal{S})$ does not occur is bounded as

$$P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})^c\right) \leq 4ne^{-m\eta^2\alpha_{X,n}} \leq \delta. \tag{360}$$

Note that this would conclude the proof, as then we would have that

$$\begin{aligned} &P\left(\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \leq \varepsilon\right) \\ &\geq P\left(\left|\widehat{h}_n(\mathcal{S}) - h_n(X)\right| \leq \varepsilon \;\middle|\; \mathfrak{A}_{n,\eta}(\mathcal{S})\right) P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) \\ &= P\left(\mathfrak{A}_{n,\eta}(\mathcal{S})\right) > 1 - \delta. \end{aligned} \tag{361}$$

The rest of the proof is devoted to showing that $\eta \geq \varepsilon c_{X,n}$ holds.

Let $\rho = (1-\mu_X)/(6d_n)$. We will show that

$$\left(\frac{1-s\mu_X}{1-s}\right)^{1/(2d_n)} \geq \frac{1+s\rho}{1-s\rho}. \tag{362}$$

Inequality (362) is equivalent to

$$(1-s\mu_X)(1-s\rho)^{2d_n} \geq (1+\rho s)^{2d_n}(1-s). \tag{363}$$

By Bernoulli's inequality, since $0 \leq s\rho \leq 1$, we have that $(1-s\rho)^{2d_n} \geq 1-2d_n\rho s$. In addition, the inequality $1+2az \geq e^{az} \geq (1+a)^z$ for $a, z \geq 0$ satisfying $az \leq \log 2$ implies, in view of $2d_n\rho s \leq 1/9 < \log 2$, that

$$1 + 4d_n\rho s \geq (1+\rho s)^{2d_n}. \tag{364}$$

Therefore, to show (363), it suffices to show that

$$(1-s\mu_X)(1-2d_n\rho s) \geq (1+4d_n\rho s)(1-s). \tag{365}$$

Now, using the definition $\rho = (1-\mu_X)/(6d_n)$, inequality (365) follows as

$$\begin{aligned} &(1-s\mu_X)(1-2d_n\rho s) \\ &= (1-s\mu_X)(1-s(1-\mu_X)/3) \\ &= (1+2(1-\mu_X)s/3)(1-s) + s^2(1-\mu_X)(\mu_X+2)/3 \\ &\geq (1+2(1-\mu_X)s/3)(1-s) = (1+4d_n\rho s)(1-s). \end{aligned}$$

Since (365) holds, we conclude that inequality (362) holds.

Now, by the definition of $\kappa_{X,n}$ in (350) there are four possible values $\kappa_{X,n}$ can take. First, if $\kappa_{X,n} = 3$, then

$$\eta = \frac{1}{4} = \varepsilon \cdot \frac{1}{4\varepsilon} \geq \varepsilon \cdot \frac{1}{8\gamma_{X,n}} \geq \varepsilon c_{X,n} \tag{366}$$

since $\varepsilon < \varepsilon_{X,n} \leq 2\gamma_{X,n}$. Now, if $\kappa_{X,n} = \tau_{X,n}$ (so $B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}) > 0$), then

$$\eta = \frac{1}{2} \cdot \frac{\tau_{X,n}-1}{\tau_{X,n}+1} \geq \frac{\varepsilon}{4\gamma_{X,n}} \cdot \frac{\tau_{X,n}-1}{\tau_{X,n}+1} \tag{367}$$

since $\varepsilon < 2\gamma_{X,n}$. Next, suppose that

$$\kappa_{X,n} = \left(\frac{1-s\mu_X}{1-s}\right)^{1/(2d_n)}. \tag{368}$$

By (362) and (368), we deduce that

$$\kappa_{X,n} \geq \frac{1+s\rho}{1-s\rho}. \tag{369}$$

Recall that, for $0 < a < 1 < b$, the inequalities $(1+a)/(1-a) \leq b$ and $(b-1)/(b+1) \geq a$ are equivalent. Therefore, the definition of η in (357) yields from (369) that $\eta \geq s\rho/2$. Plugging in the definitions of s and ρ, we conclude that

$$\eta \geq \varepsilon \cdot \frac{1-\mu_X}{72\gamma_{X,n}d_n} \geq \varepsilon c_{X,n}. \tag{370}$$

Finally, when

$$\kappa_{X,n} = \frac{1+\varepsilon/(2\beta_{X,n})}{1-\varepsilon/(2\beta_{X,n})}, \tag{371}$$

the definition of η implies that $\eta \geq \varepsilon/(4\beta_{X,n}) \geq \varepsilon c_{X,n}$. Combining these four cases, we conclude that we must have $\eta \geq \varepsilon c_{X,n}$ independently of the value of $\kappa_{X,n}$. The proof is thus complete.

B. Proof of the Uniform Bounds (326)

Being polynomials of degree at most d_n with non-negative coefficients, the functions u_X and v_X satisfy $u_X(\nu t) \leq \nu^{d_n}u_X(t)$ and $v_X(\nu t) \leq \nu^{d_n}v_X(t)$ for every $\nu \geq 1$ and $t \geq 0$. Note also that both u_X and v_X are nondecreasing. In addition, we have $v_X(t) < u_X(t)$ for every $t \geq 0$, because $u_X(t) - v_X(t) = \det \boldsymbol{M}_{\sqrt{t}X+N,n} > 0$. We have also shown that $\mu_X < 1$, where μ_X is defined in (314) as

$$\mu_X := \sup_{t\in[0,\infty)} \frac{v_X(t)}{u_X(t)}. \tag{372}$$

These facts will be enough to deduce the bounds in (326).

We show first the bounds on φ_X in (326). It suffices to consider the case $\mu_X > 0$, for otherwise v_X vanishes identically and $\varphi_X \equiv 1$ identically. We show that for every $s > 0$ and $1 \leq \nu \leq \nu'_{X,n,s}$, where $\nu'_{X,n,s} := ((1/s + 1/\mu_X)/(1/s+1))^{1/d_n}$, the uniform bound $1 \leq \varphi_X(t;\nu) \leq 1+s$ in (326) holds.

Consider the lower bound on φ_X. For every $1 \leq \nu < \mu_X^{-1/d_n}$, we have the uniform bound

$$\frac{v_X(\nu t)}{u_X(t)} \leq \frac{\nu^{d_n}v_X(t)}{u_X(t)} \leq \nu^{d_n}\mu_X < 1 \tag{373}$$

over $t \in [0,\infty)$. In particular,

$$u_X(t) - v_X(\nu t) > 0 \tag{374}$$

for every $1 \leq \nu < \mu_X^{-1/d_n}$ and $t \geq 0$. Since v_X is nondecreasing, we conclude that $\varphi_X(t;\nu) = (u_X(t) - v_X(t))/(u_X(t) - v_X(\nu t)) \geq 1$ whenever $1 \leq \nu < \mu_X^{-1/d_n}$. Note that $\nu'_{X,n,s} < \mu_X^{-1/d_n}$ for every $s > 0$ since $\mu_X \in (0,1)$.

Next, we show the upper bound on φ_X. Fix $s > 0$ and $\nu \in [1, \nu'_{X,n,s}]$. Since $v_X(t)/\mu_X \leq u_X(t)$, we have for every $t \geq 0$ the bound

$$v_X(\nu t) \leq \nu^{d_n}v_X(t) \leq \frac{1/s + 1/\mu_X}{1/s+1} \cdot v_X(t) \tag{375}$$

$$\leq \frac{v_X(t)/s + u_X(t)}{1/s+1} = v_X(t) + \frac{u_X(t) - v_X(t)}{1/s+1}. \tag{376}$$

Rearranging (376), we obtain the bound

$$\frac{-1}{1/s+1} \leq \frac{v_X(t) - v_X(\nu t)}{u_X(t) - v_X(t)}. \tag{377}$$

Adding 1 to both sides of (377) then inverting, we obtain $\varphi_X(t;\nu) \leq 1+s$; for this step, we used the fact that $u_X(t) -$

$v_X(\nu t) > 0$, which follows by (374) since $\nu \le \nu'_{X,n,s} < \mu_X^{-1/d_n}$.

Next, we prove the bounds on ψ_X in (326). We do not assume $\mu_X > 0$. The upper bound $\psi_X(t;\nu) \le 1$ follows for every $\nu \ge 1$ by monotonicity of u_X. For the lower bound on ψ_X, we show that for every $s \in (0,1)$ and $1 \le \nu \le \nu_{X,n,s}$, where $\nu_{X,n,s} := ((1-s\mu_X)/(1-s))^{1/d_n}$, the uniform bound $\psi_X(t;\nu) \ge 1-s$ holds over $t \in [0,\infty)$. We have, for every $s \in (0,1)$ and $\nu \in [1, \nu_{X,n,s}]$, the bound

$$u_X(\nu t) \le \nu^{d_n} u_X(t) \le \frac{1-s\mu_X}{1-s} \cdot u_X(t) \tag{378}$$

$$\le \frac{u_X(t) - s v_X(t)}{1-s} = \frac{u_X(t) - v_X(t)}{1-s} + v_X(t) \tag{379}$$

over $t \in [0,\infty)$. Rearranging (379), we obtain $\psi_X(t;\nu) \ge 1-s$, as desired.

Finally, note that $\nu_{X,n,s} \le \nu'_{X,n,s}$ is equivalent to $s \le (1-\mu_X)/(1+\mu_X)$. This concludes the proof that, for every $s \in (0, (1-\mu_X)/(1+\mu_X))$ and $\nu \in [1, \nu_{X,n,s}]$, the uniform bounds in (326)

$$1 - s \le \psi_X(t;\nu) \le 1 \le \varphi_X(t;\nu) \le 1 + s \tag{380}$$

hold over $t \in [0,\infty)$.

C. Proof of Inequality (343)

Recall that

$$\det \boldsymbol{M}_{X,n} = B_{d_n}(\boldsymbol{\mathcal{X}}) = B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}). \tag{381}$$

We bound the error when estimating $\log\det \boldsymbol{M}_{X,n}$ from the samples $\mathcal{S}$. Denote the random vector $\boldsymbol{\mu} := \left(\frac{\sum_{i=1}^m X_i}{m}, \cdots, \frac{\sum_{i=1}^m X_i^{2n}}{m}\right)$, and note that

$$\det \boldsymbol{M}_{U,n} = B_{d_n}(\boldsymbol{\mu}) = B^{(+)}_{d_n}(\boldsymbol{\mu}) - B^{(-)}_{d_n}(\boldsymbol{\mu}). \tag{382}$$

We assume that $m > n$. Let $\eta_{X,n}$ be as defined by (341) and (342), and fix $\eta \in (0, \eta_{X,n})$. Then we show that under $\mathfrak{A}_{n,\eta}(\mathcal{S})$

$$\left|\frac{1}{2d_n}\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| \le \frac{6\eta}{n} \cdot \frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) + B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}. \tag{383}$$

By (179) in the proof of Theorem 2, each term in the polynomials $B^{(\pm)}_{d_n}$ is a product of at most $n+1$ monomials. Thus,

$$(1-\eta)^{n+1} B^{(\pm)}_{d_n}(\boldsymbol{\mathcal{X}}) \le B^{(\pm)}_{d_n}(\boldsymbol{\mu}) \le (1+\eta)^{n+1} B^{(\pm)}_{d_n}(\boldsymbol{\mathcal{X}}). \tag{384}$$

It suffices to consider the case when $B^{(-)}_{d_n}$ is not the zero polynomial, for if $B^{(-)}_{d_n}$ is the zero polynomial then we obtain from (384) the bound

$$\left|\frac{1}{2d_n}\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| = \frac{1}{2d_n}\left|\log\frac{B^{(+)}_{d_n}(\boldsymbol{\mu})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})}\right| \tag{385}$$

$$\le \frac{\max\left(\pm\log(1\pm\eta)\right)}{n} \tag{386}$$

$$= \frac{-\log(1-\eta)}{n} < \frac{2\eta}{n} \tag{387}$$

where the last inequality follow because $-\log(1-z) < 2z$ for $z \in (0, 1/2)$, which can be verified by checking the derivative. Note that the bound $2\eta/n$ in (387) is stronger than the bound in (383). Assume that $B^{(-)}_{d_n}$ does not vanish identically, so positivity of X yields that $B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}) > 0$.

From (384), we have that

$$\log\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} + (n+1)\log(1-\eta) \le \log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \tag{388}$$

and

$$\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}} \le \log\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} + (n+1)\log(1+\eta) \tag{389}$$

where we used our assumption that

$$\nu^{\frac{n+1}{2}} = \left(1 + \frac{2}{1/\eta - 1}\right)^{n+1} < \frac{1}{2}\left(\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} + 1\right) \tag{390}$$

$$< \frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}. \tag{391}$$

Now, for every $(w,z,r) \in \mathbb{R}^3$ such that $w > z > 0$ and $w/z > r > 1$, rearranging $r + 1/r > 2$ we have that

$$\frac{w - z/r}{w - z} < \frac{w - z}{w - rz}. \tag{392}$$

Setting $(w,z,r) = (B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}), B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}}), \nu^{(n+1)/2})$, we obtain that

$$1 < \frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} \tag{393}$$

$$< \frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}. \tag{394}$$

Therefore,

$$0 < \log\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{-\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} \tag{395}$$

$$< \left|\log\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}\right|. \tag{396}$$

Applying (395)–(396) in (389) and combining that with (388), we obtain (since $\log(1+\eta) < -\log(1-\eta)$) the bound

$$\left|\log\frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| \le \log\frac{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})}{B^{(+)}_{d_n}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B^{(-)}_{d_n}(\boldsymbol{\mathcal{X}})} + (n+1)\log\frac{1}{1-\eta}. \tag{397}$$

Now, we may write

$$\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} = \left(1 - \frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \left(\nu^{\frac{n+1}{2}} - 1\right)\right)^{-1}. \tag{398}$$

The proof of (383) (or, (343)) is completed by showing that for $(w, z, r) \in \mathbb{R}^3_{>0}$ such that $(1+z)^r < 1 + \frac{1}{2w}$ we have

$$-\log\left(1 - w\left((1+z)^r - 1\right)\right) \leq (2w+1)rz. \tag{399}$$

Before showing that (399) holds, we note how it completes the proof. Setting

$$(w, z, r) = \left(\frac{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}, \frac{2\eta}{1-\eta}, n+1\right), \tag{400}$$

we obtain that

$$\log \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - \nu^{\frac{n+1}{2}} B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \leq \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \cdot (n+1) \cdot \frac{2\eta}{1-\eta} \tag{401}$$

since (see (390))

$$\nu^{\frac{n+1}{2}} < \frac{1}{2}\left(\frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} + 1\right). \tag{402}$$

Then $-\log(1-\eta) < 2\eta$ yields from (397) and (401) that

$$\frac{1}{2d_n}\left|\log \frac{\det \boldsymbol{M}_{U,n}}{\det \boldsymbol{M}_{X,n}}\right| \leq \frac{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) + B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})}{B_{d_n}^{(+)}(\boldsymbol{\mathcal{X}}) - B_{d_n}^{(-)}(\boldsymbol{\mathcal{X}})} \cdot \frac{2\eta}{n(1-\eta)} + \frac{2\eta}{n}. \tag{403}$$

Then (403) yields the desired inequality (343) as $\eta \in (0, 1/2)$.

Finally, to see that (399) holds, we consider for fixed $w, r > 0$

$$f(z) := (2w+1)rz + \log\left(1 - w\left((1+z)^r - 1\right)\right) \tag{404}$$

over $0 \leq z < (1+1/(2w))^{1/r} - 1$. Inequality (399) is restated as $f(z) \geq 0$ for every $0 < z < (1+1/(2w))^{1/r} - 1$, which follows since f is continuous, $f(0) = 0$, $f'(0^+) = (w+1)r > 0$, and

$$f'(z) = (2w+1)r - \frac{wr(1+z)^{r-1}}{1 - w((1+z)^r - 1)} \tag{405}$$

$$> (2w+1)r - \frac{wr(1+z)^r}{1 - w((1+z)^r - 1)} \tag{406}$$

$$> (2w+1)r - \frac{wr(1+1/(2w))}{1 - w((1+1/(2w)) - 1)} = 0 \tag{407}$$

for every $0 \leq z < (1+1/(2w))^{1/r} - 1$.

D. Proof of Proposition 5: Mutual Information

Let $\{(X_j, Y_j)\}_{j\in\mathbb{N}}$ be i.i.d. samples drawn according to $P_{X,Y}$. Denote $\mathcal{S}_m = \{X_j\}_{j=1}^m$. By continuity of Y, we may assume that all the Y_j, for $j \in \mathbb{N}$, are distinct. For each $x \in \mathrm{supp}(X)$, let $J_x := \{1 \leq j \leq m \ ; \ X_j = x\}$. Let $\mathfrak{D}_m$ be the event that, for every $x \in \mathrm{supp}(X)$, we have that $|J_x| > n$. We use Hoeffding's inequality to obtain a lower bound on the probability

$$P(\mathfrak{D}_m) = P\left(\min_{x\in\mathrm{supp}(X)} |J_x| > n\right). \tag{408}$$

Let $\widehat{P}_m$ be the empirical measure, i.e., define $\widehat{P}_m(x) := m^{-1}\sum_{j=1}^m \delta_x(X_j)$. Note that $|J_x| = m\widehat{P}_m(x)$.

Let $x_0 \in \mathrm{supp}(X)$ be such that $P_X(x_0)$ is minimal, set $\zeta := P_X(x_0)/2$, and suppose $m \geq \zeta^{-1}n$. Then, the union bound and $\zeta \leq P_X(x) - \zeta$ for each $x \in \mathrm{supp}(X)$ yield that

$$P\left(n \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right)$$

$$\leq P\left(m\zeta \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right) \tag{409}$$

$$\leq \sum_{x\in\mathrm{supp}(X)} P\left(m\zeta \geq |J_x|\right) \tag{410}$$

$$\leq \sum_{x\in\mathrm{supp}(X)} P\left(m(P_X(x) - \zeta) \geq |J_x|\right) \tag{411}$$

$$= \sum_{x\in\mathrm{supp}(X)} P\left(P_X(x) - \widehat{P}_m(x) \geq \zeta\right). \tag{412}$$

Since $\mathbb{E}[\widehat{P}_m(x)] = P_X(x)$ for each $x \in \mathrm{supp}(X)$, Hoeffding's inequality yields that $P\left(P_X(x) - \widehat{P}_m(x) \geq \zeta\right) \leq e^{-2\zeta^2 m}$. Therefore,

$$P\left(n \geq \min_{x\in\mathrm{supp}(X)} |J_x|\right) \leq |\mathrm{supp}(X)| \cdot e^{-2\zeta^2 m}. \tag{413}$$

In other words, for every $m \geq 2n/P_X(x_0)$, we have the bound

$$P(\mathfrak{D}_m) \geq 1 - |\mathrm{supp}(X)| \cdot e^{-mP_X(x_0)^2/2}. \tag{414}$$

Denote $\pi_X := 4/P_X(x_0)^2$ and

$$\delta_{X,n} := \min\left(\frac{1}{4|\mathrm{supp}(X)|}, e^{-P_X(x_0)n/2}\right). \tag{415}$$

We conclude from (414) that, for every $\delta \in (0, \delta_{X,n})$, if $m \geq \pi_X \log(1/\delta)$ then $P(\mathfrak{D}_m) > 1 - \delta/4$.

Consider the event $\mathfrak{P}_{m,\varepsilon}$ that the empirical measure $\widehat{P}_m$ is pointwise ε-close to the true measure P_X, i.e.,

$$\mathfrak{P}_{m,\varepsilon} := \left\{\max_{x\in\mathrm{supp}(X)} \left|\widehat{P}_m(x) - P_X(x)\right| < \varepsilon\right\}. \tag{416}$$

By the union bound, we have that

$$P\left(\mathfrak{P}_{m,\varepsilon}^c\right) \leq \sum_{x\in\mathrm{supp}(X)} P\left(\left|\widehat{P}_m(x) - P_X(x)\right| \geq \varepsilon\right). \tag{417}$$

By Hoeffding's inequality, for each $x \in \mathrm{supp}(X)$, we have that

$$P\left(\left|\widehat{P}_m(x) - P_X(x)\right| \geq \varepsilon\right) \leq 2e^{-2m\varepsilon^2}. \tag{418}$$

Therefore, we obtain the bound

$$P\left(\mathfrak{P}_{m,\varepsilon}\right) > 1 - 2|\mathrm{supp}(X)|e^{-2m\varepsilon^2}. \tag{419}$$

In particular, if $\delta \in (0, 1/(4|\mathrm{supp}(X)|))$, then $m \geq (1/\varepsilon^2)\log(1/\delta)$ implies $P(\mathfrak{P}_{m,\varepsilon}) > 1 - \delta/2$.

Recall that, if $\mathfrak{D}_m$ occurs, then we may write

$$\widehat{I}_n(\mathcal{S}_m) = \widehat{h}_n(\mathcal{A}_m) - \sum_{x\in\mathrm{supp}(X)} \widehat{P}_m(x)\, \widehat{h}_n(\mathcal{B}_{m,x}), \tag{420}$$

where $\mathcal{A}_m := \{Y_j\}_{j=1}^m$ and $\mathcal{B}_{m,x} := \{Y_j \ ; \ 1 \leq j \leq m, X_j = x\}$. Then,

$$\begin{aligned}
&\left|\widehat{I}_n(\mathcal{S}_m) - I_n(X;Y)\right| \\
&\quad\leq \left|\widehat{h}_n(\mathcal{A}_m) - h_n(Y)\right| \\
&\quad + \sum_{x\in\mathrm{supp}(X)} \widehat{P}_m(x)\left|\widehat{h}_n(\mathcal{B}_{m,x}) - h_n(Y^{(x)})\right| \\
&\quad + \left(\max_{x\in\mathrm{supp}(X)}\left|\widehat{P}_m(x) - P_X(x)\right|\right) \sum_{x\in\mathrm{supp}(X)} |h_n(Y^{(x)})|.
\end{aligned} \tag{421}$$

Denote $H_{X,Y,n} := \sum_{x\in\mathrm{supp}(X)} |h_n(Y^{(x)})|$. Consider the events

$$\mathfrak{F}_{x,\varepsilon} := \left\{\left|\widehat{h}_n(\mathcal{B}_{m,x}) - h_n(Y^{(x)})\right| < \frac{\varepsilon}{3}\right\} \tag{422}$$

$$\mathfrak{F}'_{\varepsilon} := \left\{\left|\widehat{h}_n(\mathcal{A}_m) - h_n(Y)\right| < \frac{\varepsilon}{3}\right\}. \tag{423}$$

Set $\mathfrak{F}_\varepsilon := \bigcap_{x\in\mathrm{supp}(X)} \mathfrak{F}_{x,\varepsilon}$. From Proposition 4, we know that there is a constant $C_{X,Y,n}$ such that for every small enough $\varepsilon, \delta > 0$, if $m \geq (C_{X,Y,n}/\varepsilon^2)\log(1/\delta)$ then $P(\mathfrak{F}_{x,\varepsilon} \mid \mathfrak{D}_m) \geq 1 - \delta/(8|\mathrm{supp}(X)|)$ for each $x \in \mathrm{supp}(X)$ and $P(\mathfrak{F}'_\varepsilon \mid \mathfrak{D}_m) > 1 - \delta/8$. Then, $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \mid \mathfrak{D}_m) \geq 1 - \delta/4$. We conclude, possibly after increasing $C_{X,Y,n}$, that $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m) \geq 1 - \delta/2$. Also, $P(\mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}) > 1 - \delta/2$, where we increase $C_{X,Y,n}$, if necessary, to exceed $9H^2_{X,Y,n}$. Then, $P(\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m \cap \mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}) \geq 1 - \delta$. But under the event $\mathfrak{F}_\varepsilon \cap \mathfrak{F}'_\varepsilon \cap \mathfrak{D}_m \cap \mathfrak{P}_{m,\varepsilon/(3H_{X,Y,n})}$, we have from (421) the bound

$$\left|\widehat{I}_n(\mathcal{S}_m) - I_n(X;Y)\right| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon, \tag{424}$$

and the proof is complete.

Acknowledgment

We thank the reviewers and the associate editor for their comments and suggestions. The authors would also like to thank Prof. Shahab Asoodeh (McMaster University) for pointing out that our approach for expressing the mutual information in terms of moments is applicable also to differential entropy. The authors are also grateful to Prof. Alex Dytso (NJIT) for noting that the higher-order derivatives of the conditional expectation in Gaussian channels are expressible in terms of the conditional cumulants.

References

[1] D. Guo, S. Shamai, and S. Verdu, "Mutual information and minimum mean-square error in gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[2] W. Alghamdi and F. P. Calmon, "Polynomial approximations of conditional expectations in scalar gaussian channels," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 420–425.

[3] E. M. Stein and R. Shakarchi, *Real Analysis*. Princeton University Press, 2019.

[4] K. Schmüdgen, *The Moment Problem*. Springer Cham, 2017.

[5] V. S. Adamchik, "On the Barnes function," in *Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation*, ser. ISSAC '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 15–20. [Online]. Available: https://doi.org/10.1145/384101.384104

[6] C. Berg, "On the preservation of determinacy under convolution," *Proceedings of the American Mathematical Society*, vol. 93, no. 2, pp. 351–357, 1985. [Online]. Available: https://doi.org/10.1090/S0002-9939-1985-0770553-4

[7] Y. Wu and S. Verdú, "MMSE dimension," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4857–4879, 2011.

[8] W. Alghamdi and F. P. Calmon, "Measuring information from moments," https://arxiv.org/abs/2109.00649v1, 2021.

[9] M. Zakai, "On mutual information, likelihood ratios, and estimation error for the additive Gaussian channel," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3017–3024, 2005.

[10] D. Guo, "Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation," *IEEE International Symposium on Information Theory - Proceedings*, pp. 814–818, 2009.

[11] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, 2010.

[12] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

[13] Y. Wu and S. Verdu, "Functional properties of minimum mean-square error and mutual information," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.

[14] H. Asnani, K. Venkat, and T. Weissman, "Relations Between Information and Estimation in the Presence of Feedback," *Lecture Notes in Control and Information Sciences*, vol. 450 LNCIS, pp. 157–175, 2014.

[15] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6343–6354, 2015.

[16] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai, "A view of information-estimation relations in Gaussian networks," *Entropy*, vol. 19, no. 8, pp. 1–51, 2017.

[17] A. Dytso and H. Vincent Poor, "Estimation in Poisson noise: Properties of the conditional mean estimator," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4304–4323, 2020.

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

[19] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4295–4297, 2006.

[20] S. Cha and T. Moon, "Neural adaptive image denoiser," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2981–2985.

[21] L. Devroye, D. Schäfer, L. Györfi, and H. Walk, "The estimation problem of minimum mean squared error," *Statistics & Decisions*, vol. 21, no. 1, pp. 15–28, 2003. [Online]. Available: https://doi.org/10.1524/stnd.21.1.15.20315

[22] D. L. Donoho, "One-sided inference about functionals of a density," *The Annals of Statistics*, vol. 16, no. 4, pp. 1390 – 1420, 1988. [Online]. Available: https://doi.org/10.1214/aos/1176351045

[23] M. Diaz, P. Kairouz, and L. Sankar, "Lower bounds for the minimum mean-square error via neural network-based estimation," https://arxiv.org/abs/2108.12851, 2021.

[24] F. d. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1879–1892, 2018.

[25] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[26] Z. Goldfeld, K. Greenewald, J. Weed, and Y. Polyanskiy, "Optimality of the plug-in estimator for differential entropy estimation under Gaussian convolutions," in *IEEE International Symposium on Information Theory - Proceedings*, 2019.

[27] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, 2020.

[28] L. Carleson, "On Bernstein's approximation problem," *Proceedings of the American Mathematical Society*, 1951.

[29] G. Freud, "On Markov-Bernstein-type inequalities and their applications," *Journal of Approximation Theory*, 1977.

[30] Z. Ditzian and V. Totik, *Moduli of Smoothness*. Springer New York, 1987.

[31] D. Lubinsky, "A survey of weighted polynomial approximation with exponential weights," *Surveys in Approximation Theory*, vol. 3, pp. 1–105, 2007.

[32] C. Berg and J. P. R. Christensen, "Density questions in the classical theory of moments," *Annales de l'institut Fourier*, 1981.

[33] A. Makur and L. Zheng, "Polynomial singular value decompositions of a family of source-channel models," *IEEE Transactions on Information Theory*, vol. 63, no. 12, pp. 7716–7728, 2017.

[34] P. Forrester and S. Warnaar, "The importance of the selberg integral," *Bulletin of the American Mathematical Society*, vol. 45, no. 4, pp. 489–534, 2008.

[35] T. Scharf, J. Thibon, and B. G. Wybourne, "Powers of the Vandermonde determinant and the quantum Hall effect," *Journal of Physics A: General Physics*, vol. 27, no. 12, pp. 4211–4219, 1994.

[36] R. C. King, F. Toumazet, and B. G. Wybourne, "The square of the Vandermonde determinant and its q-generalization," *Journal of Physics A: Mathematical and General*, vol. 37, no. 3, pp. 735–767, 2004.

[37] A. Dytso, H. V. Poor, and S. S. Shitz, "A general derivative identity for the conditional mean estimator in gaussian noise and some applications," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1183–1188.

[38] A. Dytso, H. V. Poor, and S. Shamai (Shitz), "A general derivative identity for the conditional mean estimator in Gaussian noise and some applications," https://arxiv.org/abs/2104.01883, 2021.

[39] H. Goodarzi, H. S. Najafabadi, P. Oikonomou, T. M. Greco, L. Fish, R. Salavati, I. M. Cristea, and S. Tavazoie, "Systematic discovery of structural elements governing stability of mammalian messenger RNAs," *Nature*, 2012.

[40] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone, "The transcriptional network for mesenchymal transformation of brain tumours," *Nature*, 2010.

[41] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, 2004.

[42] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 69, no. 6, p. 16, 2004.

[43] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 685–694, 2011.

[44] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[45] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6,

pp. 3702–3720, 2016.

[46] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5987–5998, 2017.

[47] R. E. Curto and L. A. Fialkow, "Recursiveness, positivity, and truncated moment problems," *Houston J. Math.*, no. 4, pp. 603–635, 1991. [Online]. Available: https://www.math.uh.edu/~hjm/v017n4/0603CURTO.pdf

[48] V. I. Bogachev, *Measure theory*. Berlin: Springer, 2007.

[49] OEIS Foundation Inc. (2021), The On-Line Encyclopedia of Integer Sequences. [Online]. Available: http://oeis.org/A032181

[50] Z. Ditzian and D. S. Lubinsky, "Jackson and smoothness theorems for Freud weights in L_p $(0 < p \leq \infty)$," *Constructive approximation*, vol. 13, no. 1, pp. 99–152, 1997.

[51] H. N. Mhaskar, "Introduction to the theory of weighted polynomial approximation," in *Series in approximations and decompositions*, 1996.

[52] A. Dytso and M. Cardone, "A general derivative identity for the conditional expectation with focus on the exponential family," in *2021 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–6.

[53] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *The Annals of Statistics*, vol. 48, no. 6, pp. 3228 – 3250, 2020. [Online]. Available: https://doi.org/10.1214/19-AOS1927

[54] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[55] W. Alghamdi and F. P. Calmon, "Measuring Information from Moments," https://github.com/WaelAlghamdi/MIE, 2021.

[56] G. V. Steeg, "NPEET," https://github.com/gregversteeg/NPEET, 2014.

[57] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[58] W. Cao, A. Dytso, M. Fauß, and H. V. Poor, "Finite-sample bounds on the accuracy of plug-in estimators of Fisher information," *Entropy*, vol. 23, no. 5, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/5/545

[59] D. Guo, S. Shamai, and S. Verdu, "Mutual information and conditional mean estimation in Poisson channels," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1837–1849, 2008.

[60] B. Simon, "The classical moment problem as a self-adjoint finite difference operator," *Advances in Mathematics*, vol. 137, no. 1, pp. 82–203, 1998. [Online]. Available: https://doi.org/10.1006/aima.1998.1728

[61] B. Bényi and J. L. Ramírez, "Some applications of S-restricted set partitions," *Periodica mathematica Hungarica*, vol. 78, no. 1, pp. 110–127, 2019. [Online]. Available: https://doi.org/10.1007/s10998-018-0252-1

[62] L. C. Petersen, "On the relation between the multidimensional moment problem and the one-dimensional moment problem," *Mathematica Scandinavica*, vol. 51, pp. 361–366, Jun. 1982. [Online]. Available: https://doi.org/10.7146/math.scand.a-11986

Wael Alghamdi Wael Alghamdi received the S.B. degree in mathematics from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2014, and the M.S. degree in electrical engineering from King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, in 2016. Since 2017, he has been a graduate student with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. His current research interest focuses on functional properties of information measures, and applications to estimation from data, fairness, and differential privacy.

Flavio P. Calmon Flavio P. Calmon is an Assistant Professor of Electrical Engineering at Harvard's John A. Paulson School of Engineering and Applied Sciences. Before joining Harvard, he was the inaugural Data Science for Social Good Post-Doctoral Fellow at IBM Research in Yorktown Heights, New York. He received his Ph.D. in Electrical Engineering and Computer Science at MIT. His research develops information-theoretic tools for responsible, reliable, and rigorous machine learning. Prof. Calmon has received the NSF CAREER award, faculty awards from Google, IBM, Oracle, and Amazon, the NSF-Amazon Fairness in AI award, and the Harvard Data Science Initiative Bias2 award, among other grants. He also received the inaugural Título de Honra ao Mérito (Honor to the Merit Title) given to alumni from the Universidade de Brasília, being the first awardee from engineering and computer science.