Mutual Information as a Function of Moments

Wael Alghamdi and Flavio P. Calmon
Harvard University
alghamdi@g.harvard.edu, flavio@seas.harvard.edu

***Abstract*—We introduce a mutual information estimator based on the connection between estimation theory and information theory. By combining a polynomial approximation to the minimum mean-squared error estimator with the I-MMSE relationship, we derive a new formula for the mutual information $I(X;Y)$ that is a function of only the marginal distribution of X, the moments of Y, and the conditional moments of Y given X. The proposed estimator captures desirable properties that the mutual information satisfies, such as being invariant under affine transformations.**

I. Introduction

Mutual information has been widely used as a metric for discovering and quantifying associations between data (e.g., [1]–[3]), yet reliably estimating mutual information directly from samples is a non-trivial task. The naive route of first estimating the underlying probability densities and then computing the mutual information between the estimated distributions is generally impractical and imprecise. To address this challenge, a growing number of methods for estimating mutual information and, more broadly, distribution functionals, have recently been proposed within the information theory and computer science communities (see, e.g., [4]–[8]).

We build upon this effort and propose a moments-based approach for estimating the mutual information $I(X;Y)$ from i.i.d. samples drawn from two random variables X and Y with joint distribution $P_{X,Y}$. The estimator outlined here exploits the relationship between mutual information and minimum mean-squared error (MMSE) [9] in two steps. First, instead of tackling the problem of estimating $I(X;Y)$ directly, we focus on $I(X;\sqrt{t}Y+N)$, where $N \sim \mathcal{N}(0,1)$ is independent of (X,Y) and t is a constant (in a similar vein to [10]). Via the I-MMSE relation, if $N \sim \mathcal{N}(0,1)$ and (X,Y) are independent, the mutual information $I(X;Y)$ then satisfies

$$I(X;Y) = \frac{1}{2}\int_0^\infty \mathrm{mmse}(Y|\sqrt{t}Y+N) - \mathbb{E}_X\left[\mathrm{mmse}\left(Y_X|\,\sqrt{t}Y_X+N\right)\right]\,dt, \quad (1)$$

where we use the notation Y_x to refer to the random variable whose law is $P_{Y|X}(\cdot|x)$. Second, we approximate the above mmse expressions using polynomials (dubbed the polynomial MMSE, or PMMSE for short). One of the key results results in this paper (of also independent interest) is that the PMMSE approaches the MMSE, i.e.

$$\lim_{n\to\infty} \mathrm{pmmse}_n(W|Z) = \mathrm{mmse}(W|Z), \quad (2)$$

under mild conditions on an $\mathbb{R}^2$-valued random variable (W,Z).

By combining the convergence in (2) with equation (1), we derive a new formula for expressing mutual information as a functional of the moments of the underlying random variables, given by

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty}\int_0^\gamma f_n(Y,t) - \mathbb{E}_X f_n(Y_X,t)\,dt, \quad (3)$$

with $f_n(R,t)$ a rational function in t whose coefficients are polynomials in the moments of R. Each fixed positive integer n gives an estimate via approximating the moments in f_n by sample moments. Thus, by selecting a finite (but sufficiently large) n and γ, and estimating the corresponding moments from i.i.d. samples of X and Y, we can approximate the mutual information between X and Y via (3).

We prove that the proposed estimator is asymptotically consistent, and evaluate its performance on synthetic data. Observe in (3) that, when X and Y are independent, the integrand is zero identically, implying that the estimator asserts independence accurately. More notably, we show that f_n also satisfies $f_n(\alpha R+\beta,t) = \alpha^2 f_n(R,\alpha^2 t)$ for constant α and β, yielding the affine-transformation invariance of the estimator. Note that many other estimators (e.g. estimators based on nearest-neighbor statistics [4], [8]) are not necessarily invariant to affine transformation of the data.

In Section II, we introduce a few notations and assumptions, and we briefly review the I-MMSE relation. We prove that the polynomial MMSE approaches the MMSE in Section III, as well as give an exact expression for the rational function alluded to in (3). We then develop a moments-based formula for the mutual information in Section IV. The proposed estimator is introduced in Section V, and we illustrate its performance with simulations in Section VI.

II. Preliminaries

In this section, we lay some of the mathematical groundwork used in the derivation of our results.

A. Notation and Assumptions

For $n \in \mathbb{N}$ and an $\mathbb{R}$-valued random variable R, we denote

$$\mathbf{R}^{(n)} = (1,R,\cdots,R^n)^T. \quad (4)$$

Let $\|R\|$ denote the 2-norm, i.e., $\|R\| = \left(\mathbb{E}R^2\right)^{1/2}$. We set $[n] = \{0,1,\cdots,n\}$. For two random variables A and B, we write $A \perp\!\!\!\perp B$ when A and B are independent.

We consider an $\mathbb{R}^2$-valued random variable (X,Y) for which the mutual information $I(X;Y)$ is finite. Throughout, we fix $N \sim \mathcal{N}(0,1)$ such that $(X,Y) \perp\!\!\!\perp N$. For clarity of presentation, we assume that X is discrete, taking only finitely many values that we collect in a set denoted $\mathcal{X}$; extending the results to a continuous X can be done in view of Tonelli's theorem and Lebesgue's dominated convergence. We also assume that Y and each Y_x, for $x \in \mathcal{X}$, are continuous and that the moment generating function of Y is finite everywhere.

B. The I-MMSE Relation

The starting point of our work is the I-MMSE relation, which we briefly review next. For an $\mathbb{R}$-valued random variable R and a $\gamma > 0$, we use the shorthand

$$I(R|\gamma) := I(R; \sqrt{\gamma}R + N) = I(R; R + \gamma^{-1/2}N). \tag{5}$$

One way to write the I-MMSE relationship is as follows.

Theorem 1 (I-MMSE relation, [9])**.** *For any $\mathbb{R}$-valued random variable R such that $\mathbb{E}R^2 < \infty$ and $R \perp\!\!\!\perp N$, and for any $\gamma > 0$,*

$$I(R|\gamma) = \frac{1}{2}\int_0^\gamma \mathrm{mmse}(R|\sqrt{t}R + N)\, dt. \tag{6}$$

We have the equation

$$I(X;Y) = \frac{1}{2}\int_0^\infty \mathrm{mmse}(Y|\sqrt{t}Y + N) - \mathbb{E}\left[\mathrm{mmse}\left(Y_X|\, \sqrt{t}Y_X + N\right)\right]\, dt. \tag{7}$$

Due to the difficulty of computing conditional expectations of the form $\mathbb{E}[R|\sqrt{t}R + N]$ even for $R \perp\!\!\!\perp N$, the MMSEs in (7) are difficult to compute. Hence, formula (7) cannot be used directly to estimate $I(X;Y)$ from samples of (X,Y). We thus approximate the MMSE estimator using polynomials, as described next.

III. Polynomial MMSE

To avoid calculating MMSEs, we propose in this paper viewing the MMSE as a limit of polynomial MMSEs (PMMSE), which are natural generalizations of the linear MMSE (LMMSE) to higher degree polynomial approximations.

Definition 1 (Polynomial MMSE)**.** For an $\mathbb{R}^2$-valued random variable (W,Z) and $n \in \mathbb{N}$ such that both $\mathbb{E}W^2$ and $\mathbb{E}Z^{2n}$ are finite, define the n-th order *polynomial minimum mean-squared error* for estimating W given Z by

$$\mathrm{pmmse}_n(W|Z) := \inf\left\{\mathbb{E}\left[\left(W - \mathbf{c}^T\mathbf{Z}^{(n)}\right)^2\right] \; ; \; \mathbf{c} \in \mathbb{R}^{n+1}\right\}. \tag{8}$$

Unlike the case of the MMSE, working with the PMMSE is tractable and allows for explicit formulas that can be used for the purpose of computation from samples. An explicit formula for $t \mapsto \mathrm{pmmse}_n(R|\sqrt{t}R + N)$ is given in Theorem 3, which reveals that this mapping is a rational function of t. Further, the procedure of approximating the MMSE with the PMMSE is valid under the assumptions in Section II, as shown in the following theorem.

Theorem 2. *Let R be an $\mathbb{R}$-valued random variable such that $R \perp\!\!\!\perp N$. If the MGF of R exists everywhere, then we have the uniform convergence*

$$\sup_{t \geq 0}\left|\mathrm{pmmse}_n(R|\sqrt{t}R + N) - \mathrm{mmse}(R|\sqrt{t}R + N)\right| \to 0 \tag{9}$$

as $n \to \infty$.

The proof of Theorem 2 can be broken down into two parts. First, the pointwise convergence is a corollary of the following general PMMSE-to-MMSE convergence result.

Proposition 1. *If (W,Z) is an $\mathbb{R}^2$-valued measurable function such that $\mathbb{E}[W^2] < \infty$, the moment generating function of Z is finite everywhere, and $|\mathrm{supp}(Z)| = \infty$, then*

$$\lim_{n\to\infty} \mathrm{pmmse}_n(W|Z) = \mathrm{mmse}(W|Z). \tag{10}$$

Then, the uniformity of the convergence in Theorem 2 follows by a compactness argument via the explicit formula for the PMMSE that we give in Theorem 3. We provide the proofs of Proposition 1 and Theorem 2 in Appendices B and D, respectively.

We first discuss a geometric interpretation of the PMMSE before presenting explicit formulas in Theorem 3. The next definition will be useful for our exposition.

Definition 2. For a positive integer n and an $\mathbb{R}$-valued random variable Z such that $\mathbb{E}\left[Z^{2n}\right] < \infty$, we define the Hankel matrix[1] of moments

$$M_{Z,n} := (\mathbb{E}Z^{i+j})_{(i,j)\in[n]^2}. \tag{11}$$

A. Geometric Interpretation

We note that the PMMSE bears a geometric meaning analogous to that of the MMSE. First, the infimum in the defining equation (8) may be replaced with a minimum, as a minimizer always exists. Indeed, being finite-dimensional, the subspace of polynomials in Z of degree at most n is closed; hence, by Riesz's lemma, the projection of W onto this subspace exists and is unique [11]. We denote this unique projection of W by $E_n[W|Z]$, and refer to it as the PMMSE estimate. It follows that, since it is a projection, $E_n[W|Z]$ is the closest element to W

$$\mathrm{pmmse}_n(W|Z) = \mathbb{E}\left[(W - E_n[W|Z])^2\right], \tag{12}$$

and it also satisfies the orthogonality relation

$$\mathbb{E}[(W - E_n[W|Z])p(Z)] = 0 \tag{13}$$

for any polynomial $p(Z)$ of degree at most n. In particular,

$$\mathbb{E}[E_n[W|Z]] = \mathbb{E}[W], \tag{14}$$

resembling the law of total expectation. Further, As $\mathbb{E}[W|Z]$ is also a projection,

$$E_n[\mathbb{E}[W|Z]|Z] = E_n[W|Z]. \tag{15}$$

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

On the other hand, the coefficients defining the polynomial $E_n[W|Z]$ may be not unique. For example, if Z takes only two values, then there is a linear function in Z that vanishes, so adding multiples of this linear function to $E_n[W|Z]$ leaves it invariant. In fact, it is true that the minimizing coefficients are unique if and only if $1, Z, \cdots, Z^n$ are linearly independent, i.e., if and only if Z does not lie almost surely in an n-dimensional hyperplane (or, $|\mathrm{supp}(Z)| > n$). In such case, we obtain the projection formula

$$E_n[W|Z] = \mathbb{E}\left[W\mathbf{Z}^{(n)}\right]^T M_{Z,n}^{-1}\mathbf{Z}^{(n)}. \tag{16}$$

From this formula for the PMMSE estimate, we obtain that the PMMSE satisfies

$$\mathrm{pmmse}_n(W|Z) = \mathbb{E}W^2 - \mathbb{E}\left[W\mathbf{Z}^{(n)}\right]^T M_{Z,n}^{-1}\mathbb{E}\left[W\mathbf{Z}^{(n)}\right]. \tag{17}$$

With the interpretation that $\mathrm{pmmse}_n(A|B)$ is the L_2-distance between A and the subspace of polynomials in B of degree at most n, we have the following properties regarding affine transformations. For any $(\alpha,\beta) \in \mathbb{R}^2$,

$$\mathrm{pmmse}_n(W+\alpha|Z+\beta) = \mathrm{pmmse}_n(W|Z) \tag{18}$$

and, when $\alpha\beta \neq 0$,

$$\mathrm{pmmse}_n(\alpha W|\beta Z) = \alpha^2\,\mathrm{pmmse}_n(W|Z). \tag{19}$$

We analytically re-derive all these facts concerning both the PMMSE and the PMMSE estimate in Appendix A. It is also worth noting that the polynomial $\sum_{k=0}^n d_k q_k(Z)$ in the proof of Proposition 1 is just the projection $E_n[W|Z]$, so we also obtain the L_2-convergence of the PMMSE estimates to the MMSE estimate

$$\lim_{n\to\infty} \mathbb{E}\left[\left(E_n[W|Z] - \mathbb{E}[W|Z]\right)^2\right] = 0. \tag{20}$$

B. From PMMSE to Mutual Information

A main result of our work is showing that the mapping defined by $t \mapsto \mathrm{pmmse}_n(R|\sqrt{t}R+N)$ is a rational function of a special type. We state the result as a theorem here, and prove it in Appendix C. The characterization we shall give helps in proving the pointwise convergence in Theorem 2, and is used to express the formula for mutual information and the ensuing estimator presented in the next sections.

In the statement of the theorem, we will slightly abuse standard terminology: We say that an expression is a homogeneous polynomial in the first ℓ moments of R of degree d if that expression is an $\mathbb{R}$-linear combination of terms of the form

$$\prod_{i=1}^{\ell} \mathbb{E}\left[R^i\right]^{f_i}$$

for nonnegative integers f_i satisfying $\sum_{i=1}^{\ell} i f_i = d$ (e.g., the variance is a homogeneous polynomial in the first 2 moments of degree 2).

Theorem 3. *For any $\mathbb{R}$-valued random variable R such that $R \perp\!\!\!\perp N$ and $E[R^{2n}] < \infty$, the mapping defined by $t \mapsto \mathrm{pmmse}_n(R|\sqrt{t}R+N)$ is a rational function that can be expressed as*

$$\mathrm{pmmse}_n(R|\sqrt{t}R+N) = \frac{\sum_{j=1}^{\binom{n+1}{2}-2} a_j^{(n)}(R)t^j}{c^{(n)} + \sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(R)t^j} + \frac{1}{\binom{n+1}{2}}\frac{d}{dt}\log\left(c^{(n)} + \sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(R)t^j\right) \tag{21}$$

for real numbers $a_j^{(n)}(R)$, $b_j^{(n)}(R)$, and $c^{(n)}$ where

- *each $a_j^{(n)}(R)$ is a homogeneous polynomial in the first $2n$ moments of R of degree $2j+2$,*
- *each $b_j^{(n)}(R)$ is a homogeneous polynomial in the first $2n$ moments of R of degree $2j$,*
- *the constant term satisfies $c^{(n)} = G(n+2) = \det M_{N,n}$, with G denoting the Barnes G function, which satisfies $G(n+2) = \prod_{k=1}^n k!$,*
- *the denominator satisfies*
$$c^{(n)} + \sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(R)t^j = \det M_{\sqrt{t}R+N,n} \tag{22}$$
and is strictly positive for every $t \in [0,\infty)$,
- *the numerator satisfies*
$$\sum_{j=1}^{\binom{n+1}{2}-2} a_j^{(n)}(R)t^j = -\frac{1}{\binom{n+1}{2}}\frac{d}{dt}\det M_{\sqrt{t}R+N,n} + \det\left(M_{\sqrt{t}R+N,n}\right)\left(\mathbb{E}R^2 - \mathbf{v}^T M_{\sqrt{t}R+N,n}^{-1}\mathbf{v}\right) \tag{23}$$
with
$$\mathbf{v} = \mathbb{E}\left[R\left(\left(\sqrt{t}R+N\right)^k\right)_{k\in[n]}\right], \tag{24}$$
- *the leading coefficient is nonnegative, satisfies*
$$b_{\binom{n+1}{2}}^{(n)}(R) = \det M_{R,n}, \tag{25}$$
and is strictly positive if and only if $|\mathrm{supp}(R)| > n$.

For brevity, we use the notation

$$\Theta_n(R;t) := \frac{\sum_{j=1}^{\binom{n+1}{2}-2} a_j^{(n)}(R)t^j}{c^{(n)} + \sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(R)t^j} \tag{26}$$

in the sequel. The mutual information estimation problem we consider is solved once we have a method of recovering the functions $\Theta_n(R;t)$ (as functions of $t \geq 0$, for every $n \in \mathbb{N}$) from samples of R. Indeed, having the Θ_n amounts to having the pmmse_n, from which the MMSEs are obtained, thereby giving the mutual information in view of the I-MMSE relation.

One way to accomplish the recovery of the Θ_n is via a direct expansion of the expressions in Theorem 3, which is feasible for small n via standard symbolic computations. For larger n, Theorem 3 indicates that $a_j^{(n)}(R)$ and $b_j^{(n)}(R)$ can be approximated numerically. In particular, both the denominator and numerator of Θ_n may be obtained as a result of interpolating at $O(n^2)$ distinct values of t. For brevity, we omit further numerical considerations for computing Θ_n, but provide code for numerical evaluation accompanying this paper.

IV. A Formula for Mutual Information

Combining Theorems 1-3 reveals a formula for the mutual information in the form given in equation (3). We present the formula next, and provide the proof in Appendix E.

Theorem 4. *The mutual information satisfies*

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty}\frac{1}{2}\int_0^{\gamma}\Theta_n(Y;t) - \mathbb{E}_X\Theta_n(Y_X;t)\,dt + \frac{1}{n(n+1)}\log\frac{c^{(n)}+\sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(Y)\gamma^j}{\prod_{x\in\mathcal{X}}\left(c^{(n)}+\sum_{j=1}^{\binom{n+1}{2}} b_j^{(n)}(Y_x)\gamma^j\right)^{P_X(x)}}. \tag{27}$$

We note that the additional properties that the expressions in (27) show, e.g., uniform convergence of the PMMSE to the MMSE and monotonicity of the PMMSE, lead us to conjecture that the order of limits in (27) may be interchanged.

Equipped with the relationship between the moments and $I(X;Y)$ given in Theorem 4, we will introduce a moments-based estimator of mutual information in the next section. Specifically, we approximate the mutual information by fixing n, then further approximate the ensuing expression by replacing moments with sample moments. The estimator makes use of the following definition.

Definition 3. For $n\in\mathbb{N}$ and $\gamma>0$, we define

$$I_n(X;Y|\gamma) := \frac{1}{2}\int_0^{\gamma}\mathrm{pmmse}_n(Y|\sqrt{t}Y+N) - \mathbb{E}_X\left[\mathrm{pmmse}_n(Y_X|\sqrt{t}Y_X+N)\right]\,dt \tag{28}$$

and let

$$I_n(X;Y) := \lim_{\gamma\to\infty} I_n(X;Y|\gamma). \tag{29}$$

Remark 1. From expression (21) in Theorem 3, these expressions are well-defined and finite. Further, as in Theorem 4 and its proof, we have the following limits

$$I(X;Y+\gamma^{-1/2}N) = \lim_{n\to\infty} I_n(X;Y|\gamma), \tag{30}$$

$$I(X;Y) = \lim_{\gamma\to\infty} I(X;Y+\gamma^{-1/2}N), \tag{31}$$

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty} I_n(X;Y|\gamma). \tag{32}$$

Next, we discuss some desirable properties of the approximant I_n. As X enters into I_n only in terms of its probability, I_n is invariant under any bijective mapping of X. Further, the behavior of the PMMSE under affine transformations (equations (18) and (19)) show that I_n is also invariant under (injective) affine transformations of Y. To sum up, for a bijection $f:\mathcal{X}\to\mathcal{X}$ and constants $(\alpha,\beta)\in\mathbb{R}^2$ with $\alpha\neq 0$,

$$I_n(f(X);\alpha Y+\beta) = I_n(X,Y). \tag{33}$$

In addition, we note that if $X \perp\!\!\!\perp Y$, then $I_n(X;Y)=0$ for every $n\in\mathbb{N}$.

We give full expressions for the first two approximants of mutual information that are generated by the LMMSE and quadratic MMSE. When $n=1$, we obtain (with σ denoting the standard deviation)

$$I_1(X;Y) = \log\sigma(Y) - \mathbb{E}_X\log\sigma(Y_X), \tag{34}$$

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

For $n=2$, we obtain the formula (dropping the superscripts for readability)

$$\begin{aligned} I_2(X;Y) &= \frac{1}{6}\log\frac{b_3(Y)}{\prod_{x\in\mathcal{X}} b_3(Y_x)^{P_X(x)}} \\ &\quad + \frac{1}{2}\int_0^{\infty}\frac{a_1(Y)t}{2+b_1(Y)t+b_2(Y)t^2+b_3(Y)t^3} \\ &\qquad - \mathbb{E}_X\frac{a_1(Y_X)t}{2+b_1(Y_X)t+b_2(Y_X)t^2+b_3(Y_X)t^3}\,dt \end{aligned}$$

where we may compute

$$\begin{aligned} b_3(R) &:= \begin{vmatrix} 1 & \mathbb{E}R & \mathbb{E}R^2 \\ \mathbb{E}R & \mathbb{E}R^2 & \mathbb{E}R^3 \\ \mathbb{E}R^2 & \mathbb{E}R^3 & \mathbb{E}R^4 \end{vmatrix} \\ &= \sigma(R)^2\mathbb{E}R^4 + 2(\mathbb{E}R)(\mathbb{E}R^2)\mathbb{E}R^3 - (\mathbb{E}R^2)^3 - (\mathbb{E}R^3)^2, \end{aligned}$$

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

$$\begin{aligned} b_2(R) &= -4(\mathbb{E}R)\mathbb{E}R^3 + 3(\mathbb{E}R^2)^2 + \mathbb{E}R^4 \\ b_1(R) &= 6\sigma(R)^2 \\ a_1(R) &= 4(\mathbb{E}R)^4 - 8(\mathbb{E}R)^2\mathbb{E}R^2 + \frac{8}{3}(\mathbb{E}R)\mathbb{E}R^3 + 2(\mathbb{E}R^2)^2 \\ &\quad - \frac{2}{3}\mathbb{E}R^4. \end{aligned}$$

V. The Estimator

As sample moments converge almost surely to the moments, and as Theorem 4 shows that the mutual information depends continuously on the moments, the continuous mapping theorem allows for the introduction of a consistent moments-based estimator of mutual information.

Definition 4. For $m\in\mathbb{N}$, fix a sequence of m i.i.d. samples $\mathcal{S} = \{(X_j,Y_j)\sim(X,Y)\}_{1\le j\le m}$. Define the decreasing sequence of multi-sets $\mathcal{S}_1\supseteq\mathcal{S}_2\supseteq\cdots$ as follows. For each $n\in\mathbb{N}$,

$$\mathcal{S}_n := \{(X_j,Y_j)\ ;\ |\{1\le i\le m\ ;\ X_i=X_j\}|>n\}. \tag{35}$$

For each $n\in\mathbb{N}$, let $(U^{(n)},V^{(n)})\sim\mathrm{Unif}(\mathcal{S}_n)$ be independent of N. We define, for each $\gamma>0$ and each $n\in\mathbb{N}$ such that $\mathcal{S}_n$ is nonempty,

$$\widehat{I_n}(\mathcal{S}|\gamma) := I_n(U^{(n)},V^{(n)}|\gamma), \tag{36}$$

and we set

$$\widehat{I_n}(\mathcal{S}) := \lim_{\gamma\to\infty}\widehat{I_n}(\mathcal{S}|\gamma). \tag{37}$$

We have the following convergence result, whose proof is given in Appendix F.

Theorem 5. *For a positive integer n and a sequence of i.i.d. samples $\{(X_j, Y_j) \sim (X,Y)\}_{j\in\mathbb{N}}$, we have that*

$$\widehat{I}_n\left(\{(X_j,Y_j)\}_{1\le j\le m}|\gamma\right) \to I_n(X;Y|\gamma) \tag{38}$$

for every $\gamma > 0$ and

$$\widehat{I}_n\left(\{(X_j,Y_j)\}_{1\le j\le m}\right) \to I_n(X;Y) \tag{39}$$

both almost surely as $m\to\infty$. Furthermore,

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty}\lim_{m\to\infty}\widehat{I}_n\left(\{(X_j,Y_j)\}_{1\le j\le m}|\gamma\right), \tag{40}$$

where the convergence in m is almost sure convergence.

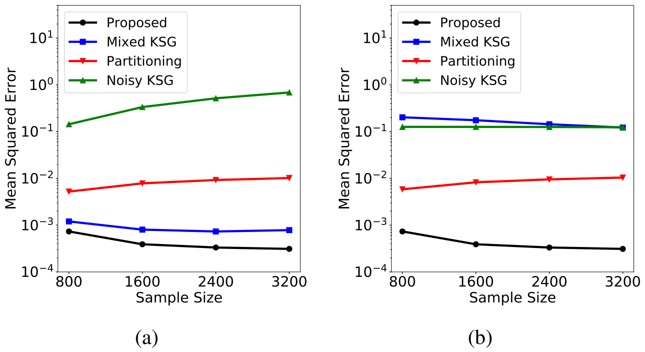
If Y is of bounded support, Hoeffding's inequality implies that, for some constant d, having $|\mathcal{S}| > (d/\varepsilon^2)\log(|\mathcal{X}|/\delta)$ as $\varepsilon \to 0$ ensures that $\Pr\left\{|\widehat{I}_n(\mathcal{S}) - I_n(X;Y)| < \varepsilon\right\} > 1-\delta$ (see Appendix G).

VI. Simulations

We compare via synthetic experiments the performance of our estimator to that of the partitioning estimator, the Noisy KSG estimator based on the estimator in [4], and the Mixed KSG estimator [8]. We utilize the implementation in [8] for all of these three estimators. For fairness of comparison, the parameters are fixed throughout, namely, we set $n=5$ for our estimator and utilize the parameters used in [8] ($k=5$ for both the Noisy KSG and the Mixed KSG, $\sigma = 0.01$ for the Noisy KSG, and 8 bins per dimension for the partitioning estimator). So, for samples $\mathcal{S}$ of (X,Y), our estimate for $I(X;Y)$ will be $\widehat{I}_5(\mathcal{S})$ as defined in equation (37). We perform 250 independent trials, then plot the mean squared error of the estimations of $I(X;Y)$ against the sample size.

Experiment I: We replicate the mixture-distribution part of the zero-inflated Poissonization experiment of [8]. 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 ground truth is approximately 0.25606. The comparison of estimators' performance is plotted in Figure 1(a). We also test the affine-transformation invariance property of the proposed estimator. Plotted in Figure 1(b) is a comparison of the same estimators using the same samples as those used to generate Figure 1(a), but where Y is processed through an affine transformation. Specifically, each Y sample is scaled by 10^4. The ground truth stays unchanged, and so do our estimator and the partitioning estimator, but the Noisy KSG and Mixed KSG change. Although the setup is more general than the assumptions we prove our results under in this paper (e.g., the MGF of Y does not exist everywhere, and X is not finitely supported), the proposed estimator outperforms the other estimators.

Experiment II: We test for independence under the following settings. We let $X\sim \mathrm{Bernoulli}(0.5)$ and $Y \sim \mathrm{Unif}([0,2])$ be such that $X \perp\!\!\!\perp Y$. The results are plotted in Figure 2.



(a)

(b)

Figure 1. Mean Squared Error vs. Sample Size for Experiment I for (a) unscaled and (b) scaled samples. The proposed estimator is resilient to scaling.

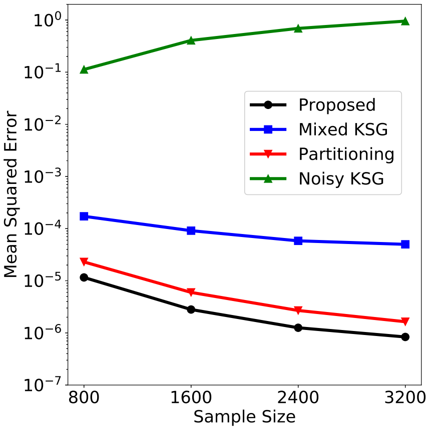


Figure 2. Mean Squared Error vs. Sample Size when estimating $I(X;Y)$ for $X \perp\!\!\!\perp Y$ with X uniform over $\{0,1\}$ and Y uniform over $[0,2]$.

Appendix A
PMMSE Basics

We prove in this appendix that the PMMSE satisfies the properties mentioned in Section III-A.

A. PMMSE Behavior Under Affine Transformations

Lemma 1. *For any $(\alpha,\beta)\in\mathbb{R}^2$, one has that*

$$\mathrm{pmmse}_n(W+\alpha|Z+\beta) = \mathrm{pmmse}_n(W|Z) \tag{41}$$

and, when $\alpha\beta \neq 0$,

$$\mathrm{pmmse}_n(\alpha W|\beta Z) = \alpha^2\,\mathrm{pmmse}_n(W|Z). \tag{42}$$

Proof. Set $U = Z+\beta$. For any $\mathbf{c}\in\mathbb{R}^{n+1}$,

$$W+\alpha-\mathbf{c}^T\mathbf{U}^{(n)} = W - (M\mathbf{c}-\alpha\mathbf{e}_1)^T\mathbf{Z}^{(n)} \tag{43}$$

where we define the matrix

$$M := \left(\beta^{i-j}\binom{i}{j}\right)_{(i,j)\in[n]^2} \tag{44}$$

with the understanding that $\beta^{i-j}\binom{i}{j} = 0$ when $j>i$, and $\beta^0\binom{i}{i} = 1$ when $\beta = 0$. Then M is lower-triangular with an all-1 diagonal, so the inverse M^{-1} exists. Thus, the mapping

$\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defined by $\mathbf{c} \mapsto M\mathbf{c}-\alpha\mathbf{e}_1$ is invertible (where $\mathbf{d} \mapsto M^{-1}(\mathbf{d}+\alpha\mathbf{e}_1)$ is the inverse mapping). By the definition of the PMMSE, then, equality (18) holds.

Equation (19) may be treated similarly. Setting $V = \beta Z$, one has that for any $\mathbf{c} \in \mathbb{R}^{n+1}$

$$\alpha W - \mathbf{c}^T \mathbf{V}^{(n)} = \alpha \left(W - (L\mathbf{c})^T \mathbf{Z}^{(n)} \right) \tag{45}$$

where we define the invertible matrix

$$L := \operatorname{diag}\left(\left(\beta^k / \alpha \right)_{k \in [n]} \right). \tag{46}$$

As $\mathbf{c} \mapsto L\mathbf{c}$ is a bijection of $\mathbb{R}^{n+1}$, the definition of the PMMSE yields equation (19). □

B. A Preliminary Formula for the PMMSE

The following lemma about the Hankel matrix of moments is instrumental for the proofs in this paper.

Lemma 2. *For any $\mathbb{R}$-valued random variable Z and $n \in \mathbb{N}$ such that $\mathbb{E}\left[Z^{2n}\right] < \infty$, the inverse $M_{Z,n}^{-1}$ exists if and only if $|\operatorname{supp}(Z)| > n$.*

Proof. First, note that $M_{Z,n}$ is symmetric. For any $\mathbf{0} \neq \mathbf{d} \in \mathbb{R}^{n+1}$,

$$\begin{aligned} \mathbf{d}^T M_{Z,n} \mathbf{d} &= \mathbf{d}^T (\mathbb{E} Z^{i+j})_{(i,j)} \mathbf{d} = \mathbf{d}^T \mathbb{E}\left[\mathbf{Z}^{(n)} \left(\mathbf{Z}^{(n)} \right)^T \right] \mathbf{d} \\ &= \mathbb{E}\left[\mathbf{d}^T \mathbf{Z}^{(n)} \left(\mathbf{Z}^{(n)} \right)^T \mathbf{d} \right] = \mathbb{E}\left| \mathbf{d}^T \mathbf{Z}^{(n)} \right|^2 \geq 0, \end{aligned} \tag{47}$$

so $M_{Z,n}$ is positive semi-definite. Furthermore, $M_{Z,n}$ is positive definite if and only if $\mathbf{Z}^{(n)}$ does not lie almost surely in a hyperplane, i.e., if and only if $|\operatorname{supp}(Z)| > n$. □

Next, we prove a preliminary formula for the PMMSE.

Lemma 3. *For a measurable (W, Z) and $n \in \mathbb{N}$ such that $\mathbb{E}\left[W^2\right], \mathbb{E}\left[Z^{2n}\right] < \infty$ and $|\operatorname{supp}(Z)| > n$,*

$$\operatorname{pmmse}_n(W|Z) = \mathbb{E}\left[\left(W - \mathbf{c}_{W,Z,n}^T \mathbf{Z}^{(n)} \right)^2 \right], \tag{48}$$

where we define

$$\mathbf{c}_{W,Z,n} := M_{Z,n}^{-1} \mathbb{E}\left[W \mathbf{Z}^{(n)} \right]. \tag{49}$$

Proof. Consider the function $h : \mathbb{R}^{n+1} \to [0, \infty)$ defined by

$$h(\mathbf{d}) = \mathbb{E}\left[\left(W - \mathbf{d}^T \mathbf{Z}^{(n)} \right)^2 \right].$$

For any $\mathbf{d} \in \mathbb{R}^{n+1}$, linearity of expectation implies that the gradient of h is

$$\nabla h(\mathbf{d}) = \left(\mathbb{E}\left[2Z^k \left(\mathbf{d}^T \mathbf{Z}^{(n)} - W \right) \right] \right)_{0 \leq k \leq n},$$

so the Hessian of h is the constant $2M_{Z,n}$. As $M_{Z,n}$ is positive-definite, h is strictly convex. As $\nabla h(\mathbf{d}) = \mathbf{0}$ is equivalent to $M_{Z,n}\mathbf{d} = \mathbb{E}\left[W\mathbf{Z}^{(n)}\right]$, i.e., to $\mathbf{d} = \mathbf{c}_{W,Z,n}$, the desired result follows. □

We may rewrite (48) as

$$\begin{aligned} \operatorname{pmmse}_n(W|Z) &= \mathbb{E}\left[\left(W - \mathbf{c}_{W,Z,n}^T \mathbf{Z}^{(n)} \right)^2 \right] \\ &= \mathbb{E}W^2 - 2\,\mathbf{c}_{W,Z,n}^T \mathbb{E}\left[W\mathbf{Z}^{(n)} \right] \\ &\qquad + \mathbf{c}_{W,Z,n}^T M_{Z,n} \mathbf{c}_{W,Z,n} \\ &= \mathbb{E}W^2 - \mathbb{E}\left[W\mathbf{Z}^{(n)} \right]^T M_{Z,n}^{-1} \mathbb{E}\left[W\mathbf{Z}^{(n)} \right]. \end{aligned} \tag{50}$$

C. Geometric Properties of the PMMSE Estimate

The proof of Lemma 3 shows the uniqueness of the PMMSE estimate, which we denote by $E_n[W|Z]$.

Definition 5. For a measurable (W, Z) and $n \in \mathbb{N}$ such that $\mathbb{E}\left[W^2\right], \mathbb{E}\left[Z^{2n}\right] < \infty$ and $|\operatorname{supp}(Z)| > n$, set

$$E_n[W|Z] = \mathbf{c}_{W,Z,n}^T \mathbf{Z}^{(n)}. \tag{51}$$

Plugging in equation (49) and rearranging, we obtain that

$$\mathbf{c}_{E_n[W|Z],Z,n} = \mathbf{c}_{W,Z,n} = \mathbf{c}_{\mathbb{E}[W|Z],Z,n}. \tag{52}$$

In particular, then, equation (51) yields that

$$E_n[\mathbb{E}[W|Z]|Z] = E_n[W|Z]. \tag{53}$$

Further, as $M_{Z,n}^{-1}\mathbb{E}\left[\mathbf{Z}^{(n)}\right]$ is the vector $(1, 0, \cdots, 0)^T$, we get that

$$\mathbb{E}[E_n[W|Z]] = \mathbb{E}[W]. \tag{54}$$

More generally, as $M_{Z,n}^{-1}\mathbb{E}\left[Z^j\mathbf{Z}^{(n)}\right]$, for $j \in [n]$, is the vector with a 1 at the j-th entry and zero elsewhere, we get that for any polynomial $p(Z)$ of degree at most n

$$\mathbb{E}[(W - E_n[W|Z])p(Z)] = 0. \tag{55}$$

Appendix B
Proof of Proposition 1

Let $\mathcal{S} = \operatorname{supp}(P_Z)$, and consider the weighted-L^2 space $L^2(\mathcal{S}, P_Z)$ of functions $f : \mathcal{S} \to \mathbb{R}$ such that

$$\int_{\mathcal{S}} f(z)^2 P_Z(z)\, dz < \infty$$

equipped with the inner product

$$\langle f, g \rangle = \int_{\mathcal{S}} f(z) g(z) P_Z(z)\, dz.$$

We know that $L^2(\mathcal{S}, P_Z)$ is a separable Hilbert space. We will show that there exists a complete orthonormal basis of $L^2(\mathcal{S}, P_Z)$ consisting of polynomials.

Fix $f \in L^2(\mathcal{S}, P_Z)$ satisfying $\langle f, z^k \rangle = 0$ for every $k \in \mathbb{N}$, and we'll show that $f = 0$. Let $f_1 : \mathbb{R} \to \mathbb{R}$ be the extension of f such that $f_1(z) = 0$ for $z \notin \mathcal{S}$. Consider $\varphi : \mathbb{C} \to \mathbb{C}$ defined by

$$\varphi(s) := \int_{\mathbb{R}} e^{sz} f_1(z) P_Z(z)\, dz = \mathbb{E}[e^{sZ} f_1(Z)]. \tag{56}$$

By Morera's theorem, φ is an entire function. We have that $\varphi^{(k)}(0) = 0$ for every $k \in \mathbb{N}$. Considering the power series of φ around 0, we obtain that $\varphi(s) = 0$ for every $s \in \mathbb{C}$. In particular, for $s = -i\tau$ and $\tau \in \mathbb{R}$, we have that the Fourier transform of the function $g(z) := f_1(z)P_Z(z)$ satisfies $\widehat{g}(\tau) = 0$ for every $\tau \in \mathbb{R}$. Hence, $g(z) = 0$ for every $z \in \mathbb{R}$, i.e., $f = 0$.

Further, since $|\mathrm{supp}(Z)| = \infty$, the monomials are linearly independent. Hence, applying Gram-Schmidt, one obtains an orthonormal basis consisting of polynomials $\{q_k\}_{k\in\mathbb{N}}$ such that $\deg q_k = k$ for each $k \in \mathbb{N}$.

Then, for some $\{d_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\sum_{k=0}^{n} d_k q_k(Z) - \mathbb{E}[W|Z]\right|^2\right] = 0. \tag{57}$$

Writing $u_n(Z) = \sum_{k=0}^{n} d_k q_k(Z)$, one sees that

$$\begin{aligned}\|W - \mathbb{E}[W|Z]\| &\le \mathrm{pmmse}_n(W|Z)^{1/2} \\ &\le \|W - u_n(Z)\| \\ &\le \|W - \mathbb{E}[W|Z]\| + \|u_n(Z) - \mathbb{E}[W|Z]\|,\end{aligned}$$

so $\mathrm{pmmse}_n(W|Z) \to \mathrm{mmse}(W|Z)$.

Appendix C
Proof of Theorem 3: PMMSE is a Rational Function

Equation (50) in Appendix A gives the preliminary formula for the PMMSE

$$\mathrm{pmmse}_n(W|Z) = \mathbb{E}W^2 - \mathbb{E}\left[W\mathbf{Z}^{(n)}\right]^T M_{Z,n}^{-1}\mathbb{E}\left[W\mathbf{Z}^{(n)}\right]. \tag{58}$$

when both W and Z^n are square-integrable and $|\mathrm{supp}(Z)| > n$. We utilize equation (58) to derive explicit formulas for the function $t \mapsto \mathrm{pmmse}_n(R|\sqrt{t}R + N)$ when $R \perp\!\!\!\perp N$.

Our derivations will be combinatorial in nature. Specifically, we analyze the ensuing permutations that arise from the Leibniz formula for determinants. We begin with some notation.

For $n \in \mathbb{N}$, we let $S_n^{(0)}$ denote the symmetric group of permutations on the $n+1$ elements $[n]$. It will be useful to let, for $(i,j) \in [n]^2$, the collection of permutations in $S_n^{(0)}$ sending i to j be denoted by $T_n^{(i,j)}$, i.e., set

$$T_n^{(i,j)} := \{\pi \in S_n^{(0)} \ ; \ \pi(i) = j\}. \tag{59}$$

We also introduce the following shorthands.

Definition 6. Fix an $\mathbb{R}$-valued random variable R such that $R \perp\!\!\!\perp N$, an integer $n > 0$, and let $t > 0$. Assume that $\mathbb{E}R^{2n} < \infty$. We define

- the function $F_{R,n} : [0,\infty) \to [0,\infty)$ by
$$\begin{aligned}F_{R,n}(t) := \ &\mathbb{E}\left[R\left(\left(\sqrt{t}R+N\right)^k\right)_{k\in[n]}\right]^T M_{\sqrt{t}R+N,n}^{-1} \\ &\mathbb{E}\left[R\left(\left(\sqrt{t}R+N\right)^k\right)_{k\in[n]}\right],\end{aligned} \tag{60}$$
- for each $(i,j) \in [n]^2$, the cofactor of $M_{\sqrt{t}R+N,n}$
$$c_{R,n}^{(i,j)}(t) := \sum_{\pi\in T_n^{(i,j)}} \mathrm{sgn}(\pi) \prod_{k\neq i}\left(M_{\sqrt{t}R+N,n}\right)_{k,\pi(k)}, \tag{61}$$
- the cofactor matrix
$$C_{R,n}(t) = \left(c_{R,n}^{(i,j)}(t)\right)_{(i,j)\in[n]^2} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{62}$$
- the function $D_{R,n} : [0,\infty) \to [0,\infty)$ by
$$\begin{aligned}D_{R,n}(t) := \ &\mathbb{E}\left[R\left(\left(\sqrt{t}R+N\right)^k\right)_{k\in[n]}\right]^T C_{R,n}(t) \\ &\mathbb{E}\left[R\left(\left(\sqrt{t}R+N\right)^k\right)_{k\in[n]}\right],\end{aligned} \tag{63}$$
- for each $(i,j) \in [n]^2$,
$$\begin{aligned}d_{R,n}^{(i,j)}(t) := \ &\mathbb{E}\left[R\left(\sqrt{t}R+N\right)^i\right] c_{R,n}^{(i,j)}(t) \\ &\mathbb{E}\left[R\left(\sqrt{t}R+N\right)^j\right],\end{aligned} \tag{64}$$
- for each $k \in [2n]$,
$$\mathcal{R}_k := \mathbb{E}R^k, \tag{65}$$
- for each $\pi \in S_n^{(0)}$, $m \in \mathbb{N}$, and $\{i_1,\cdots,i_m\} \subseteq [n]$, the product
$$Q_R(\pi; i_1,\cdots,i_m) := \prod_{k\notin\{i_1,\cdots,i_m\}} \mathcal{R}_{k+\pi(k)}. \tag{66}$$

Remark 2. There are a few relationships between these shorthands. For example,

$$D_{R,n}(t) = \sum_{(i,j)\in[n]^2} d_{R,n}^{(i,j)}(t), \tag{67}$$

and by Cramer's rule

$$F_{R,n}(t) = \frac{D_{R,n}(t)}{\det M_{\sqrt{t}R+N,n}}. \tag{68}$$

The relation (68) implies, in view of equation (58), that

$$\mathrm{pmmse}_n(R|\sqrt{t}R+N) = \mathcal{R}_2 - \frac{D_{R,n}(t)}{\det M_{\sqrt{t}R+N,n}}. \tag{69}$$

We first show that the PMMSE is rational, then show that the specific properties in Theorem 3 hold.

A. Rationality

Lemma 4. *Fix $(S,R) \perp\!\!\!\perp N$ and $\ell \in \mathbb{N}$. If ℓ is even then $\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^\ell\right]$ is a polynomial in t, and if ℓ is odd then $t^{-1/2}\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^\ell\right]$ is a polynomial in t. Further, both polynomials are of degree at most $\lfloor \ell/2 \rfloor$, and (for $\ell \ge 2$) subtracting $t^{\lfloor \ell/2\rfloor}\mathbb{E}SR^\ell$ produces a polynomial of degree at most $\lfloor \ell/2\rfloor - 1$ (for $\ell \in \{0,1\}$, we produce a zero).*

Proof. Note that $\mathbb{E}N^r = 0$ for odd $r \in \mathbb{N}$. Assume first that ℓ is even. Then,

$$\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^\ell\right] = \sum_{k \text{ even}} \binom{\ell}{k} t^{k/2}\mathbb{E}SR^k\mathbb{E}N^{\ell-k},$$

so it is a polynomial in t.

Now, assume instead that ℓ is odd. Then,

$$t^{-1/2}\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right] = \sum_{k \text{ odd}} \binom{\ell}{k} t^{(k-1)/2}\mathbb{E}SR^{k}\mathbb{E}N^{\ell-k},$$

so it also is a polynomial in t. The second statement of the lemma follows from the expressions above. □

Definition 7. Fix $(S,R) \perp\!\!\!\perp N$ and $\ell \in \mathbb{N}$. If ℓ is even, set

$$e_{S,R,\ell}(t) := \mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right]. \tag{70}$$

If ℓ is odd, set

$$o_{S,R,\ell}(t) := t^{-1/2}\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right]. \tag{71}$$

From lemma 4, all $e_{S,R,\ell}$ and $o_{S,R,\ell}$ are polynomials. Further, if $i+j$ is even then $\left(M_{\sqrt{t}R+N,n}\right)_{i,j} = e_{1,R,i+j}(t)$, and if $i+j$ is odd then $\left(M_{\sqrt{t}R+N,n}\right)_{i,j} = \sqrt{t}o_{1,R,i+j}(t)$.

Lemma 5. *For $n \in \mathbb{N}$, a permutation $\pi \in S_n$, and a partition $\{1,\cdots,n\} = A \cup B$, the set*

$$T := \{i \in \{1,\cdots,n\} \ ; \ (i,\pi(i)) \in (A\times B)\cup(B\times A)\}$$

has even cardinality.

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

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

and define B_π similarly. Then, $T = A_\pi \cup B_\pi$ is a partition. As $|A_\pi| = |B_\pi|$, the desired result follows. □

Corollary 1. *For any permutation $\pi \in S_n$, the number $|\{i \in \{1,\cdots,n\} \ ; \ i+\pi(i)$ is odd $\}|$ is even.*

Proof. As $i+\pi(i)$ is odd if and only if i and $\pi(i)$ have opposite parities, the result follows from lemma 5 by partitioning $\{1,\cdots,n\}$ into even and odd numbers. □

Definition 8. For $n \in \mathbb{N}$ and $\pi \in S_n^{(0)}$, set

$$o(\pi) := \left|\{i \in \{0,\cdots,n\} \ ; \ i+\pi(i) \text{ is odd }\}\right|. \tag{72}$$

Proposition 2. *For $R \perp\!\!\!\perp N$ and $n \in \mathbb{N}$, $\det M_{\sqrt{t}R+N,n}$ is a polynomial in t of degree at most $n(n+1)/2$. Furthermore, the polynomial $\det M_{\sqrt{t}Y+N,n}$ is of degree exactly $n(n+1)/2$ if and only if $|\mathrm{supp}(R)| > n$, in which case the leading coefficient is $\det M_{R,n}$.*

Proof. We may write

$$\det M_{\sqrt{t}R+N,n} \tag{73}$$
$$= \sum_{\pi\in S_n^{(0)}} \mathrm{sgn}(\pi) t^{o(\pi)/2} \prod_{i \, : \, i+\pi(i) \text{ odd}} o_{1,R,i+\pi(i)}(t) \prod_{j \, : \, j+\pi(j) \text{ even}} e_{1,R,j+\pi(j)}(t), \tag{74}$$

thereby showing that $\det M_{\sqrt{t}Y+N,n}$ is a polynomial in t by evenness of each $o(\pi)$. Furthermore, for each $\pi \in S_n^{(0)}$,

$$\begin{aligned}&\deg t^{o(\pi)/2} \prod_{i \, : \, i+\pi(i) \text{ odd}} o_{1,R,i+\pi(i)}(t) \prod_{j \, : \, j+\pi(j) \text{ even}} e_{1,R,j+\pi(j)}(t)\\ &\leq \frac{o(\pi)}{2} + \sum_{i \, : \, i+\pi(i) \text{ odd}} \frac{i+\pi(i)-1}{2} + \sum_{j \, : \, j+\pi(j) \text{ even}} \frac{j+\pi(j)}{2}\\ &= \frac{1}{2}\sum_{k=0}^{n} k+\pi(k) = \frac{n(n+1)}{2}.\end{aligned}$$

Finally, as

$$\deg\left(\det M_{\sqrt{t}R+N,n} - \det M_{\sqrt{t}R,n}\right) < \frac{n(n+1)}{2}$$

and $\det M_{\sqrt{t}R,n} = t^{n(n+1)/2}\det M_{R,n}$, the last statement of the lemma follows. □

Lemma 6. *Fix $R \perp\!\!\!\perp N$, $n \in \mathbb{N}$, and $(i,j) \in \{0,\cdots,n\}^2$. If $i+j$ is even, then $c_{R,n}^{(i,j)}(t)$ is a polynomial in t, and if $i+j$ is odd then $\sqrt{t}c_{R,n}^{(i,j)}(t)$ is a polynomial in t. Further, in both cases the polynomial is of degree at most $n(n+1)/2 - \lfloor (i+j)/2 \rfloor$ and subtracting (for the case $(i,j,n) \neq (1,1,1)$) a term*

$$t^{n(n+1)/2-\lfloor (i+j)/2\rfloor} \sum_{\pi \in T_n^{(i,j)}} \mathrm{sgn}(\pi) Q_R(\pi;i)$$

produces a polynomial of degree at most $n(n+1)/2 - \lfloor (i+j)/2 \rfloor - 1$.

Proof. If $i+j$ is even, then

$$\begin{aligned}&c_{R,n}^{(i,j)}(t)\\ &= \sum_{\pi\in T_n^{(i,j)}} \mathrm{sgn}(\pi) t^{o(\pi)/2} \prod_{k \, : \, k+\pi(k) \text{ odd}} o_{1,R,k+\pi(k)}(t) \prod_{r \, : \, r+\pi(r) \text{ even}, \, r\neq i} e_{1,R,r+\pi(r)}(t),\end{aligned}$$

whereas if $i+j$ is odd then

$$\begin{aligned}&c_{R,n}^{(i,j)}(t)\\ &= \sum_{\pi\in T_n^{(i,j)}} \mathrm{sgn}(\pi) t^{(o(\pi)-1)/2} \prod_{k \, : \, k+\pi(k) \text{ odd}, \, k\neq i} o_{1,R,k+\pi(k)}(t) \prod_{r \, : \, r+\pi(r) \text{ even}} e_{1,R,r+\pi(r)}(t),\end{aligned}$$

Thus, the first statement of the lemma follows. The second statement of the lemma follows from analyzing the the degrees and leading coefficients of each o and e, since for even $i+j$

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

and for odd $i+j$

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

□

Lemma 7. *For $R \perp\!\!\!\perp N$ and $n \in \mathbb{N}$, the function $D_{R,n}(t)$ is a polynomial in t. Further, the polynomial $D_{R,n}(t)$ is of degree at most $n(n+1)/2$ and subtracting a term*

$$t^{n(n+1)/2} \sum_{0 \le i,j \le n} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{j+1} Q_R(\pi; i)$$

produces a polynomial of degree at most $n(n+1)/2 - 1$.

Proof. We will show that each $d_{R,n}^{(i,j)}(t)$ is a polynomial in t. Fix $(i,j) \in \{0, \cdots, n\}^2$. Consider separately the parity of $i+j$.

Assume first that $i+j$ is even, so i and j have the same parity. Then, by lemma 4,

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

is a polynomial in t of degree at most $(i+j)/2$ and subtracting (for $i+j \ge 2$) a term $t^{(i+j)/2}\mathcal{R}_{i+1}\mathcal{R}_{j+1}$ produces a polynomial of degree at most $(i+j)/2-1$ (if $i+j=0$, then we produce a zero).

Now, assume instead that $i+j$ is odd, so i and j have different parities. Then,

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

is a polynomial in t of degree at most $(i+j-1)/2$ and subtracting (for $i+j \ge 3$) a term $t^{(i+j-1)/2}\mathcal{R}_{i+1}\mathcal{R}_{j+1}$ produces a polynomial of degree at most $(i+j-1)/2-1$ (if $i+j=1$, then we produce a zero). □

B. Leading Coefficients

Lemma 8. *For $R \perp\!\!\!\perp N$ and $n \in \mathbb{N}$,*

$$\sum_{0 \le i,j \le n} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{j+1} Q_R(\pi; i) = \mathcal{R}_2 \det M_{R,n}. \tag{75}$$

Proof. Note that, for each fixed $i \in [n]$, we have a partition

$$S_n^{(0)} = \bigcup_{j=0}^{n} T_n^{(i,j)}.$$

Thus, denoting $U := \{0, \cdots, n\} \times S_n^{(0)}$, we may rewrite

$$\begin{aligned}\sum_{0 \le i,j \le n} \sum_{\pi \in T_n^{(i,j)}} & \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{j+1} Q_R(\pi; i) \\ &= \sum_{(i,\pi)\in U} \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{\pi(i)+1} Q_R(\pi; i).\end{aligned}$$

For $(i,\pi) \in U$, set $\pi_i := \pi(1\ i)$. Then, we have the partition

$$U = \bigcup_{(i,\pi)\in U} \{(i,\pi_i)\}.$$

Define $f : U \to \mathbb{R}$ by

$$f(i,\pi) = \mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1} Q_Y(\pi; i).$$

We may write

$$\sum_{(i,\pi)\in U} \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{\pi(i)+1} Q_R(\pi; i) = \sum_{(i,\pi)\in U} \operatorname{sgn}(\pi) f(i,\pi).$$

Note that $f(1,\pi) = f(1,\pi_1)$, and for $i \ne 1$,

$$\begin{aligned} f(i,\pi) &= \mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{\pi(1)+1} \prod_{k \notin \{1,i\}} \mathcal{R}_{\pi(k)+k} \\ &= \mathcal{R}_{i+1}\mathcal{R}_{\pi_i(1)+1}\mathcal{R}_{\pi_i(i)+1} \prod_{k \notin \{1,i\}} \mathcal{R}_{\pi_i(k)+k} = f(i,\pi_i). \end{aligned}$$

Furthermore, for $i \ne 1$,

$$\operatorname{sgn}(\pi_i) = -\operatorname{sgn}(\pi).$$

Denote $U' := U \setminus \bigcup_{\pi \in S_n^{(0)}} \{(1,\pi)\}$. Hence,

$$\begin{aligned} \sum_{(i,\pi)\in U'} \operatorname{sgn}(\pi) f(i,\pi) &= \sum_{(i,\pi_i)\in U'} \operatorname{sgn}(\pi_i) f(i,\pi_i) \\ &= -\sum_{(i,\pi_i)\in U'} \operatorname{sgn}(\pi) f(i,\pi) \\ &= -\sum_{(i,\pi)\in U'} \operatorname{sgn}(\pi) f(i,\pi). \end{aligned}$$

Thus,

$$\sum_{(i,\pi)\in U'} \operatorname{sgn}(\pi) f(i,\pi) = 0,$$

implying that

$$\begin{aligned} \sum_{(i,\pi)\in U} \operatorname{sgn}(\pi) \mathcal{R}_{i+1} \mathcal{R}_{\pi(i)+1} Q_R(\pi; i) &= \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) f(1,\pi) \\ &= \mathcal{R}_2 \det M_{R,n} \end{aligned}$$

where the last equality follows since we may write

$$\det M_{R,n} = \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+\pi(i)} Q_R(\pi; i)$$

for any $i \in [n]$. □

Thus, we have proved the following fact.

Proposition 3. *For $R \perp\!\!\!\perp N$ and $n \in \mathbb{N}$ such that $|\operatorname{supp}(R)| > n$, and for $t > 0$, we may write*

$$\operatorname{pmmse}_n(R|\sqrt{t}R+N) = \frac{\gamma_{R,n}(t)}{\delta_{R,n}(t)},$$

where $\gamma_{R,n}$ is a polynomial of degree at most $n(n+1)/2-1$ and $\delta_{R,n}$ is a polynomial of degree exactly $n(n+1)/2$ and leading coefficient $\det M_{R,n}$.

Finally, to see that the leading term of $\gamma_{R,n}$ is also $\det M_{R,n}$, we apply the same trick in Lemma 8 to the leading term

$$\begin{aligned}\mathcal{R}_2 \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi) \sum_{r=0}^{n} \binom{r+\pi(r)}{2} \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; r) \\ - \sum_{0\le i,j\le n} \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1}\mathcal{R}_{j+1} \\ \sum_{r\neq i} \binom{r+\pi(r)}{2} \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; i, r) \\ - \sum_{0\le i,j\le n} \sum_{\pi\in T_n^{(i,j)}} \operatorname{sgn}(\pi) Q_R(\pi; i) \\ \left(\binom{i}{2} \mathcal{R}_{i-1}\mathcal{R}_{j+1} + \binom{j}{2} \mathcal{R}_{i+1}\mathcal{R}_{j-1} \right).\end{aligned}$$

Appendix D
Proof of Theorem 2

The pointwise convergence in Theorem 2 follows from Proposition 1 as a direct result of the following two facts. First, $|\text{supp}(\sqrt{t}R+N)| = \infty$ regardless of what R is. Second, the moment generating function of $\sqrt{t}R+N$ is the product of those of $\sqrt{t}R$ and N by assumption of independence. Hence, we get that for every $t \ge 0$

$$\lim_{n\to\infty} \text{pmmse}_n(Y|\sqrt{t}Y+N) = \text{mmse}(Y|\sqrt{t}Y+N). \quad (76)$$

Now, we show that the convergence is uniform.

From expression (21) in Theorem 3, we have that

$$\lim_{t\to\infty} \text{pmmse}_n(Y|\sqrt{t}Y+N) = 0. \quad (77)$$

Further, by the convergence of the integral in the I-MMSE relation, we also know that

$$\lim_{t\to\infty} \text{mmse}(Y|\sqrt{t}Y+N) = 0. \quad (78)$$

Hence,

$$\lim_{t\to\infty} \text{pmmse}_n(Y|\sqrt{t}Y+N) - \text{mmse}(Y|\sqrt{t}Y+N) = 0. \quad (79)$$

By definition of the PMMSE as the minimum over sets of increasing size (in n), the sequence $\{\text{pmmse}_n(Y|\sqrt{t}Y+N)\}_{n\ge1}$ is decreasing. These properties are enough to conclude that the convergence in (76) is uniform, as we show next. Set

$$g_n(t) := \text{pmmse}_n(Y|\sqrt{t}Y+N) - \text{mmse}(Y|\sqrt{t}Y+N) \quad (80)$$

for short. Note that the g_n are nonnegative.

We have that $\{g_n\}_{n\ge1}$ is decreasing, and

$$\lim_{t\to\infty} g_n(t) = 0 = \lim_{n\to\infty} g_n(t). \quad (81)$$

Fix $\varepsilon > 0$. For each $n \ge 1$, let $C_{\varepsilon,n} = g_n^{-1}([\varepsilon,\infty))$. As $\{g_n\}_{n\ge1}$ 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, so each $C_{\varepsilon,n}$ is bounded. Hence, each $C_{\varepsilon,n}$ is compact. But, the intersection $\bigcap_{n\ge1} C_{\varepsilon,n}$ is empty, for if t_0 were in the intersection then $\liminf_{n\to\infty} g_n(t_0) \ge \varepsilon$ 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, i.e., there is an $m \in \mathbb{N}$ such that $\sup_{t\in[0,\infty)} g_n(t) < \varepsilon$ for every $n > m$. This is precisely uniform convergence.

Appendix E
Proof of Theorem 4

First, as $(X,Y) \perp\!\!\!\perp N$ by assumption,

$$I(X; Y+\gamma^{-1/2}N) = I(Y|\gamma) - \mathbb{E}_X\left[I(Y_X|\gamma)\right]. \quad (82)$$

Hence, by the I-MMSE formula,

$$\begin{aligned}I(X; Y+\gamma^{-1/2}N) = \frac{1}{2}\int_0^\gamma \text{mmse}(Y|\sqrt{t}Y+N) \\ - \mathbb{E}_X\left[\text{mmse}(Y_X|\sqrt{t}Y_X+N)\right] dt. \quad (83)\end{aligned}$$

Thus,

$$I(X;Y) = \lim_{\gamma\to\infty} I(X; Y+\gamma^{-1/2}N). \quad (84)$$

The uniform convergence of Theorem 2 applies to Y and each Y_x. Thus, interchanging the order of integration over $[0.\gamma]$ and taking the limit as $n \to \infty$, we obtain that

$$I(X; Y+\gamma^{-1/2}N) = \lim_{n\to\infty} I_n(X;Y|\gamma). \quad (85)$$

Hence,

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty} I_n(X;Y|\gamma), \quad (86)$$

as desired.

Appendix F
Proof of Theorem 5

By the strong law of large numbers, the sample k-th moment for each positive integer k converges almost surely to the k-th moment. Then, the continuous mapping theorem yields the desired result. Indeed, the coefficients $a_j^{(n)}$ and $b_j^{(n)}$ are continuous in the sample moments, so the rational function Θ_n along with its denominator are continuous functions of the sample moments too. Further, the numerator of Θ_n is of degree 2-less than the that of its denominator, which is strictly positive for any values of the sample moments. Thus, evaluations of Θ_n at arbitrary small perturbations of the samples are all majorized by an integrable function. Hence, we may apply Lebesgue's dominated convergence theorem to conclude that the integral of Θ_n over $[0,\infty)$ is continuous in the sample moments too; in particular, the integral over each $[0,\gamma]$ is continuous in the sample moments also. The continuous mapping theorem then yields the almost sure convergences.

Appendix G
Number of Samples

Lemma 9. *Let $f : [0,\infty) \to (0,\infty)$ and $g : [0,\infty) \to [0,\infty)$ be two nondecreasing continuous functions such that*

- *$f(t) > g(t)$ for every $t \in [0,\infty)$,*
- *$\limsup_{t\to\infty} g(\alpha t)/f(t) < 1$ for some $\alpha > 1$, and*
- *there is an $n > 0$ such that $g(xy) \le x^n g(y)$ for every $x \in [1,\infty)$ and $y \in [0,\infty)$.*

Then, for any $\varepsilon > 0$, there exists a $\beta > 1$ such that for any $(t,\gamma) \in [0,\infty) \times [1,\beta]$ one has

$$1 \le \frac{f(t)-g(t)}{f(t)-g(\gamma t)} \le 1+\varepsilon. \quad (87)$$

(which occurs, e.g., for $\beta = (1+\varepsilon(m-1)/(1+\varepsilon))^{1/n}$ where $m = \inf_{t\in[0,\infty)} f(t)/g(t)$).

Proof. First, we establish the existence of a $\beta \in (1,\alpha]$ such that $f(t) > g(\beta t)$ for every $t \in [0,\infty)$. Then, the left hand side of (87) is bounded; indeed, for such β and any $\gamma \in [1,\beta]$, the function $h(t) := (f(t)-g(t))/(f(t)-g(\gamma t))$ over $[0,\infty)$ is positive and continuous, and it satisfies $\limsup_{t\to\infty} h(t) < \infty$. Then, we will show that uniform bounds on h arbitrarily close to 1 are attainable.

For each $k \in \mathbb{N}$, consider the number

$$\delta_k := 1 + \frac{\alpha - 1}{k}, \tag{88}$$

the function $\ell_k : [0,\infty) \to \mathbb{R}$ defined by

$$\ell_k(t) = f(t) - g(\delta_k t), \tag{89}$$

and the pre-image

$$C_k := \ell_k^{-1}((-\infty, 0]). \tag{90}$$

As g is nondecreasing, the C_k are decreasing. Since $\delta_k \to 1$, continuity of g and the fact that $f(t) > g(t)$ for every $t \in [0,\infty)$ imply that the intersection $\bigcap_{k\in\mathbb{N}} C_k$ is empty. Since a decreasing sequence of nonempty compact sets is nonempty, it suffices to show that each C_k is compact; then, it must be that some C_{k_0} is empty, i.e., that $f(t) > g(\delta_{k_0} t)$ for every $t \in [0,\infty)$.

Fix $k \in \mathbb{N}$. To see that C_k is compact, one could note that it is both closed in $\mathbb{R}$ and bounded. As ℓ_k is continuous, C_k is closed in $[0,\infty)$, which is closed in $\mathbb{R}$, so C_k is closed in $\mathbb{R}$. It remains to show that C_k is bounded, for which it suffices to show that C_1 is bounded. Note that $\delta_1 = \alpha$. Since $\limsup_{t\to\infty} g(\alpha t)/f(t) < 1$ and f is positive, $f(t) - g(\alpha t) > 0$ for all large enough t, i.e., C_1 is bounded. Thus, C_k is compact.

Now that we established the existence of a $\beta \in (1,\alpha]$ such that $f(t) > g(\beta t)$ for every $t \in [0,\infty)$, we show that decreasing β, if necessary, ensures that the bound in (87) is attained. Let

$$m = \inf_{t\in[0,\infty)} \frac{f(t)}{g(t)}, \tag{91}$$

and

$$\beta = \min\left(\beta_0, \left(1 + \frac{\varepsilon(m-1)}{1+\varepsilon}\right)^{1/n}\right). \tag{92}$$

Note that for any $(\gamma, t) \in [1,\beta] \times [0,\infty)$ one has

$$\begin{aligned} g(\gamma t) &\le \gamma^n g(t) \\ &\le \left(1 + \frac{\varepsilon(m-1)}{1+\varepsilon}\right) g(t) \\ &\le g(t) + \frac{\varepsilon(f(t)-g(t))}{1+\varepsilon}, \end{aligned}$$

implying that

$$\begin{aligned} \left|\frac{f(t)-g(t)}{f(t)-g(\gamma t)}\right| &= \frac{f(t)-g(t)}{f(t)-g(\gamma t)} \\ &\le \frac{f(t)-g(t)}{f(t)-g(t)-\frac{\varepsilon(f(t)-g(t))}{1+\varepsilon}} \\ &= 1+\varepsilon. \end{aligned}$$

□

Lemma 10. *Let $f : [0,\infty) \to (0,\infty)$ and $g : [0,\infty) \to [0,\infty)$ be two nondecreasing continuous functions such that*

- *$f(t) > g(t)$ for every $t \in [0,\infty)$, and*
- *there is an $n > 0$ such that $f(xy) \le x^n f(y)$ for every $x \in [1,\infty)$ and $y \in [0,\infty)$.*

Then, for any $\varepsilon > 0$, there exists a $\beta > 1$ such that for any $(t,\gamma) \in [0,\infty) \times [1,\beta]$ one has

$$1 - \varepsilon \le \frac{f(t)-g(t)}{f(\gamma t)-g(t)} \le 1. \tag{93}$$

(which occurs, e.g., for $\beta = (1+\varepsilon(1-\mu)/(1-\varepsilon))^{1/n}$ where $\mu = \sup_{t\in[0,\infty)} g(t)/f(t)$).

Proof.

$$\begin{aligned} \frac{f(t)-g(t)}{f(\gamma t)-g(t)} &\ge \frac{f(t)-g(t)}{\gamma^n f(t)-g(t)} \ge \frac{f(t)-g(t)}{\beta^n f(t)-g(t)} \\ &= \frac{f(t)-g(t)}{\frac{1-\varepsilon\cdot\sup_{x\in[0,\infty)}\frac{g(x)}{f(x)}}{1-\varepsilon} f(t) - g(t)} \\ &\ge \frac{f(t)-g(t)}{\frac{1-\varepsilon\cdot\frac{g(t)}{f(t)}}{1-\varepsilon} f(t) - g(t)} \\ &= \frac{(f(t)-g(t))(1-\varepsilon)}{f(t)-\varepsilon g(t)-(1-\varepsilon)g(t)} = 1-\varepsilon. \end{aligned}$$

□

For an $\mathbb{R}$-valued random variable R that is supported over $[a,b]$, and for $r_1, \cdots, r_N \sim R$ i.i.d., Hoeffding's inequality says that

$$\Pr\left\{\left|\mathbb{E}R - \frac{1}{N}\sum_{i=1}^N r_i\right| \ge \delta\right\} \le 2e^{-2N\delta^2/(b-a)^2}.$$

Then, for $a > 0$,

$$\begin{aligned} \Pr&\left\{(1-\eta)\mathbb{E}R \le \frac{1}{N}\sum_{i=1}^N r_i \le (1+\eta)\mathbb{E}R\right\} \\ &\ge 1 - 2e^{-2N\eta^2/(b/a-1)^2}. \end{aligned}$$

For $n \in \mathbb{N}$, we obtain

$$\begin{aligned} \Pr&\left\{(1-\eta)\mathbb{E}R^k \le \frac{1}{N}\sum_{i=1}^N r_i^k \le (1+\eta)\mathbb{E}R^k \text{ for every } k \in [2n]\right\} \\ &\ge 1 - 4ne^{-2N\eta^2/\left((b/a)^{2n}-1\right)^2}. \end{aligned}$$

Lemma 11. *Let $f_0, g_0, u_0,$ and v_0 be four functions defined on $[0,\infty)$ such that f_0 and u_0 are strictly positive while g_0 and v_0 are nonnegative. Assume that*

- *u_0 and v_0 are monotonically nondecreasing,*
- *$u_0(t) > v_0(t)$ for every $t \in [0, \infty)$,*
- *$\limsup_{t\to\infty} v_0(\alpha t)/u_0(t) < 1$ for some $\alpha > 1$, and*
- *there is an $n > 0$ such that for every $x \geq 1$ and $y \geq 0$ one has $u_0(xy) \leq x^n u_0(y)$ and $v_0(xy) \leq x^n v_0(y)$.*

Then, for any functions

Now, consider positive f, g, u, and v, and

$$\frac{f(t) - g(t)}{u(t) - v(t)}.$$

Then,

$$\begin{aligned}
&\frac{(1-\eta)^2 f_0((1-\eta)^2 t) - (1+\eta)^2 g_0((1+\eta)^2 t)}{u_0((1+\eta)^2 t) - v_0((1-\eta)^2 t)} \\
&\leq \frac{f(t) - g(t)}{u(t) - v(t)} \\
&\leq \frac{(1+\eta)^2 f_0((1+\eta)^2 t) - (1-\eta)^2 g_0((1-\eta)^2 t)}{u_0((1-\eta)^2 t) - v_0((1+\eta)^2 t)}.
\end{aligned}$$

Integrating with respect to t over $[0, \infty)$ then replacing t with $(1-\eta)^2 t$, one obtains, with $\nu := (1+\eta)^2/(1-\eta)^2$,

$$\begin{aligned}
&\int_0^\infty \frac{f_0(t) - \nu g_0(\nu t)}{u_0(\nu t) - v_0(t)}\, dt \\
&\leq \int_0^\infty \frac{f(t) - g(t)}{u(t) - v(t)}\, dt \\
&\leq \int_0^\infty \frac{\nu f_0(\nu t) - g_0(t)}{u_0(t) - v_0(\nu t)}\, dt.
\end{aligned}$$

Set

$$\begin{aligned}
\delta(t) &:= \frac{u_0(t) - v_0(t)}{u_0(t) - v_0(\nu t)}, \quad \text{and} \\
\gamma(t) &:= \frac{u_0(t) - v_0(t)}{u_0(\nu t) - v_0(t)}.
\end{aligned}$$

Note that

$$\frac{\nu f_0(\nu t) - g_0(t)}{u_0(t) - v_0(\nu t)} = \delta(t) \left(\frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} + \frac{\nu f_0(\nu t) - f_0(t)}{u_0(t) - v_0(t)} \right).$$

Further,

$$\delta(t) \cdot \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} \leq \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} + \varepsilon \cdot \frac{f_0(t)}{u_0(t) - v_0(t)}$$

and

$$\delta(t) \cdot \frac{\nu f_0(\nu t) - f_0(t)}{u_0(t) - v_0(t)} \leq \frac{(1+\varepsilon)(\nu^{\binom{n+1}{2}-1} - 1) f_0(t)}{u_0(t) - v_0(t)}.$$

Hence,

$$\begin{aligned}
&\int_0^\infty \frac{f(t) - g(t)}{u(t) - v(t)}\, dt \\
&\leq \int_0^\infty \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)}\, dt \\
&+ \left((1+\varepsilon)\nu^{\binom{n+1}{2}-1} - 1 \right) \int_0^\infty \frac{f_0(t)}{u_0(t) - v_0(t)}\, dt.
\end{aligned}$$

On the other hand,

$$\frac{f_0(t) - \nu g_0(\nu t)}{u_0(\nu t) - v_0(t)} = \gamma(t) \left(\frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} + \frac{g_0(t) - \nu g_0(\nu t)}{u_0(t) - v_0(t)} \right),$$

while

$$\gamma(t) \cdot \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} \geq \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)} - \varepsilon \cdot \frac{f_0(t)}{u_0(t) - v_0(t)}$$

and

$$\gamma(t) \cdot \frac{g_0(t) - \nu g_0(\nu t)}{u_0(t) - v_0(t)} \geq \left(1 - \nu^{\binom{n+1}{2}-1} \right) \frac{g_0(t)}{u_0(t) - v_0(t)}.$$

Hence,

$$\begin{aligned}
&\int_0^\infty \frac{f(t) - g(t)}{u(t) - v(t)}\, dt \\
&\geq \int_0^\infty \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)}\, dt - \varepsilon \int_0^\infty \frac{f_0(t)}{u_0(t) - v_0(t)}\, dt \\
&- \left(\nu^{\binom{n+1}{2}-1} - 1 \right) \int_0^\infty \frac{g_0(t)}{u_0(t) - v_0(t)}\, dt.
\end{aligned}$$

Thus,

$$\begin{aligned}
&-\varepsilon A - \left(\nu^{\binom{n+1}{2}-1} - 1 \right) B \\
&\leq \int_0^\infty \frac{f(t) - g(t)}{u(t) - v(t)}\, dt - \int_0^\infty \frac{f_0(t) - g_0(t)}{u_0(t) - v_0(t)}\, dt \\
&\leq \left((1+\varepsilon)\nu^{\binom{n+1}{2}-1} - 1 \right) A
\end{aligned}$$

where

$$\begin{aligned}
A &:= \int_0^\infty \frac{f_0(t)}{u_0(t) - v_0(t)}\, dt, \quad \text{and} \\
B &:= \int_0^\infty \frac{g_0(t)}{u_0(t) - v_0(t)}\, dt.
\end{aligned}$$

References

[1] H. Goodarzi et al. “Systematic Discovery of Structural Elements Governing Stability of Mammalian Messenger RNAs,” *Nature*, vol. 485, iss. 7397, pp. 264-268, May 2012.

[2] M. S. Carro et al. “The Transcriptional Network for Mesenchymal Transformation of Brain Tumours,” *Nature*, vol. 463, iss. 7279, pp. 318-325, Jan. 2010.

[3] F. Fleuret, “Fast Binary Feature Selection with Conditional Mutual Information,” *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[4] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating Mutual Information,” *Physical Review E*, vol. 69, iss. 6, 066138, Jun. 2004.

[5] G. Valiant and P. Valiant, “Estimating the Unseen: an $n/\log(n)$-sample Estimator for Entropy and Support Size, Shown Optimal via New CLTs,” in *Proc. 43rd STOC*, pp. 685–694, ACM, 2011.

[6] J. Jiao, “Minimax Estimation of Functionals of Discrete Distributions,” *IEEE Trans. Information Theory*, vol. 61, iss. 5, pp. 2835-2885, 2015.

[7] Y. Wu and P. Yang, “Minimax Rates of Entropy Estimation on Large Alphabets via Best Polynomial Approximation,” *IEEE Trans. Information Theory*, vol. 62, iss. 6, pp. 3702-3720, 2016.

[8] W. Gao et al. “Estimating Mutual Information for Discrete-Continuous Mixtures,” in *Advances in Neural Information Processing Systems*, pp. 5988–5999, 2017.

[9] D. Guo, S. Shamai, and S. Verdu, “Mutual Information and Minimum Mean-squared Error in Gaussian Channels,” *IEEE Trans. Information Theory*, vol. 51, iss. 4, pp. 1261-1282, 2005.

[10] Z. Goldfeld, K. Greenewald, and Y. Polyanskiy, “Estimating Differential Entropy under Gaussian Convolutions,” arXiv:1810.11589v2, Oct. 2018.

[11] W. Rudin, *Functional Analysis*, 2nd ed. McGraw-Hill, 2006.