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

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

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.

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

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

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

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 $\gamma > 0$,*

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

We have the equation

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

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

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

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 $t \mapsto \operatorname{pmmse}_n(R|\sqrt{t}R + N)$ 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*

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

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 $|\operatorname{supp}(Z)| = \infty$, then*

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

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

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

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

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

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.

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

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

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)

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

where we may compute

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

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

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

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 $n \in \mathbb{N}$ such that $\mathcal{S}_n$ is nonempty, let $(U^{(n)}, V^{(n)}) \sim \text{Unif}(\mathcal{S}_n)$ 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.

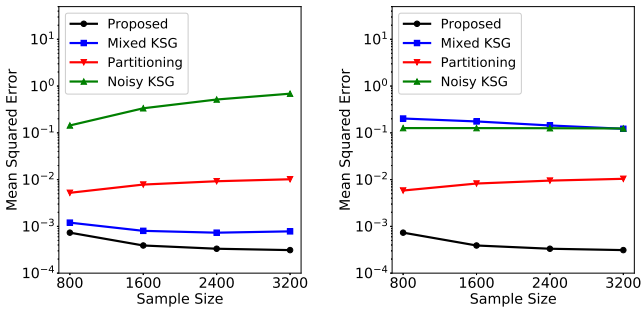
If Y is of bounded support, Hoeffding's inequality implies that there exists a constant $C = C_{X,Y,n}$ such that having $|\mathcal{S}| > (C/\varepsilon^2)\log(1/\delta)$ as $\varepsilon, \delta \to 0$ ensures that $\Pr\left\{|\widehat{I}_n(\mathcal{S}) - I_n(X;Y)| < \varepsilon\right\} \ge 1 - \delta$ (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.

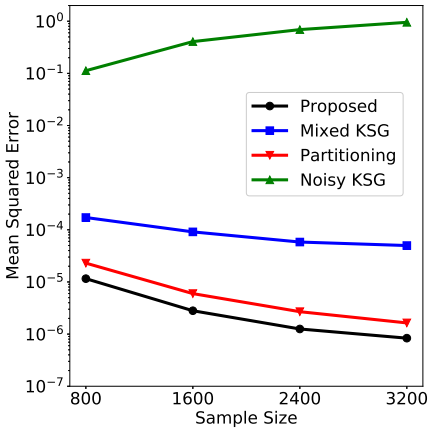


Figure 2. Mean Squared Error vs. Sample Size when estimating $I(X;Y)$ for $X \perp\!\!\!\perp Y$ with X uniform over $\{0, 1\}$ and Y uniform over $[0, 2]$.

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 Formula

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 W and Z^n are square-integrable and $|\mathrm{supp}(Z)| > n$. We utilize equation (58) to derive explicit formulas for the function $t \mapsto \mathrm{pmmse}_n(R|\sqrt{t}R + N)$ by setting $W = R$ and $Z = \sqrt{t}R+N$. 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 $[n]$. We define

- for each $(i,j) \in [n]^2$, the subset $T_n^{(i,j)} \subset S_n^{(0)}$ as the collection of permutations sending i to j, i.e.,

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

- 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

$$Q_R(\pi; i_1, \cdots, i_m) := \prod_{k\notin\{i_1,\cdots,i_m\}} \mathcal{R}_{k+\pi(k)}, \tag{66}$$

and

$$Q_R(\pi) := \prod_{k\in[n]} \mathcal{R}_{k+\pi(k)}. \tag{67}$$

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

$$D_{R,n}(t) = \sum_{(i,j)\in[n]^2} d_{R,n}^{(i,j)}(t), \tag{68}$$

and by Cramer's rule

$$F_{R,n}(t) = \frac{D_{R,n}(t)}{\det M_{\sqrt{t}R+N,n}}. \tag{69}$$

The relation (69) implies, in view of equation (58), that

$$\mathrm{pmmse}_n(R|\sqrt{t}R+N) = \mathcal{R}_2 - \frac{D_{R,n}(t)}{\det M_{\sqrt{t}R+N,n}}. \tag{70}$$

We first show that the PMMSE is rational, then show that the specific properties in Theorem 3 hold. To show rationality, we prove that both $D_{R,n}(t)$ and $\det M_{\sqrt{t}R+N,n}$ are polynomials in Appendix C-A, then prove further characterization about the coefficients in Appendix C-B.

A. Rationality

We start by showing that the mapping $t \mapsto \det M_{\sqrt{t}R+N,n}$ is a polynomial of degree $\binom{n+1}{2}$ with leading coefficient $\det M_R$.

The fact that $\mathbb{E}[N^r] = 0$ for odd naturals r simplifies the expressions involved. Let ℓ be a nonnegative integer and S be an $\mathbb{R}$-valued random variable such that $S \perp\!\!\!\perp N$. If ℓ is even, then

$$\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^\ell\right] = \sum_{k \text{ even}} \binom{\ell}{k} t^{k/2}\mathbb{E}SR^k\mathbb{E}N^{\ell-k}, \tag{71}$$

whereas if ℓ is odd then

$$t^{-1/2}\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right]=\sum_{k\text{ odd}}\binom{\ell}{k}t^{(k-1)/2}\mathbb{E}SR^{k}\mathbb{E}N^{\ell-k}. \tag{72}$$

Both expressions on the right hand side of (71) and (72) are polynomials of degree at most $\lfloor \ell/2 \rfloor$. Further, the coefficient of $t^{\lfloor \ell/2 \rfloor}$ in either polynomial is $\mathbb{E}SR^{\ell}$. As the polynomials in (71) and (72) will occur repeatedly, we use the following shorthand. For ℓ even, we set

$$e_{S,R,\ell}(t):=\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right], \tag{73}$$

and for ℓ odd we set

$$o_{S,R,\ell}(t):=t^{-1/2}\mathbb{E}\left[S\left(\sqrt{t}R+N\right)^{\ell}\right]. \tag{74}$$

Both $e_{S,R,\ell}$ and $o_{S,R,\ell}$ are polynomials.

We may write the entries of $M_{\sqrt{t}R+N,n}$ in terms of the auxiliary polynomials $e_{1,R,\ell}$ and $o_{1,R,\ell}$. If $i+j$ is even, then

$$\left(M_{\sqrt{t}R+N,n}\right)_{i,j}=e_{1,R,i+j}(t), \tag{75}$$

while if $i+j$ is odd then

$$\left(M_{\sqrt{t}R+N,n}\right)_{i,j}=\sqrt{t}o_{1,R,i+j}(t). \tag{76}$$

The key ingredient in proving that $\det M_{\sqrt{t}R+N,n}$ and $D_{R,n}(t)$ are polynomials is the observation that, for a permutation $\pi\in S_n^{(0)}$, there is an even number of indices $i\in[n]$ such that the integer $i+\pi(i)$ is odd. We state this fact as a lemma.

Lemma 4. *For any permutation $\pi\in S_n^{(0)}$, the number*

$$|\{i\in[n]\,;\ i+\pi(i)\textit{ is odd }\}| \tag{77}$$

is even.

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

$$I:=\{i\in\{1,\cdots,n\}\,;\ (i,\pi(i))\in(A\times B)\cup(B\times A)\} \tag{78}$$

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.,

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

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

For $\pi\in S_n^{(0)}$, we use the notation

$$\delta(\pi):=|\{i\in[n]\,;\ i+\pi(i)\text{ is odd }\}|. \tag{79}$$

By Lemma 4, the integer $\delta(\pi)$ is always even.

We now show that $\det M_{\sqrt{t}R+N,n}$ is a polynomial. By the Leibniz formula for the determinant, we may write

$$\det M_{\sqrt{t}R+N,n}=\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\prod_{r\in[n]}\mathbb{E}\left[\left(\sqrt{t}R+N\right)^{r+\pi(r)}\right]. \tag{80}$$

With the auxiliary polynomials defined in (73) and (74), we may write

$$\det M_{\sqrt{t}R+N,n} \tag{81}$$

$$=\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)t^{\delta(\pi)/2}\prod_{i\,:\,i+\pi(i)\text{ odd}}o_{1,R,i+\pi(i)}(t)\prod_{j\,:\,j+\pi(j)\text{ even}}e_{1,R,j+\pi(j)}(t), \tag{82}$$

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

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

Finally, taking the terms of highest degrees in (80), we obtain that the coefficient of $t^{n(n+1)/2}$ in $\det M_{\sqrt{t}R+N,n}$ is

$$\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\prod_{r\in[n]}\mathcal{R}_{r+\pi(r)}, \tag{83}$$

which is $\det M_{R,n}$ by the Leibniz determinant formula. This coefficient is nonnegative, and it is nonzero if and only if $|\operatorname{supp}(R)|>n$.

Next, we show that $D_{R,n}$ is a polynomial. We start with a characterization of the cofactors $c_{R,n}^{(i,j)}$. Namely, we show that 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.

If $i+j$ is even, then

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

whereas if $i+j$ is odd then

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

Thus, evenness of $\delta(\pi)$ for each $\pi\in S_n^{(0)}$ implies that each $c_{R,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is even and that each

$\sqrt{t}c_{R,n}^{(i,j)}(t)$ is a polynomial when $i+j$ is odd. Further, the degree of $c_{R,n}^{(i,j)}$ for even $i+j$ is upper bounded by

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

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

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

We note that both upper bounds are equal to

$$\frac{n(n+1)}{2} - \left\lfloor \frac{i+j}{2} \right\rfloor. \tag{84}$$

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

$$\sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) Q_R(\pi; i). \tag{85}$$

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

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

If $i+j$ is even, so i and j have the same parity, then

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

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

Thus, from the characterization of $c_{R,n}^{(i,j)}$, regardless of the parity of $i+j$ we obtain that $d_{R,n}^{(i,j)}(t)$ is a polynomial in t of degree at most $n(n+1)/2$ with the coefficient of $t^{n(n+1)/2}$ being

$$\mathcal{R}_{i+1}\mathcal{R}_{j+1} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) Q_R(\pi; i). \tag{86}$$

Hence, the coefficient of $t^{n(n+1)/2}$ in $D_{R,n}(t)$ is

$$\sum_{(i,j)\in[n]^2} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1}\mathcal{R}_{j+1} Q_R(\pi; i). \tag{87}$$

We show in Appendix C-B that this coefficient is equal to

$$\mathcal{R}_2 \det M_{R,n}, \tag{88}$$

thereby allowing for expressions for the PMMSE to simplify.

B. Leading Coefficients

We first show that the coefficient of $t^{n(n+1)/2}$ in $D_{R,n}(t)$ is $\mathcal{R}_2 \det M_{R,n}$, i.e., that

$$\sum_{(i,j)\in[n]^2} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1}\mathcal{R}_{j+1} Q_R(\pi; i) = \mathcal{R}_2 \det M_{R,n}. \tag{89}$$

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

$$S_n^{(0)} = \bigcup_{j\in[n]} T_n^{(i,j)}. \tag{90}$$

Thus, we may write

$$\sum_{(i,j)\in[n]^2} \sum_{\pi \in T_n^{(i,j)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1}\mathcal{R}_{j+1} Q_R(\pi; i) = \sum_{i\in[n]} \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1} Q_R(\pi; i). \tag{91}$$

In the sequel, we will denote for $i \in [n]$ and $\pi \in S_n^{(0)}$

$$\pi_i := \pi \circ (1\ i). \tag{92}$$

If $i = 1$, then $\pi_i = \pi$. Further, for each $i \in [n]$, as multiplication by $(1\ \ i)$ is an automorphism of $S_n^{(0)}$, the mapping $\pi \mapsto \pi_i$ is a bijection of $S_n^{(0)}$. In addition, when $i \neq 1$,

$$\operatorname{sgn}(\pi_i) = -\operatorname{sgn}(\pi). \tag{93}$$

Now, for each $i \in [n] \setminus \{1\}$, we have that

$$\begin{aligned}
&\sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{\pi(i)+1} Q_R(\pi; i) \\
&\quad= \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{\pi(i)+1}\mathcal{R}_{\pi(1)+1} Q_R(\pi; 1, i) \\
&\quad= \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi_i) \mathcal{R}_{\pi_i(i)+1}\mathcal{R}_{\pi_i(1)+1} Q_R(\pi_i; 1, i) \\
&\quad= -\sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{\pi(1)+1}\mathcal{R}_{\pi(i)+1} Q_R(\pi; 1, i) \\
&\quad= -\sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{\pi(i)+1} Q_R(\pi; i).
\end{aligned} \tag{94}$$

Hence,

$$\sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_{\pi(i)+1} Q_R(\pi; i) = 0. \tag{95}$$

Thus, only $i = 1$ could give a nonzero sum in (91). Furthermore, when $i = 1$ in (91), we obtain the sum

$$\begin{aligned}
&\sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) \mathcal{R}_2 \mathcal{R}_{\pi(1)+1} Q_R(\pi; 1) \\
&\quad= \mathcal{R}_2 \sum_{\pi \in S_n^{(0)}} \operatorname{sgn}(\pi) Q_R(\pi) \\
&\quad= \mathcal{R}_2 \det M_{R,n}.
\end{aligned} \tag{96}$$

Thus, (89) follows.

Therefore, equation (70) yields that the equation

$$\mathrm{pmmse}_n(R|\sqrt{t}R+N) = \frac{\mathcal{R}_2 \det M_{\sqrt{t}R+N,n} - D_{R,n}(t)}{\det M_{\sqrt{t}R+N,n}} \tag{97}$$

is a representation of the PMMSE as a rational function, where the numerator is of degree at most $n(n+1)/2-1$ and the denominator is of degree at most $n(n+1)/2$. Further, if $|\mathrm{supp}(R)| > n$ then the degree of denominator is exactly $n(n+1)/2$, and its leading coefficient is $\det M_{R,n}$. Next, we apply similar bijectivity tricks to show that the coefficient of $t^{n(n+1)/2-1}$ in the numerator is $\det M_{R,n}$.

Via equation (80) and the defining equations for $D_{R,n}$, a preliminary formula for the coefficient of $t^{n(n+1)/2-1}$ in the numerator $\mathcal{R}_2 \det M_{\sqrt{t}R+N,n} - D_{R,n}(t)$ is

$$\begin{aligned}
&\mathcal{R}_2 \sum_{r\in[n]} \sum_{\pi\in S_n^{(0)}} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; r) \\
&\quad - \sum_{(i,j)\in[n]^2} \sum_{\pi\in T_n^{(i,j)}} \sum_{r\neq i} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \\
&\qquad\qquad \mathcal{R}_{i+1}\mathcal{R}_{j+1}\mathcal{R}_{r+\pi(r)-2} Q_R(\pi; i, r) \\
&\quad - \sum_{(i,j)\in[n]^2} \sum_{\pi\in T_n^{(i,j)}} \mathrm{sgn}(\pi) Q_R(\pi; i) \\
&\qquad\qquad \left(\binom{i}{2} \mathcal{R}_{i-1}\mathcal{R}_{j+1} + \binom{j}{2} \mathcal{R}_{i+1}\mathcal{R}_{j-1} \right),
\end{aligned}$$

where we set $\mathcal{R}_\ell = 0$ when $\ell < 0$. We will deal with each of the three sums in this preliminary formula separately; so, denote the three sums, in order, by $\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3$ (where, for $\mathfrak{S}_1$, we absorb the factor $\mathcal{R}_2$ inside the sum), i.e., define

$$\mathfrak{S}_1 := \sum_{r\in[n]} \sum_{\pi\in S_n^{(0)}} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \mathcal{R}_2 \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; r),$$

$$\begin{aligned}
\mathfrak{S}_2 := - \sum_{(i,j)\in[n]^2} \sum_{\pi\in T_n^{(i,j)}} \sum_{r\neq i} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \\
\mathcal{R}_{i+1}\mathcal{R}_{j+1}\mathcal{R}_{r+\pi(r)-2} Q_R(\pi; i, r),
\end{aligned}$$

$$\begin{aligned}
\mathfrak{S}_3 := - \sum_{(i,j)\in[n]^2} \sum_{\pi\in T_n^{(i,j)}} \mathrm{sgn}(\pi) Q_R(\pi; i) \\
\left(\binom{i}{2} \mathcal{R}_{i-1}\mathcal{R}_{j+1} + \binom{j}{2} \mathcal{R}_{i+1}\mathcal{R}_{j-1} \right).
\end{aligned}$$

Thus, the coefficient of $t^{n(n+1)/2-1}$ in the numerator $\mathcal{R}_2 \det M_{\sqrt{t}R+N,n} - D_{R,n}(t)$ is

$$\mathfrak{S}_1 + \mathfrak{S}_2 + \mathfrak{S}_3. \tag{98}$$

We show that this coefficient is equal to $\det M_{R,n}$ by showing that

$$\mathfrak{S}_1 + \mathfrak{S}_2 = \det M_{R,n} \tag{99}$$

and that

$$\mathfrak{S}_3 = 0. \tag{100}$$

For $i_1, \cdots, i_\ell \in [n]$, let $[n]_{i_1,\cdots,i_\ell} = [n] \setminus \{i_1, \cdots, i_\ell\}$.

For the first sum, $\mathfrak{S}_1$, we partition $[n] \times S_n^{(0)}$ into three parts as

$$\left(\{1\} \times T_n^{(1,1)}\right) \cup \left(\{1\} \times \bigcup_{j\in[n]_1} T_n^{(1,j)} \right) \cup \left([n]_1 \times S_n^{(0)} \right), \tag{101}$$

and we let $\mathfrak{S}_1 = \mathfrak{S}_{1,1} + \mathfrak{S}_{1,2} + \mathfrak{S}_{1,3}$ be the ensuing decomposition, which we express next. For the first part in (101), we obtain

$$\mathfrak{S}_{1,1} = \sum_{\pi\in T_n^{(1,1)}} \mathrm{sgn}(\pi) Q_R(\pi), \tag{102}$$

whereas the second and third parts give

$$\mathfrak{S}_{1,2} = \sum_{j\in[n]_1} \sum_{\pi\in T_n^{(1,j)}} \mathrm{sgn}(\pi) \binom{j+1}{2} \mathcal{R}_2 \mathcal{R}_{j-1} Q_R(\pi; 1) \tag{103}$$

and

$$\mathfrak{S}_{1,3} = \sum_{(r,\pi)\in[n]_1\times S_n^{(0)}} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \mathcal{R}_2 \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; r), \tag{104}$$

respectively. We will show that both $\mathfrak{S}_{1,2}$ and $\mathfrak{S}_{1,3}$ cancel out identically when summed with parts of the sum $\mathfrak{S}_2$. We also note that $\mathfrak{S}_{1,1}$ provides part of the sum that will ultimately produce $\det M_{R,n}$; the remaining part lies in $\mathfrak{S}_2$, which we treat next.

Let

$$U_n^{(i,j)} = \left\{ (i,j,\pi) \; ; \; \pi \in T_n^{(i,j)} \right\}. \tag{105}$$

For the second sum, $\mathfrak{S}_2$, we employ the partition

$$\bigcup_{(i,j)\in[n]^2} \left(U_n^{(i,j)} \times [n]_i \right) = \mathfrak{p}_1 \cup \mathfrak{p}_2 \cup \mathfrak{p}_3 \cup \mathfrak{p}_4 \tag{106}$$

where

$$\mathfrak{p}_1 := \bigcup_{j\in[n]} \left(U_n^{(1,j)} \times [n]_1 \right) \tag{107}$$

$$\mathfrak{p}_2 := \bigcup_{i\in[n]_1} \left(U_n^{(i,1)} \times \{1\} \right) \tag{108}$$

$$\mathfrak{p}_3 := \bigcup_{(i,j)\in[n]_1\times[n]} \left(U_n^{(i,j)} \times [n]_{1,i} \right) \tag{109}$$

$$\mathfrak{p}_4 := \bigcup_{(i,j)\in[n]_1^2} \left(U_n^{(i,j)} \times \{1\} \right). \tag{110}$$

We will denote the ensuing sums by $\mathfrak{S}_{2,1}, \mathfrak{S}_{2,2}, \mathfrak{S}_{2,3}, \mathfrak{S}_{2,4}$, which we express next. We will denote a generic element $((i,j,\pi),r) \in U_n^{(i,j)} \times [n]_i$ as (i,j,π,r) for short. The $\mathfrak{p}_1$-part yields

$$\mathfrak{S}_{2,1} = - \sum_{(r,\pi)\in[n]_1\times S_n^{(0)}} \mathrm{sgn}(\pi) \binom{r+\pi(r)}{2} \mathcal{R}_2 \mathcal{R}_{r+\pi(r)-2} Q_R(\pi; r), \tag{111}$$

the $\mathfrak{p}_2$-part yields

$$\mathfrak{S}_{2,2} = -\sum_{i\in[n]_1}\sum_{\pi\in T_n^{(i,1)}}\operatorname{sgn}(\pi)\binom{\pi(1)+1}{2}\mathcal{R}_2\mathcal{R}_{\pi(1)-1}Q_R(\pi;1), \tag{112}$$

and the $\mathfrak{p}_3$-part yields

$$\mathfrak{S}_{2,3} = -\sum_{i\in[n]_1}\sum_{r\in[n]_{1,i}}\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\binom{r+\pi(r)}{2} \mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{r+\pi(r)-2}Q_R(\pi;i,r). \tag{113}$$

We further partition the $\mathfrak{p}_4$-part according to whether permutations fix 1, namely,

$$\mathfrak{p}_4 = \mathfrak{p}_{4,1}\cup\mathfrak{p}_{4,2} \tag{114}$$

where

$$\mathfrak{p}_{4,1} := \{(i,j,\pi,1)\in D\ ;\ \pi(1)=1\} \tag{115}$$
$$\mathfrak{p}_{4,2} := \{(i,j,\pi,1)\in D\ ;\ \pi(1)\neq 1\}. \tag{116}$$

We denote the ensuing sums by $\mathfrak{S}_{2,4,1},\mathfrak{S}_{2,4,2}$. The $\mathfrak{p}_{4,1}$-part gives

$$\mathfrak{S}_{2,4,1} = -\sum_{(i,j)\in[n]_1^2}\sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,1)}}\operatorname{sgn}(\pi)\mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1}Q_R(\pi;i,1), \tag{117}$$

whereas the $\mathfrak{p}_{4,2}$-part gives

$$\mathfrak{S}_{2,4,2} = -\sum_{(i,j,k)\in[n]_1^3}\sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,k)}}\operatorname{sgn}(\pi)\binom{\pi(1)+1}{2} \mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{\pi(1)-1}Q_R(\pi;1,i). \tag{118}$$

With this decomposition of the coefficient of $t^{n(n+1)/2-1}$ in the numerator $\mathcal{R}_2\det M_{\sqrt{t}R+N,n} - D_{R,n}(t)$ at hand, we proceed to show that equations (99) and (100) holds by showing that the following six equations hold. We will show that

$$\mathfrak{S}_{1,1}+\mathfrak{S}_{2,4,1} = \det M_{R,n}, \tag{119}$$
$$\mathfrak{S}_{1,2}+\mathfrak{S}_{2,2} = 0, \tag{120}$$
$$\mathfrak{S}_{1,3}+\mathfrak{S}_{2,1} = 0, \tag{121}$$
$$\mathfrak{S}_{2,3} = 0, \tag{122}$$
$$\mathfrak{S}_{2,4,2} = 0, \tag{123}$$
$$\mathfrak{S}_3 = 0. \tag{124}$$

We first show that (119) holds. From (102), we have that

$$\mathfrak{S}_{1,1} = \sum_{\pi\in T_n^{(1,1)}}\operatorname{sgn}(\pi)Q_R(\pi). \tag{125}$$

We show that $\mathfrak{S}_{2,4,1}$ complements this summation to give $\det M_{R,n}$, i.e., to get that (119) holds. From Leibniz formula for the determinant, we have that

$$\det M_{R,n} = \sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)Q_R(\pi). \tag{126}$$

From the partition $S_n^{(0)} = \bigcup_{i\in[n]} T_n^{(i,1)}$ (similar to the partition in (90)), then, it suffices to show that

$$\mathfrak{S}_{2,4,1} = \sum_{i\in[n]_1}\sum_{\pi\in T_n^{(i,1)}}\operatorname{sgn}(\pi)Q_R(\pi). \tag{127}$$

We proceed to show that (127) holds. We employ a similar technique to how we showed (89). Fix $i\in[n]_1$. By the change of variables $\sigma=\pi_i$ (equivalently, $\pi=\sigma_i$, since $(1\ i)^{-1} = (1\ i)$) we have that

$$\begin{aligned}
&\sum_{j\in[n]_1}\sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,1)}}\operatorname{sgn}(\pi)\mathcal{R}_{i+1}\mathcal{R}_{\pi(i)+1}Q_R(\pi;i,1)\\
&=\sum_{j\in[n]_1}\sum_{\sigma\in T_n^{(i,1)}\cap T_n^{(1,j)}}\operatorname{sgn}(\sigma_i)\mathcal{R}_{i+1}\mathcal{R}_{\sigma_i(i)+1}Q_R(\sigma_i;i,1)\\
&=-\sum_{j\in[n]_1}\sum_{\sigma\in T_n^{(i,1)}\cap T_n^{(1,j)}}\operatorname{sgn}(\sigma)\mathcal{R}_{i+\sigma(i)}\mathcal{R}_{\sigma(1)+1}Q_R(\sigma;i,1)\\
&=-\sum_{j\in[n]_1}\sum_{\sigma\in T_n^{(i,1)}\cap T_n^{(1,j)}}\operatorname{sgn}(\sigma)Q_R(\sigma)\\
&=-\sum_{\sigma\in T_n^{(i,1)}}\operatorname{sgn}(\sigma)Q_R(\sigma).
\end{aligned}$$

Summing over all $i\in[n]_1$ and noting the minus sign in the definition of $\mathfrak{S}_{2,4,1}$ in (117), we obtain that (127) holds. Hence, equation (119) holds. We now show that the other parts give a vanishing contribution, i.e., that (120)-(124) all hold.

Equations (120) and (121) follow from the expressions we we give in (103),(104),(111),(112). For (120), we note that each j in the summand in (103) may be replaced with $\pi(1)$ as $\pi\in T_n^{(1,j)}$. Then, as

$$\bigcup_{j\in[n]_1} T_n^{(1,j)} = \bigcup_{i\in[n]_1} T_n^{(i,1)} \tag{128}$$

are both partitions of the same set, namely, the set of permutations that do not fix 1, equations (103) and (112) yield that $\mathfrak{S}_{1,2} = -\mathfrak{S}_{2,2}$, i.e., (120) holds. For (121), the expressions in (104) and (111) show that $\mathfrak{S}_{1,3} = -\mathfrak{S}_{2,1}$, i.e., that (121) holds.

Next, we show that (122) holds. Fix $i\in[n]_1$ and $r\in[n]_{1,i}$. We will show that the following sum vanishes

$$\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\binom{r+\pi(r)}{2}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{r+\pi(r)-2}Q_R(\pi;i,r) = 0. \tag{129}$$

As $\mathfrak{S}_{2,3}$, according to equation (113), is a linear combination of such sums, we would obtain that $\mathfrak{S}_{2,3}=0$, i.e., that (122) holds. To show that (129) holds, we utilize that $\pi\mapsto\pi_i$ is an

automorphism of $S_n^{(0)}$, as follows. We have that

$$\begin{aligned}
&\sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\binom{r+\pi(r)}{2}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{r+\pi(r)-2}Q_R(\pi;i,r)\\
&= \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\binom{r+\pi(r)}{2}\mathcal{R}_{\pi(i)+1}\mathcal{R}_{r+\pi(r)-2}\\
&\qquad\qquad \mathcal{R}_{\pi(1)+1}Q_R(\pi;1,i,r)\\
&= \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi_i)\binom{r+\pi_i(r)}{2}\mathcal{R}_{\pi_i(i)+1}\mathcal{R}_{r+\pi_i(r)-2}\\
&\qquad\qquad \mathcal{R}_{\pi_i(1)+1}Q_R(\pi_i;1,i,r)\\
&= -\sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\binom{r+\pi(r)}{2}\mathcal{R}_{\pi(1)+1}\mathcal{R}_{r+\pi(r)-2}\\
&\qquad\qquad \mathcal{R}_{\pi(i)+1}Q_R(\pi;1,i,r)\\
&= -\sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\binom{r+\pi(r)}{2}\mathcal{R}_{r+\pi(r)-2}\mathcal{R}_{\pi(i)+1}Q_R(\pi;i,r),
\end{aligned}$$

so the vanishing in (129) holds. Hence, (122) holds.

Next, we show that (123) holds. We rewrite (118) as

$$\mathfrak{S}_{2,4,2} = -\sum_{(i,j,k,\ell)\in[n]_1^4}\ \sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,k)}\cap T_n^{(\ell,1)}} \operatorname{sgn}(\pi)\binom{k+1}{2} \mathcal{R}_{i+1}\mathcal{R}_{j+1}\mathcal{R}_{k-1}\mathcal{R}_{\ell+1}Q_R(\pi;1,i,\ell). \tag{130}$$

We fix $(j,k)\in[n]_1^2$ and show the vanishing of each of the following sums

$$\sum_{(i,\ell)\in[n]_1^2} \mathcal{R}_{i+1}\mathcal{R}_{\ell+1} \sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,k)}\cap T_n^{(\ell,1)}} \operatorname{sgn}(\pi)Q_R(\pi;1,i,\ell) = 0. \tag{131}$$

From (130), we may write $\mathfrak{S}_{2,4,2}$ as a linear combination of such sums, so we would obtain that $\mathfrak{S}_{2,4,2}=0$. We show (132) next. We change variables as $\sigma=\pi\circ(i\ \ \ell)$ in the inner sum in (132) to obtain that

$$\begin{aligned}
&\sum_{\pi\in T_n^{(i,j)}\cap T_n^{(1,k)}\cap T_n^{(\ell,1)}} \operatorname{sgn}(\pi)Q_R(\pi;1,i,\ell)\\
&= \sum_{\sigma\in T_n^{(\ell,j)}\cap T_n^{(1,k)}\cap T_n^{(i,1)}} \operatorname{sgn}(\sigma\circ(i\ \ \ell))Q_R(\sigma\circ(i\ \ \ell);1,i,\ell)\\
&= -\sum_{\sigma\in T_n^{(\ell,j)}\cap T_n^{(1,k)}\cap T_n^{(i,1)}} \operatorname{sgn}(\sigma)Q_R(\sigma;1,i,\ell).
\end{aligned} \tag{132}$$

Multiplying by $\mathcal{R}_{i+1}\mathcal{R}_{\ell+1}$ then summing over $(i,\ell)\in[n]_1^2$, we obtain that the quantity on the left hand side of (132) is equal to its negative. Hence (132) holds, and we obtain that $\mathfrak{S}_{2,4,2}=0$.

Finally, we show that $\mathfrak{S}_3=0$. We may write

$$\begin{aligned}
\mathfrak{S}_3 = &-\sum_{i\in[n]_1}\binom{i}{2}\mathcal{R}_{i-1}\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(i)+1}Q_R(\pi;i)\\
&-\sum_{j\in[n]_1}\binom{j}{2}\mathcal{R}_{j-1}\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}Q_R(\pi;\pi^{-1}(j)).
\end{aligned} \tag{133}$$

We will show that for each $(i,j)\in[n]_1^2$,

$$\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(i)+1}Q_R(\pi;i)=0 \tag{134}$$

and

$$\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}Q_R(\pi;\pi^{-1}(j))=0. \tag{135}$$

Together, equations (134) and (135) imply in view of (133) that $\mathfrak{S}_3=0$. To show that (134) holds, we apply the automorphism $\pi\mapsto\pi_i$ of $S_n^{(0)}$, from which we obtain

$$\begin{aligned}
&\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(i)+1}Q_R(\pi;i)\\
&= \sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(i)+1}\mathcal{R}_{\pi(1)+1}Q_R(\pi;1,i)\\
&= \sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi_i)\mathcal{R}_{\pi_i(i)+1}\mathcal{R}_{\pi_i(1)+1}Q_R(\pi_i;1,i)\\
&= -\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(1)+1}\mathcal{R}_{\pi(i)+1}Q_R(\pi;1,i)\\
&= -\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(i)+1}Q_R(\pi;i)
\end{aligned} \tag{136}$$

and (134) follows. Now, we show that (135) reduces to (134) via the automorphism $\pi\mapsto\pi^{-1}$ of $S_n^{(0)}$. First, note that for any $\pi\in S_n^{(0)}$ we have

$$Q_R(\pi;\pi^{-1}(k_1),\cdots,\pi^{-1}(k_\ell)) = Q_R(\pi^{-1};k_1,\cdots,k_\ell). \tag{137}$$

Hence, the left hand side of (135) may be rewritten as

$$\begin{aligned}
&\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}Q_R(\pi;\pi^{-1}(j))\\
&=\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}\mathcal{R}_{\pi^{-1}(1)+1}Q_R(\pi;\pi^{-1}(j),\pi^{-1}(1))\\
&=\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}\mathcal{R}_{\pi^{-1}(1)+1}Q_R(\pi^{-1};j,1).
\end{aligned} \tag{138}$$

Further, the bijection $\pi\mapsto\pi^{-1}$ yields that

$$\begin{aligned}
&\sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}\mathcal{R}_{\pi^{-1}(1)+1}Q_R(\pi^{-1};j,1)\\
&= \sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi^{-1})\mathcal{R}_{\pi(j)+1}\mathcal{R}_{\pi(1)+1}Q_R(\pi;j,1)\\
&= \sum_{\pi\in S_n^{(0)}}\operatorname{sgn}(\pi)\mathcal{R}_{\pi(j)+1}\mathcal{R}_{\pi(1)+1}Q_R(\pi;j,1).
\end{aligned} \tag{139}$$

Combining (138) and (139), we get that

$$\sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\mathcal{R}_{\pi^{-1}(j)+1}Q_R(\pi;\pi^{-1}(j)) = \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\mathcal{R}_{\pi(j)+1}\mathcal{R}_{\pi(1)+1}Q_R(\pi;j,1), \tag{140}$$

i.e., the left hand side of (135) is equal to the left hand side of (134) with j in place of i. As (134) holds, (135) holds too. Therefore, $\mathfrak{S}_3 = 0$. This concludes the proof that the coefficient of $t^{n(n+1)/2-1}$ in the numerator $\mathcal{R}_2 \det M_{\sqrt{t}R+N,n} - D_{R,n}(t)$ in the representation (97) of the PMMSE is equal to $\det M_{R,n}$.

C. Constant Terms

We have proven that the PMMSE satisfies

$$\operatorname{pmmse}_n(R|\sqrt{t}R+N) - \frac{1}{\binom{n+1}{2}}\frac{d}{dt}\log\det M_{\sqrt{t}R+N,n} = \frac{\mathcal{R}_2 \det M_{\sqrt{t}R+N,n} - D_{R,n}(t) - \frac{1}{\binom{n+1}{2}}\frac{d}{dt}\det M_{\sqrt{t}R+N,n}}{\det M_{\sqrt{t}R+N,n}} \tag{141}$$

where both the numerator and denominator of the right hand side are polynomials. We show next that, in the right hand side of (141), the constant term of the numerator is 0 and the constant term of the denominator is $G(n+2)$. Denote the constant terms of the numerator and denominator in the right hand side of (141) by $a_0^{(n)}(R)$ and $c^{(n)}$, respectively. Setting $t=0$ in $\det M_{\sqrt{t}R+N,n}$, we obtain that

$$c^{(n)} = \det M_{N,n}. \tag{142}$$

As in the statement of the theorem, for each $j \in \{1,\cdots,n(n+1)/2\}$, let $b_j^{(n)}(R)$ be the coefficient of t^j in $\det M_{\sqrt{t}R+N,n}$. Setting $t=0$ in (141), we obtain that

$$a_0^{(n)} = \mathcal{R}_2 c^{(n)} - D_{R,n}(0) - \frac{1}{\binom{n+1}{2}} b_1^{(n)}(R). \tag{143}$$

Furthermore, by definition of $D_{R,n}$ and $F_{R,n}$ (see equations (60)-(63)), we have that

$$D_{R,n}(0) = c^{(n)} F_{R,n}(0), \tag{144}$$

which simplifies into

$$D_{R,n}(0) = \mathcal{R}_1^2 c^{(n)}. \tag{145}$$

Indeed,

$$F_{R,n}(0) = \mathcal{R}_1^2 \mathbb{E}[\mathbf{N}^{(n)}]^T M_{N,n}^{-1}\mathbb{E}[\mathbf{N}^{(n)}] = \mathcal{R}_1^2, \tag{146}$$

where use the fact that $M_{N,n}^{-1}\mathbb{E}[\mathbf{N}^{(n)}] = (1,0,\cdots,0)^T$ since $\mathbb{E}[\mathbf{N}^{(n)}]$ is the first column of the invertible matrix $M_{N,n}$. Therefore,

$$a_0^{(n)} = \sigma(R)^2 c^{(n)} - \frac{1}{\binom{n+1}{2}} b_1^{(n)}(R). \tag{147}$$

We will utilize equation (147) to prove that $a_0^{(n)} = 0$, namely, we will show that

$$b_1^{(n)}(R) = \binom{n+1}{2}\sigma(R)^2 c^{(n)}. \tag{148}$$

We note that equation (148) would imply that $b_1^{(n)}$ is shift-invariant, i.e., $b_1^{(n)}(R+\alpha) = b_1^{(n)}(R)$ for any $\alpha \in \mathbb{R}$. In fact, we will show first that $b_1^{(n)}$ is shift-invariant, as this fact greatly simplifies the proof of (148).

For any $\mathbb{R}$-valued random variable such that $\mathbb{E}[Z^{2n}] < \infty$, considering for $j \in [n]$ i.i.d. random variables $Z_j \sim Z$ we have that

$$\det M_{Z,n} = \frac{1}{(n+1)!}\mathbb{E}\left[\prod_{0\le i<j\le n}(Z_i - Z_j)^2\right]. \tag{149}$$

From equation (149), we obtain that $\det M_{Z+\beta,n} = \det M_{Z,n}$ for any $\beta \in \mathbb{R}$. Hence, for any $\alpha \in \mathbb{R}$ we have that

$$\det M_{\sqrt{t}(R+\alpha)+N,n} = \det M_{\sqrt{t}R+N,n} \tag{150}$$

identically for every $t \in [0,\infty)$. As both sides of (150) are polynomials in t, we obtain that $b_j^{(n)}(R+\alpha) = b_j^{(n)}(R)$ for every $j \in \{1,\cdots,n(n+1)/2\}$.

Thus, we have that

$$b_1^{(n)}(R) = b_1^{(n)}(R-\mathcal{R}_1). \tag{151}$$

For any $\mathbb{R}$-valued random variable W, each entry in the matrix $M_{\sqrt{t}W+N,n}$ is a polynomial in $\sqrt{t}$; hence, the coefficient of t in $\det M_{\sqrt{t}W+N,n}$ is equal to the coefficient of t in $\det H_W$ (suppressing the dependence on N, n, and t) where

$$H_W := \left(\binom{i+j}{2}\mathcal{W}_2 t\mathcal{N}_{i+j-2} + (i+j)\mathcal{W}_1\sqrt{t}\mathcal{N}_{i+j-1} + \mathcal{N}_{i+j}\right)_{(i,j)\in[n]^2}$$

is the matrix obtained from $M_{\sqrt{t}W+N,n}$ by ignoring the terms of order $(\sqrt{t})^3$ and higher. Letting $W = R - \mathcal{R}_1$, we obtain that $b_1^{(n)}(R-\mathcal{R}_1)$ is equal to the coefficient of t in $\det H_{R-\mathcal{R}_1}$ where

$$H_{R-\mathcal{R}_1} = \left(\binom{i+j}{2}\sigma(R)^2 t\mathcal{N}_{i+j-2} + \mathcal{N}_{i+j}\right)_{(i,j)\in[n]^2}. \tag{152}$$

By Leibniz formula for the determinant, $b_1^{(n)}(R-\mathcal{R}_1)$ is equal to the coefficient of t in the polynomial

$$\sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\prod_{i\in[n]}\left(H_{R-\mathcal{R}_1}\right)_{i,\pi(i)}. \tag{153}$$

Then, we have that

$$b_1^{(n)}(R-\mathcal{R}_1) = \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi)\sum_{k\in[n]}\binom{k+\pi(k)}{2}\sigma(R)^2\mathcal{N}_{k+\pi(k)-2}\prod_{i\in[n]_k}\mathcal{N}_{i+\pi(i)}. \tag{154}$$

As we have that for any nonnegative integer m

$$\binom{m}{2}\mathcal{N}_{m-2} = \frac{m}{2}\mathcal{N}_m, \tag{155}$$

we may simplify (154) to

$$b_1^{(n)}(R-\mathcal{R}_1) = \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi) \sum_{k\in[n]} \frac{k+\pi(k)}{2}\sigma(R)^2 Q_N(\pi). \tag{156}$$

Evaluating the summation over k, we obtain that

$$b_1^{(n)}(R-\mathcal{R}_1) = \binom{n+1}{2}\sigma(R)^2 \sum_{\pi\in S_n^{(0)}} \operatorname{sgn}(\pi) Q_N(\pi). \tag{157}$$

From Leibniz formula for $\det M_{N,n}$, then,

$$b_1^{(n)}(R-\mathcal{R}_1) = \binom{n+1}{2}\sigma(R)^2 \det M_{N,n}. \tag{158}$$

Combining (142) and (142), we obtain (148). Hence, in view of formula (147), we obtain that $a_0^{(n)} = 0$, as desired.

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, $|\operatorname{supp}(\sqrt{t}R+N)| = \infty$ 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 $t \geq 0$

$$\lim_{n\to\infty} \operatorname{pmmse}_n(R|\sqrt{t}R+N) = \operatorname{mmse}(R|\sqrt{t}R+N). \tag{159}$$

Now, we show that the convergence is uniform.

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

$$\lim_{t\to\infty} \operatorname{pmmse}_n(R|\sqrt{t}R+N) = 0 \tag{160}$$

for every $n \in \mathbb{N}$ Further, by the convergence of the integral in the I-MMSE relation, we also know that

$$\lim_{t\to\infty} \operatorname{mmse}(R|\sqrt{t}R+N) = 0. \tag{161}$$

Hence, for each $n \in \mathbb{N}$,

$$\lim_{t\to\infty} \operatorname{pmmse}_n(R|\sqrt{t}Y+N) - \operatorname{mmse}(R|\sqrt{t}Y+N) = 0. \tag{162}$$

By definition of the PMMSE as the minimum over sets of increasing size (in n), the sequence $\{\operatorname{pmmse}_n(R|\sqrt{t}R+N)\}_{n\in\mathbb{N}}$ is decreasing.

Set, for each $n \in \mathbb{N}$ and $t \in [0,\infty)$,

$$g_n(t) := \operatorname{pmmse}_n(R|\sqrt{t}R+N) - \operatorname{mmse}(R|\sqrt{t}R+N) \tag{163}$$

for short. Note that the g_n are nonnegative. We have that $\{g_n\}_{n\in\mathbb{N}}$ is decreasing, and by (159) and (162)

$$\lim_{t\to\infty} g_n(t) = \lim_{n\to\infty} g_n(t) = 0. \tag{164}$$

We will show that

$$\lim_{n\to\infty} \sup_{t\in[0,\infty)} g_n(t) = 0, \tag{165}$$

which is the uniform convergence in (9).

Fix $\varepsilon > 0$. For each $n \in \mathbb{N}$, let $C_{\varepsilon,n} = g_n^{-1}([\varepsilon,\infty))$. 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, so each $C_{\varepsilon,n}$ is bounded. Hence, 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, 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 the uniform convergence in (165), as desired.

Appendix E
Proof of Theorem 4

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

$$I(X;Y+\gamma^{-1/2}N) = I(Y|\gamma) - \mathbb{E}_X\left[I(Y_X|\gamma)\right]. \tag{166}$$

Hence, by the I-MMSE formula,

$$I(X;Y+\gamma^{-1/2}N) = \frac{1}{2}\int_0^\gamma \operatorname{mmse}(Y|\sqrt{t}Y+N) - \mathbb{E}_X\left[\operatorname{mmse}(Y_X|\sqrt{t}Y_X+N)\right] dt. \tag{167}$$

Thus, we have the convergence

$$I(X;Y) = \lim_{\gamma\to\infty} I(X;Y+\gamma^{-1/2}N). \tag{168}$$

The uniform convergence of Theorem 2 applies to Y and each Y_x. Further, the expectation $\mathbb{E}_X$ is a finite positive linear combination. Thus, interchanging the order of integration over $[0,\gamma]$ and taking the limit as $n\to\infty$, we obtain that

$$I(X;Y+\gamma^{-1/2}N) = \lim_{n\to\infty} I_n(X;Y|\gamma). \tag{169}$$

Hence, we obtain the formula

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty} I_n(X;Y|\gamma), \tag{170}$$

as desired.

Appendix F
Proof of Theorem 5

We first treat convergence of the integral part. Let R be a continuous $\mathbb{R}$-valued random variable and $\{R_j \sim R\}_{j\in\mathbb{N}}$ an i.i.d. sequence, and assume that $(R,R_1,R_2,\cdots) \perp\!\!\!\perp N$. We show that the sequence $\{W_m \sim \operatorname{Unif}(\{R_j\}_{1\leq j\leq m})\}_{m\in\mathbb{N}}$ satisfies

$$\int_0^\infty \Theta_n(W_m;t)\,dt \to \int_0^\infty \Theta_n(R;t)\,dt \tag{171}$$

almost surely as $m\to\infty$. Further, we show that (171) holds if the integrals are computed instead over $[0,\gamma]$ for any $\gamma > 0$. We prove our claim via the continuous mapping theorem, then we show that applying the claim for Y and each Y_x in place of R finishes the proof of the theorem.

The starting point of the argument for proving (171) is that, by the strong law of large numbers, we have the almost sure convergence

$$\frac{\sum_{j=1}^{m} R_j^k}{m} \to \mathcal{R}_k \tag{172}$$

as $m \to \infty$ for every integer $1 \leq k \leq 2n$. We introduce the following notation. For each $m \in \mathbb{N}$, we consider the $\mathbb{R}^{2n}$-valued random vector $\boldsymbol{\mu}^{(m)}$ consisting of the first $2n$ moments of W_m

$$\boldsymbol{\mu}^{(m)} := \left(\mathbb{E}W_m^k\right)_{1\leq k\leq 2n}, \tag{173}$$

and let $\mu_k^{(m)}$ be the k-th coordinate of $\boldsymbol{\mu}^{(m)}$ for each $1 \leq k \leq 2n$, i.e,

$$\mu_k^{(m)} := \frac{\sum_{j=1}^{m} R_j^k}{m}. \tag{174}$$

So, (172) says that $\mu_k^{(m)} \to \mathcal{R}_k$ almost surely as $m \to \infty$ for each $1 \leq k \leq 2n$. We consider the constant vector

$$\boldsymbol{\nu} := \left(\mathbb{E}R^k\right)_{1\leq k\leq 2n}. \tag{175}$$

Then, $\boldsymbol{\mu}^{(m)} \to \boldsymbol{\nu}$ almost surely as $m \to \infty$. We continue the argument by showing that the integral we have is a successive composition of continuous functions.

The first continuous mapping comes as a result of the $a_j^{(n)}$ and $b_\ell^{(n)}$ being homogeneous polynomials in the moments. For each $j \in \{1, \cdots, n(n+1)/2 - 2\}$, let $\alpha_j : \mathbb{R}^{2n} \to \mathbb{R}$ be a polynomial such that

$$\alpha_j(\mathcal{Z}_1, \cdots, \mathcal{Z}_{2n}) = a_j^{(n)}(Z) \tag{176}$$

holds identically for every $\mathbb{R}$-valued random variable Z, whose MGF is finite everywhere, that satisfies and $Z \perp\!\!\!\perp N$. Similarly, we define, for each $\ell \in \{1, \cdots, n(n+1)/2\}$, a polynomial $\beta_\ell : \mathbb{R}^{2n} \to \mathbb{R}$ such that

$$\beta_\ell(\mathcal{Z}_1, \cdots, \mathcal{Z}_{2n}) = b_\ell^{(n)}(Z) \tag{177}$$

holds identically. Being polynomials, each function α_j and β_ℓ is continuous over $\mathbb{R}^{2n}$. Then, by the continuous mapping theorem,

$$\alpha_j(\boldsymbol{\mu}^{(m)}) \to \alpha_j(\boldsymbol{\nu}) \quad \text{and} \quad \beta_\ell(\boldsymbol{\mu}^{(m)}) \to \beta_\ell(\boldsymbol{\nu}) \tag{178}$$

almost surely as $m \to \infty$ for each $1 \leq j \leq n(n+1)/2 - 2$ and $1 \leq \ell \leq n(n+1)/2$. Denoting

$$\boldsymbol{\alpha}(\boldsymbol{\nu}) := (\alpha_j(\boldsymbol{\nu}))_{1\leq j\leq n(n+1)/2-2}, \tag{179}$$

$$\boldsymbol{\beta}(\boldsymbol{\nu}) := (\beta_\ell(\boldsymbol{\nu}))_{1\leq \ell\leq n(n+1)/2}, \tag{180}$$

we finish the proof that (171) holds by proving the continuity at the point

$$\begin{pmatrix} \boldsymbol{\alpha}(\boldsymbol{\nu}) \\ \boldsymbol{\beta}(\boldsymbol{\nu}) \end{pmatrix} \tag{181}$$

of the mapping $\mathbb{R}^{n(n+1)-2} \to \mathbb{R}$ defined by

$$\begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix} \mapsto \int_0^\infty \frac{\sum_{j=1}^{n(n+1)/2-2} p_j t^j}{c^{(n)} + \sum_{\ell=1}^{n(n+1)/2} q_j t^j}\, dt \tag{182}$$

where p_j and q_ℓ are the j-th and ℓ-th coordinates of $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively, in (182). For this continuity argument, we shall show first that the mapping in (182) is well-defined on an open neighborhood of the point of interest (181).

We consider the subset $H \subset \mathbb{R}^{\binom{n+1}{2}}$ defined by

$$H := \left\{ \boldsymbol{d} \in \mathbb{R}^{\binom{n+1}{2}} \;\middle|\; d_{\binom{n+1}{2}} > 0,\ \sum_{\ell=1}^{\binom{n+1}{2}} d_\ell t^\ell > -c^{(n)}\ \forall t \geq 0 \right\} \tag{183}$$

where in this definition and the subsequent argument we set

$$\boldsymbol{d} = \left(d_1, \cdots, d_{\binom{n+1}{2}}\right)^T. \tag{184}$$

We show that H is open and that it contains $\boldsymbol{\beta}(\boldsymbol{\nu})$, then pick a small open neighborhood of $\boldsymbol{\beta}(\boldsymbol{\nu})$ that is contained within H.

Fix $\boldsymbol{d} \in H$ and let

$$\varepsilon_1 \in \left(0, d_{\binom{n+1}{2}}\right). \tag{185}$$

We have that the polynomial

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon_1) t^\ell \tag{186}$$

is eventually increasing and approaches infinity as $t \to \infty$. Let $t_0 > 1$ be any real such that

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon_1) t^\ell > -c^{(n)} \tag{187}$$

for every $t > t_0$. Being continuous, the polynomial

$$\sum_{\ell=1}^{\binom{n+1}{2}} d_\ell t^\ell \tag{188}$$

attains its minimum over the compact set $[0, t_0]$. Let s be this minimum. Let

$$\varepsilon = \frac{1}{2} \min\left(\varepsilon_1, \frac{(s + c^{(n)})(t_0 - 1)}{t_0(t_0^{\binom{n+1}{2}} - 1)}\right). \tag{189}$$

As $\varepsilon < \varepsilon_1$, inequality (187) yields that

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon) t^\ell > -c^{(n)} \tag{190}$$

for every $t > t_0$. Furthermore, for any $t \in [0, t_0]$,

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon) t^\ell = \sum_{\ell=1}^{\binom{n+1}{2}} d_\ell t^\ell - \varepsilon \sum_{\ell=1}^{\binom{n+1}{2}} t^\ell \tag{191}$$

$$\geq s - \varepsilon \sum_{\ell=1}^{\binom{n+1}{2}} t_0^\ell \tag{192}$$

$$> s - \frac{(s + c^{(n)})(t_0 - 1)}{t_0(t_0^{\binom{n+1}{2}} - 1)} \sum_{\ell=1}^{\binom{n+1}{2}} t_0^\ell \tag{193}$$

$$= s - (s + c^{(n)}) = -c^{(n)}. \tag{194}$$

Thus,

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon) t^\ell > -c^{(n)} \tag{195}$$

for every $t \in [0,\infty)$. Hence, for any $(\delta_\ell)_{1\le\ell\le n(n+1)/2} =: \boldsymbol{\delta} \in \mathbb{R}^{\binom{n+1}{2}}$ such that $\|\boldsymbol{\delta}\| < \varepsilon$, we have that

$$\sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \delta_\ell) t^\ell \ge \sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \|\boldsymbol{\delta}\|) t^\ell \ge \sum_{\ell=1}^{\binom{n+1}{2}} (d_\ell - \varepsilon) t^\ell > -c^{(n)} \tag{196}$$

for all $t \in [0,\infty)$. This completes the proof that H is open.

Continuity of R and the fact that $\det M_{\sqrt{t}R+N} > 0$ for every $t \in [0,\infty)$ imply that

$$\boldsymbol{\beta}(\boldsymbol{\nu}) \in H. \tag{197}$$

By openness of H, there is an $\eta_1 > 0$ such that

$$\prod_{\ell=1}^{2n} (\beta_\ell(\boldsymbol{\nu}) - \eta_1, \beta_\ell(\boldsymbol{\nu}) + \eta_1) \subset H \tag{198}$$

is an open set containing $\boldsymbol{\beta}(\boldsymbol{\nu})$. We necessary have $\eta_1 \le \beta_{n(n+1)/2}(\boldsymbol{\nu})$, for otherwise the element $((\beta_{n(n+1)/2}(\boldsymbol{\nu}) - \eta_1)/2, 0, \cdots, 0)$ of H would produce the polynomial $(\beta_{n(n+1)/2}(\boldsymbol{\nu}) - \eta_1) t^{n(n+1)/2}/2$ which would approach $-\infty$ as t grows without bound contradicting the defining condition of H. We set

$$\eta := \frac{\eta_1}{2}. \tag{199}$$

The strict inequality $\eta < \beta_{n(n+1)/2}(\boldsymbol{\nu})$ holds. Further, setting

$$\mathcal{O} := \prod_{\ell=1}^{2n} (\beta_\ell(\boldsymbol{\nu}) - \eta, \beta_\ell(\boldsymbol{\nu}) + \eta), \tag{200}$$

we get that for any $\boldsymbol{q} \in \mathcal{O}$

$$\left| c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} q_\ell t^\ell \right| \ge \det M_{\sqrt{t}R+N,n} - \eta \sum_{\ell=1}^{\binom{n+1}{2}} t^\ell > 0 \tag{201}$$

for every $t \in [0,\infty)$.

Thus, the function $g : \mathbb{R}^{\binom{n+1}{2}-2} \times \mathcal{O} \to \mathbb{R}$ given by

$$g(\boldsymbol{p}, \boldsymbol{q}) = \int_0^\infty \frac{\sum_{j=1}^{\binom{n+1}{2}-2} p_j t^j}{c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} q_\ell t^\ell}\, dt \tag{202}$$

is well defined, as the integrand is integrable. From (201), Lebesgue's dominated convergence shows continuity of g at $(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu}))$, as follows.

Let $\boldsymbol{w} := (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{\binom{n+1}{2}-2} \times \mathcal{O}$ be such that $\|\boldsymbol{w}\| < \eta$. The integrand of g at $(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})) - (\boldsymbol{u}, \boldsymbol{v})$ may be bounded as

$$\begin{aligned} &\left| \frac{\sum_{j=1}^{\binom{n+1}{2}-2} (\alpha_j(\boldsymbol{\nu}) - u_j) t^j}{c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} (\beta_\ell(\boldsymbol{\nu}) - v_\ell) t^\ell} \right| \\ &\le \frac{\sum_{j=1}^{\binom{n+1}{2}-2} (|\alpha_j(\boldsymbol{\nu})| + \eta) t^j}{c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} \beta_\ell(\boldsymbol{\nu}) t^\ell - \eta \sum_{\ell=1}^{\binom{n+1}{2}} t^\ell} \end{aligned} \tag{203}$$

which is integrable over $[0,\infty)$ as the denominator's degree exceed that of the numerator by 2. Hence, by Lebesgue's dominated convergence

$$\lim_{\|\boldsymbol{w}\|\to 0} g((\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})) - \boldsymbol{w}) = g(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})), \tag{204}$$

i.e., g is continuous at $(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu}))$, as desired.

Denoting, for each $m \in \mathbb{N}$,

$$\boldsymbol{\alpha}^{(m)} := \left(\alpha_j(\boldsymbol{\mu}^{(m)})\right)_{1\le j\le\binom{n+1}{2}-2}, \tag{205}$$

$$\boldsymbol{\beta}^{(m)} := \left(\beta_\ell(\boldsymbol{\mu}^{(m)})\right)_{1\le \ell\le\binom{n+1}{2}}, \tag{206}$$

we see that

$$g(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\beta}^{(m)}) = \int_0^\infty \Theta_n(W_m; t)\, dt \tag{207}$$

and

$$g(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})) = \int_0^\infty \Theta_n(R; t)\, dt. \tag{208}$$

Further,

$$(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\beta}^{(m)}) \to (\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})) \tag{209}$$

almost surely as $m \to \infty$. Hence, continuity of g implies that

$$g(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\beta}^{(m)}) \to g(\boldsymbol{\alpha}(\boldsymbol{\nu}), \boldsymbol{\beta}(\boldsymbol{\nu})) \tag{210}$$

almost surely as $m \to \infty$, i.e., (171) holds. Repeating this proof for $[0,\gamma]$ replacing $[0,\infty)$ as the integration region yields that

$$\int_0^\gamma \Theta_n(W_m; t)\, dt \to \int_0^\gamma \Theta_n(R; t)\, dt \tag{211}$$

almost surely as $m \to \infty$ for any fixed $\gamma > 0$.

We have the almost sure convergence

$$\beta_{n(n+1)/2}(\boldsymbol{\mu}^{(m)}) \to \beta_{n(n+1)/2}(\boldsymbol{\nu}) \tag{212}$$

as $m \to \infty$. As the mapping $\mathbb{R}_{>0} \to \mathbb{R}$ defined by

$$q \mapsto \log q \tag{213}$$

is continuous, the continuous mapping theorem yields that

$$\log \beta_{n(n+1)/2}(\boldsymbol{\mu}^{(m)}) \to \log \beta_{n(n+1)/2}(\boldsymbol{\nu}) \tag{214}$$

almost surely as $m \to \infty$. In other words,

$$\log\det M_{W_m,n} \to \log\det M_{R,n} \tag{215}$$

as almost surely as $m \to \infty$.

Fix $\gamma > 0$. Since we have that the almost sure convergence

$$\boldsymbol{\beta}^{(m)} \to \boldsymbol{\beta}(\boldsymbol{\nu}) \tag{216}$$

as $m \to \infty$, and since the mapping $\mathbb{R}^{\binom{n+1}{2}} \to \mathbb{R}$ defined by

$$\boldsymbol{q} \mapsto \log\left(c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} q_\ell \gamma^\ell\right) \tag{217}$$

is continuous, the continuous mapping theorem implies that

$$\log\left(c^{(n)}+\sum_{\ell=1}^{\binom{n+1}{2}} b_\ell^{(n)}(W_m)\gamma^\ell\right) \to \log\left(c^{(n)}+\sum_{\ell=1}^{\binom{n+1}{2}} b_\ell^{(n)}(R)\gamma^\ell\right) \tag{218}$$

almost surely as $m \to \infty$. In other words,

$$\log\det M_{\sqrt{\gamma}W_m+N} \to \log\det M_{\sqrt{\gamma}R+N} \tag{219}$$

almost surely as $m \to \infty$.

For each $m \in \mathbb{N}$, let

$$\mathcal{S}^{(m)} = \{(X_j, Y_j)\}_{1\le j\le m} \tag{220}$$

denote the first m samples. For each $x \in \mathcal{X}$, set

$$\mathcal{T}_x^{(m)} := \{j \in [m] \,;\, X_j = x\}. \tag{221}$$

We have the partition

$$[m] = \bigcup_{x\in\mathcal{X}} \mathcal{T}_x^{(m)}. \tag{222}$$

Let $\mathcal{S}_n^{(m)}$ be as defined in (35), i.e.,

$$\mathcal{S}_n^{(m)} = \{(X_j, Y_j) \in \mathcal{S}^{(m)} \,;\, |\{1 \le i \le m \,;\, X_i = X_j\}| > n\}. \tag{223}$$

Then, we have the equality of events

$$\{\mathcal{S}_n^{(m)} \neq \emptyset\} = \bigcup_{x\in\mathcal{X}} \{|\mathcal{T}_x^{(m)}| > n\}. \tag{224}$$

We will, in fact, require that the stricter event

$$\bigcap_{x\in\mathcal{X}} \{|\mathcal{T}_x^{(m)}| > n\} \tag{225}$$

occurs. As $\mathcal{X}$ is finite,

$$\lim_{m\to\infty} \Pr\left(\bigcap_{x\in\mathcal{X}} \{|\mathcal{T}_x^{(m)}| > n\}\right) = 1 \tag{226}$$

if each

$$\lim_{m\to\infty} \Pr\left(\{|\mathcal{T}_x^{(m)}| > n\}\right) = 1 \tag{227}$$

holds individually. Further, each (227) holds, since we are assuming that $\Pr(X = x) > 0$ and Y_x is continuous.

Finally, we note that the empirical measure of X converges almost surely to P_X. Thus, replacing R by Y and each Y_x, we obtain that

$$\widehat{I_n}\left(\{(X_j, Y_j)\}_{1\le j\le m}|\gamma\right) \to I_n(X;Y|\gamma) \tag{228}$$

for every $\gamma > 0$ and

$$\widehat{I_n}\left(\{(X_j, Y_j)\}_{1\le j\le m}\right) \to I_n(X;Y) \tag{229}$$

both almost surely as $m \to \infty$. As the MGF of Y is finite everywhere, we also have, by Theorem 4,

$$I(X;Y) = \lim_{\gamma\to\infty}\lim_{n\to\infty}\lim_{m\to\infty} \widehat{I_n}\left(\{(X_j, Y_j)\}_{1\le j\le m}|\gamma\right), \tag{230}$$

where the convergence in m is almost sure convergence.

Appendix G
Number of Samples

The following proposition shows that for a fixed $n \in \mathbb{N}$, the number of samples required for our estimate of $I_n(X;Y)$ to be within ε of the true value with probability greater than $1-\delta$ is of order at most $-\varepsilon^{-2}\log\delta$.

Proposition 2. *Assume that Y is of bounded support, and let $n \in \mathbb{N}$. Fix a sequence of i.i.d. samples*

$$\{(X_j, Y_j) \sim (X, Y)\}_{j\in\mathbb{N}}, \tag{231}$$

and for each $m \in \mathbb{N}$ set

$$\mathcal{S}^{(m)} = \{(X_j, Y_j)\}_{1\le j\le m}. \tag{232}$$

There are constants $\delta_{X,Y,n}, \varepsilon_{X,Y,n}, C_{X,Y,n} > 0$ such that for every $\delta \in (0, \delta_{X,Y,n})$ and $\varepsilon \in (0, \varepsilon_{X,Y,n})$ we have that

$$m > \frac{C_{X,Y,n}}{\varepsilon^2}\log\frac{1}{\delta} \tag{233}$$

implies

$$\Pr\left\{\left|\widehat{I_n}\left(\mathcal{S}^{(m)}\right) - I_n(X;Y)\right| < \varepsilon\right\} \ge 1-\delta. \tag{234}$$

The proof of Proposition 2 is divided into several parts throughout this appendix. In the course of the proof, we will give explicit formulas for the constants $\delta_{X,Y,n}, \varepsilon_{X,Y,n}, C_{X,Y,n}$.

First, we give a sufficient condition on m making each $x \in \mathcal{X}$ occur in the samples $\mathcal{S}^{(m)}$ more than n times. This step is necessary for the determinants in the estimator to be strictly positive. Then, we set-up the notation and the overarching idea of the proof in Appendix G-A. Finally, the separate parts of the proof are distributed across subsequent subsections.

As in Appendix F, for each $x \in \mathcal{X}$ we define

$$\mathcal{T}_x^{(m)} := \{j \in [m] \,;\, X_j = x\}, \tag{235}$$

so we have the partition

$$[m] = \bigcup_{x\in\mathcal{X}} \mathcal{T}_x^{(m)}. \tag{236}$$

By continuity of Y_x, the number of distinct elements in

$$\{(X_j, Y_j) \,;\, j \in \mathcal{T}_x^{(m)}\} \tag{237}$$

is equal to $|\mathcal{T}_x^{(m)}|$ with probability 1. Indeed, the event

$$O(\mathcal{S}^{(m)}) := \bigcap_{1\le i<j\le m} \{(X_i, Y_i) \neq (X_j, Y_j)\} \tag{238}$$

occurs with probability 1. Further, equation (226) shows that the $\mathcal{T}_x^{(m)}$ simultaneously have sizes greater than n with probability 1 as $m \to \infty$. We quantify this convergence next.

Consider the empirical measure defined by

$$\widehat{P_X}^{(m)}(x) = \frac{|\{j \in [m] \,;\, X_j = x\}|}{m} \tag{239}$$

for each $x \in \mathcal{X}$. Let δ_x denoting the Dirac measure, i.e., for $G \subset \mathbb{R}$

$$\delta_x(G) = \begin{cases} 1 & \text{if } x \in G, \\ 0 & \text{otherwise.} \end{cases} \tag{240}$$

Then we may write

$$\widehat{P_X}^{(m)}(x) = \frac{\sum_{j=1}^m \delta_x(\{X_j\})}{m}. \tag{241}$$

We have that for each $x \in \mathcal{X}$

$$\mathbb{E}\left[\delta_x(\{X\})\right] = P_X(x). \tag{242}$$

Thus, by Hoeffding's inequality we have that for any $\zeta > 0$

$$\Pr\left\{P_X(x) - \widehat{P_X}^{(m)}(x) \geq \zeta\right\} \leq e^{-2m\zeta^2}. \tag{243}$$

But, for each $x \in \mathcal{X}$,

$$m\widehat{P_X}^{(m)}(x) = |\mathcal{T}_x^{(m)}|. \tag{244}$$

Then, for every $\zeta > 0$ and $x \in \mathcal{X}$,

$$\Pr\left\{m(P_X(x) - \zeta) \geq |\mathcal{T}_x^{(m)}|\right\} \leq e^{-2m\zeta^2}. \tag{245}$$

Let $x_0 \in \mathcal{X}$ be a minimizer of $P_X(x)$, i.e., x_0 is any element in $\mathcal{X}$ satisfying

$$P_X(x_0) = \min_{x \in \mathcal{X}} P_X(x). \tag{246}$$

We have that, for every $\zeta > 0$ and $x \in \mathcal{X}$

$$\Pr\left\{m(P_X(x_0) - \zeta) \geq |\mathcal{T}_x^{(m)}|\right\} \leq e^{-2m\zeta^2}. \tag{247}$$

By the union bound,

$$\Pr\left\{m(P_X(x_0) - \zeta) \geq \min_{x \in \mathcal{X}} |\mathcal{T}_x^{(m)}|\right\} \leq |\mathcal{X}| e^{-2m\zeta^2}. \tag{248}$$

Let $\zeta = P_X(x_0)/2$. Then,

$$\Pr\left\{mP_X(x_0)/2 \geq \min_{x \in \mathcal{X}} |\mathcal{T}_x^{(m)}|\right\} \leq |\mathcal{X}| e^{-mP_X(x_0)^2/2}. \tag{249}$$

Then, for each $\delta > 0$ and $\ell > 0$, whenever

$$m > \max\left(\frac{2\ell}{P_X(x_0)}, \frac{2}{P_X(x_0)^2} \log \frac{3|\mathcal{X}|}{\delta}\right) \tag{250}$$

we will also have that

$$\Pr\left\{\min_{x \in \mathcal{X}} |\mathcal{T}_x^{(m)}| > \ell\right\} > 1 - \frac{\delta}{3}. \tag{251}$$

We will let

$$D_\ell(\mathcal{S}^{(m)}) := \left\{\min_{x \in \mathcal{X}} |\mathcal{T}_x^{(m)}| > \ell\right\} \cap O(\mathcal{S}^{(m)}) \tag{252}$$

denote the event that every $x \in \mathcal{X}$ occurs in the samples $\mathcal{S}^{(m)}$ in more than ℓ distinct pairs. We will requires that $D_n(\mathcal{S}^{(m)})$ occurs so that determinants in the estimator are nonzero, and we will also requires that $D_\ell(\mathcal{S}^{(m)})$ for other values of ℓ occur so that the estimated moments are accurate.

A. Set-up

We prove general results for a continuous $\mathbb{R}$-valued random variable R of finite support $\mathrm{supp}(R) \subset [p, q] \subset (0, \infty)$ that satisfies $R \perp\!\!\!\perp N$. For any sequence of i.i.d. random variables $\{R_j \sim R\}_{j \in \mathbb{N}}$ such that $(R_1, R_2, \cdots) \perp\!\!\!\perp N$, and any $m > n$, we consider the random variables

$$\Delta_{R,n}(R_1, \cdots, R_m) := \frac{1}{n(n+1)} \log \frac{\det M_{W_m,n}}{\det M_{R,n}} + \frac{1}{2}\left(\int_0^\infty \Theta_n(W_m; t)\, dt - \int_0^\infty \Theta_n(R; t)\, dt\right) \tag{253}$$

where we define a sequence

$$\{W_m \sim \mathrm{Unif}(\{R_j\}_{1 \leq j \leq m})\}_{m > n}. \tag{254}$$

Obtaining bounds on $\Delta_{R,n}$ and quantifying the convergence of the empirical measure of the samples of X to P_X together will suffice to prove Proposition 2. Indeed, we may write

$$\begin{aligned}
&\widehat{I_n}(\mathcal{S}^{(m)}) - I_n(X;Y) \\
&= \Delta_{Y,n}(Y_1, \cdots, Y_m) - \sum_{x \in \mathcal{X}} \widehat{P_X}^{(m)}(x) \Delta_{Y_x,n}\left((Y_j)_{j \in \mathcal{T}_x^{(m)}}\right) \\
&+ \sum_{x \in \mathcal{X}} \left(\frac{1}{n(n+1)} \log\det M_{Y_x,n} + \frac{1}{2}\int_0^\infty \Theta_n(Y_x; t)\, dt\right) \cdot \left(P_X(x) - \widehat{P_X}^{(m)}(x)\right).
\end{aligned} \tag{255}$$

Via bounds on $\Delta_{R,n}$ applied to Y and Y_x (for each $x \in \mathcal{X}$) in place of R, along with an upper bound on the deviation of the empirical measure of X from the true measure, we will deduce an upper bound on the deviation of $\widehat{I_n}(\mathcal{S}^{(m)})$ from $I_n(X;Y)$. When treating $\Delta_{R,n}$, we will consider the determinant part and the integral part separately. Nevertheless, the proof technique is the same. Let α_j and β_ℓ be the polynomials in equations (176) and (177) in Appendix F. The main technique is to split each of these polynomials into a positive part and a negative part, as follows.

Let $\alpha_j^{(+)}, \alpha_j^{(-)}, \beta_\ell^{(+)}, \beta_\ell^{(-)}$ be polynomials in $2n$ variables of minimal total degrees and with all strictly positive coefficients such that

$$\alpha_j(\xi_1, \cdots, \xi_{2n}) = \alpha_j^{(+)}(\xi_1, \cdots, \xi_{2n}) - \alpha_j^{(-)}(\xi_1, \cdots, \xi_{2n}) \tag{256}$$

$$\beta_\ell(\xi_1, \cdots, \xi_{2n}) = \beta_\ell^{(+)}(\xi_1, \cdots, \xi_{2n}) - \beta_\ell^{(-)}(\xi_1, \cdots, \xi_{2n}) \tag{257}$$

all hold identically over $(\xi_1, \cdots, \xi_{2n}) \in \mathbb{R}^{2n}$. By positivity of R, each moment $\mathbb{E}R^k$ is strictly positive. Then, we may write

$$\Theta_n(R; t) = \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} \tag{258}$$

with the polynomials in t

$$f_R(t) := \sum_{j=1}^{\binom{n+1}{2}-2} \alpha_j^{(+)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) t^j \tag{259}$$

$$g_R(t) := \sum_{j=1}^{\binom{n+1}{2}-2} \alpha_j^{(-)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) t^j \tag{260}$$

$$u_R(t) := c^{(n)} + \sum_{\ell=1}^{\binom{n+1}{2}} \beta_\ell^{(+)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) t^\ell \tag{261}$$

$$v_R(t) := \sum_{\ell=1}^{\binom{n+1}{2}} \beta_\ell^{(-)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) t^\ell, \tag{262}$$

having all nonnegative coefficients, where we keep the notation $\mathcal{R}_k$ from (65) in Appendix C, i.e.,

$$\mathcal{R}_k := \mathbb{E}R^k \tag{263}$$

for $k \in [2n]$. We note that we have suppressed the dependence on n in the notation used for these polynomials.

For $h \in \{f, g, u, v\}$, let h_{W_m} be the random variable whose value is what is obtained via h_R when the moments of R are replaced with the sample moments obtained from the samples $\{R_i\}_{1 \le i \le m}$, e.g.,

$$f_{W_m}(t) := \sum_{j=1}^{\binom{n+1}{2}-2} \alpha_j^{(+)} \left(\frac{\sum_{i=1}^m R_i}{m}, \cdots, \frac{\sum_{i=1}^m R_i^{2n}}{m} \right) t^j. \tag{264}$$

As $W \perp\!\!\!\perp N$,

$$u_{W_m}(t) - v_{W_m}(t) = \det M_{\sqrt{t}W_m + N, n} > 0. \tag{265}$$

Then the function

$$\Theta_n(W_m; t) = \frac{f_{W_m}(t) - g_{W_m}(t)}{u_{W_m}(t) - v_{W_m}(t)} \tag{266}$$

is well-defined over $t \in [0, \infty)$.

Hoeffding's inequality yields that, for any $z > 0$

$$\Pr\left\{ \left| \mathbb{E}R - \frac{1}{m} \sum_{i=1}^m R_i \right| \ge z \right\} \le 2e^{-2mz^2/(q-p)^2}. \tag{267}$$

Setting $z = \eta \mathbb{E}R \ge \eta p > 0$ for $\eta \in (0, 1)$ yields that

$$\Pr\left\{ (1-\eta)\mathbb{E}R < \frac{1}{m} \sum_{i=1}^m R_i < (1+\eta)\mathbb{E}R \right\} \ge 1 - 2e^{-2m\eta^2/(q/p-1)^2}.$$

Further, the union bound yields that

$$\Pr\left\{ (1-\eta)\mathbb{E}R^k < \frac{1}{m} \sum_{i=1}^m R_i^k < (1+\eta)\mathbb{E}R^k, \ \forall\, k \in [2n] \right\} \ge 1 - 4ne^{-2m\eta^2/\left((q/p)^{2n}-1\right)^2}. \tag{268}$$

We would like to be able use inequality (268) for Y and each Y_x in place of R. Assuming boundedness of Y ensures that each Y_x is bounded. Further, shift-invariance of I_n yields that we may assume that Y (and each Y_x) is supported inside some interval $[p, q] \subset (0, \infty)$. For each $\eta \in (0, 1)$, we denote the event

$$E_{n,\eta}(\{R_i\}_{i\in[m]}) := \left\{ 1 - \eta \le \frac{\sum_{i=1}^m R_i^k}{m\mathcal{R}^k} \le 1 + \eta, \ \forall\, k \in [2n] \right\}, \tag{269}$$

so (268) says that

$$\Pr\left(E_{n,\eta}(\{R_i\}_{i\in[m]})\right) \ge 1 - 4ne^{-2m\eta^2/\left((q/p)^{2n}-1\right)^2}. \tag{270}$$

We will show that there is a threshold on η depending only R and n below which the event $E_{n,\eta}(\{R_i\}_{i\in[m]})$ implies the bounds

$$\int_0^\infty \frac{f_R(t) - \nu g_R(\nu t)}{u_R(\nu t) - v_R(t)}\, dt \le \int_0^\infty \Theta_n(W_m; t)\, dt \tag{271}$$

$$\le \int_0^\infty \frac{\nu f_R(\nu t) - g_R(t)}{u_R(t) - v_R(\nu t)}\, dt \tag{272}$$

where we denote

$$\nu := \left(\frac{1+\eta}{1-\eta} \right)^2. \tag{273}$$

For the upper bound (272), we show that the denominator $u_R(t) - v_R(\nu t)$ of the integrand is strictly positive for all $t \in [0, \infty)$. We further derive refined bounds by uniformly bounding the two functions

$$\varphi_R(t; \nu) := \frac{u_R(t) - v_R(t)}{u_R(t) - v_R(\nu t)}, \tag{274}$$

$$\psi_R(t; \nu) := \frac{u_R(t) - v_R(t)}{u_R(\nu t) - v_R(t)}. \tag{275}$$

Specifically, we show that as $s \searrow 0$ there are thresholds $\nu_{R,n,s} > 1$ such that for any fixed $\nu \le \nu_{R,n,s}$ we have the uniform bounds

$$1 - s \le \psi_R(t; \nu) \le 1 \le \varphi_R(t; \nu) \le 1 + s \tag{276}$$

for every $t \in [0, \infty)$. Then, we will deduce the following bound on the deviation of the integral part of the estimator

$$\left| \int_0^\infty \Theta_n(W_m; t)\, dt - \int_0^\infty \Theta_n(R; t)\, dt \right| \le \left((1+s)\nu^{\binom{n+1}{2}-1} - 1 \right) \int_0^\infty \frac{f_R(t) + g_R(t)}{u_R(t) - v_R(t)}\, dt. \tag{277}$$

The upper bound in (277) may be made as small as needed by choosing a small s then choosing a small ν.

Uniform bounds on φ_R and ψ_R are given in Appendix G-B, then a proof of inequality (277) is derived in Appendix G-C. To finish the proof, we will derive error bounds on estimating $\det M_{R,n}$ and P_X from samples. In Appendix G-D we obtain the bound

$$\left| \frac{1}{n(n+1)} \log \frac{\det M_{W_m,n}}{\det M_{R,n}} \right| \le \frac{6\left(\beta_{R,n}^{(+)} + \beta_{R,n}^{(-)}\right)\eta}{\left(\beta_{R,n}^{(+)} - \beta_{R,n}^{(-)}\right) n} \tag{278}$$

where we write

$$\beta_{R,n}^{(+)} := \beta_{\binom{n+1}{2}}^{(+)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{279}$$

$$\beta_{R,n}^{(-)} := \beta_{\binom{n+1}{2}}^{(-)}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{280}$$

for short. Then in Appendix G-E we bound the error in estimating P_X.

B. Uniform Bounds on φ_R and ψ_R: Inequalities (276)

We show that the following bounds on φ_R and ψ_R hold.

Lemma 5. *Let R be a nonnegative continuous $\mathbb{R}$-valued random variable whose MGF is finite everywhere, and assume that $R \perp\!\!\!\perp N$. Define*

$$\mu_R := \sup_{t\in[0,\infty)} \frac{v_R(t)}{u_R(t)}. \tag{281}$$

Then $\mu_R < 1$. Furthermore, for every

$$s \in \left(0, \frac{1-\mu_R}{1+\mu_R}\right), \tag{282}$$

if

$$1 \le \nu \le \left(1 + \frac{s(1-\mu_R)}{1-s}\right)^{2/(n(n+1))} \tag{283}$$

then we have the uniform bounds

$$1 - s \le \psi_R(t;\nu) \le 1 \le \varphi_R(t;\nu) \le 1 + s \tag{284}$$

for all $t \in [0,\infty)$.

We deduce Lemma 5 from two separate results for slightly larger classes of functions than φ_R and ψ_R. The following lemma shows that φ_R can be made uniformly close to 1, and it also provides a threshold on η below which the denominator $u_R(t) - v_R(\nu t)$ is strictly positive for all $t \in [0,\infty)$.

Lemma 6. *Let $u : [0,\infty) \to (0,\infty)$ and $v : [0,\infty) \to [0,\infty)$ be two nondecreasing continuous functions such that v does not identically vanish. Suppose that*

- *$u(t) > v(t)$ for every $t \in [0,\infty)$,*
- *$\limsup_{t\to\infty} v(t)/u(t) < 1$, and*
- *there is a $k > 0$ such that $v(xy) \le x^k v(y)$ for every $x \in [1,\infty)$ and $y \in [0,\infty)$,*

and define

$$\mu := \sup_{t\in[0,\infty)} \frac{v(t)}{u(t)}. \tag{285}$$

For every $s > 0$, if

$$1 \le \nu \le \left(1 + \frac{s(1-\mu)}{\mu(1+s)}\right)^{1/k}. \tag{286}$$

then we have the uniform bound

$$1 \le \frac{u(t)-v(t)}{u(t)-v(\nu t)} \le 1 + s \tag{287}$$

for all $t \in [0,\infty)$.

Proof. First, we note that monotonicity of v implies that for any $\nu \ge 1$ and $t \ge 0$

$$1 \le \frac{u(t)-v(t)}{u(t)-v(\nu t)} \tag{288}$$

whenever the denominator $u(t) - v(\nu t)$ is strictly positive. We show that $\mu \in (0,1)$, and that this fact implies that the denominator in (287) is strictly positive.

That v does not vanish identically yields that $\mu > 0$. On the other hand, as

$$\xi := \limsup_{t\to\infty} \frac{v(t)}{u(t)} < 1, \tag{289}$$

we get that there is a $t_0 \ge 0$ such that

$$\frac{v(t)}{u(t)} < \frac{1+\xi}{2} < 1 \tag{290}$$

for every $t > t_0$. Further, by continuity of $v(t)/u(t)$, the extreme value theorem yields that $v(t)/u(t)$ attains its maximum over $[0,t_0]$, which is necessarily strictly less than 1 because $v(t) < u(t)$ for all $t \ge 0$. Hence, $\mu < 1$. Then, with

$$\nu_0 := \frac{1}{\mu^{1/k}} \tag{291}$$

we have for any $\nu \in [1,\nu_0)$ the uniform bound

$$\frac{v(\nu t)}{u(t)} \le \frac{\nu^k v(t)}{u(t)} \le \nu^k \mu < \nu_0^k \mu = 1 \tag{292}$$

for every $t \in [0,\infty)$.

What we have proven thus far implies that the ratio in the middle of (287) is bounded. Indeed, for any $\nu \in [1,\nu_0)$,

$$1 \le \frac{u(t)-v(t)}{u(t)-v(\nu t)} \le \frac{1}{1-\nu^k\mu} \tag{293}$$

uniformly for $t \in [0,\infty)$. We will show that uniform bounds arbitrarily close to 1 are attainable when the condition in (286) is satisfied; we note that the upper bound in (293) cannot go below $1+s$ if $s < \mu/(1-\mu)$.

So, fix $s > 0$, and assume that (286) holds, i.e., assume that ν is a real number satisfying

$$1 \le \nu \le \left(1 + \frac{s(1-\mu)}{\mu(1+s)}\right)^{1/k}. \tag{294}$$

Then, for every $t \in [0,\infty)$, that $\mu \ge v(t)/u(t)$ yields

$$v(\nu t) \le \nu^k v(t) \tag{295}$$

$$\le \left(1 + \frac{s(1-\mu)}{\mu(1+s)}\right) v(t) \tag{296}$$

$$\le v(t) + \frac{s(u(t)-v(t))}{1+s}. \tag{297}$$

Rearranging (297), we obtain that

$$\frac{-s}{1+s} \le \frac{v(t)-v(\nu t)}{u(t)-v(t)} \tag{298}$$

for every $t \in [0,\infty)$. Since $\mu \in (0,1)$ and $s > 0$,

$$\left(1 + \frac{s(1-\mu)}{\mu(1+s)}\right)^{1/k} < \nu_0. \tag{299}$$

Thus, by assumption (286), $\nu < \nu_0$. Hence, (292) yields that

$$u(t) > v(\nu t) \tag{300}$$

for every $t \in [0, \infty)$. Inequality (300) allows us to conclude from inequality (298), upon adding 1 to both sides then inverting, that

$$\frac{u(t) - v(t)}{u(t) - v(\nu t)} \leq 1 + s \tag{301}$$

for every $t \in [0, \infty)$. Inequality (300) further implies, by (288) that

$$1 \leq \frac{u(t) - v(t)}{u(t) - v(\nu t)} \tag{302}$$

uniformly for $t \in [0, \infty)$. Combining (301) and (302) we obtain the desired result. □

We note that when v vanishes identically, then inequality (287) is automatically satisfied, as then

$$\frac{u(t) - v(t)}{u(t) - v(\nu t)} = 1 \tag{303}$$

identically for every $\nu, t \geq 0$. We prove in the next lemma a more straightforward uniform bound on ψ_R.

Lemma 7. *Let $u : [0, \infty) \to (0, \infty)$ and $v : [0, \infty) \to [0, \infty)$ be two nondecreasing continuous functions such that*

- *$u(t) > v(t)$ for every $t \in [0, \infty)$, and*
- *there is a $k > 0$ such that $u(xy) \leq x^k u(y)$ for every $x \in [1, \infty)$ and $y \in [0, \infty)$,*

and define

$$\mu := \sup_{t \in [0, \infty)} \frac{v(t)}{u(t)}. \tag{304}$$

For every $s \in (0, 1)$, if

$$1 \leq \nu \leq \left(1 + \frac{s(1-\mu)}{1-s}\right)^{1/k} \tag{305}$$

then we have the uniform bound

$$1 - s \leq \frac{u(t) - v(t)}{u(\nu t) - v(t)} \leq 1 \tag{306}$$

for every $t \in [0, \infty)$.

Proof. Fix $s \in (0, 1)$ and ν satisfying (305). As $\nu \geq 1$, monotonicity of u yields that

$$\frac{u(t) - v(t)}{u(\nu t) - v(t)} \leq 1 \tag{307}$$

for every $t \in [0, \infty)$. Further, we have the lower bound

$$\frac{u(t) - v(t)}{u(\nu t) - v(t)} \geq \frac{u(t) - v(t)}{\nu^k u(t) - v(t)} \tag{308}$$

$$= \frac{u(t) - v(t)}{\frac{1 - s \cdot \sup_{x \in [0, \infty)} \frac{v(x)}{u(x)}}{1-s} u(t) - v(t)} \tag{309}$$

$$\geq \frac{u(t) - v(t)}{\frac{1 - s \cdot \frac{v(t)}{u(t)}}{1-s} u(t) - v(t)} \tag{310}$$

$$= \frac{(u(t) - v(t))(1-s)}{u(t) - sv(t) - (1-s)v(t)} = 1 - s, \tag{311}$$

as desired. □

We now to deduce Lemma 5 from Lemmas 6 and 7.

Proof of Lemma 5. We shall first show that u_R and v_R satisfy the premises of Lemmas 6 and 7. For any single-variable polynomial h with nonnegative coefficents

$$h(xy) \leq x^{\deg h} h(y) \tag{312}$$

holds for every $x \in [1, \infty)$ and $y \in [0, \infty)$. We also note that (312) still holds if the exponent $\deg h$ is increased. So, we use $k = n(n+1)/2$ for both u_R and v_R.

We have that

$$\det M_{\sqrt{t}R + N, n} = u_R(t) - v_R(t). \tag{313}$$

Hence, $\det M_{\sqrt{t}R+N,n} > 0$ implies that

$$u_R(t) > v_R(t) \tag{314}$$

for all $t \geq 0$. In fact, since

$$\det M_{R,n} = \beta^{(+)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) - \beta^{(-)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{315}$$

and $\det M_{R,n} > 0$, we have that

$$\lim_{t \to \infty} \frac{v_R(t)}{u_R(t)} = \frac{\beta^{(-)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n})}{\beta^{(+)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n})} < 1. \tag{316}$$

Then, as condition (305) is more restrictive than condition (286) for all small s, specifically, for all $s < (1-\mu)/(1+\mu)$, Lemma 5 follows. □

C. Error in Estimating the Integral: Inequality (277)

We return to proving (277) via proving (271) and (272) and utilizing the uniform bound (284) in Lemma 5. So, defining

$$\mu_R := \sup_{t \in [0, \infty)} \frac{v_R(t)}{u_R(t)}, \tag{317}$$

we assume that ν satisfies

$$\nu < \mu_R^{-2/(n(n+1))} \tag{318}$$

and that the event $E_{n,\eta}(\{R_i\}_{i \in [m]})$ holds. Then,

$$\frac{(1-\eta)^2 f_R((1-\eta)^2 t) - (1+\eta)^2 g_R((1+\eta)^2 t)}{u_R((1+\eta)^2 t) - v_R((1-\eta)^2 t)}$$

$$\leq \frac{f_{W_m}(t) - g_{W_m}(t)}{u_{W_m}(t) - v_{W_m}(t)} \tag{319}$$

$$\leq \frac{(1+\eta)^2 f_R((1+\eta)^2 t) - (1-\eta)^2 g_R((1-\eta)^2 t)}{u_R((1-\eta)^2 t) - v_R((1+\eta)^2 t)} \tag{320}$$

where the positivity of the denominator in the upper bound (320) follows from (292) in Lemma 6. Integrating with respect to t over $[0, \infty)$ then performing a change of variables from t to $(1-\eta)^2 t$, we obtain the bounds

$$\int_0^\infty \frac{f_R(t) - \nu g_R(\nu t)}{u_R(\nu t) - v_R(t)}\, dt \leq \int_0^\infty \frac{f_{W_m}(t) - g_{W_m}(t)}{u_{W_m}(t) - v_{W_m}(t)}\, dt \tag{321}$$

$$\leq \int_0^\infty \frac{\nu f_R(\nu t) - g_R(t)}{u_R(t) - v_R(\nu t)}\, dt, \tag{322}$$

which are the bounds (271), (272).

Fix a real s such that

$$s \in \left(0, \frac{1-\mu_R}{1+\mu_R}\right), \tag{323}$$

and assume that ν further satisfies (283), i.e., that

$$\nu \le \left(1 + \frac{s(1-\mu_R)}{1-s}\right)^{2/(n(n+1))}. \tag{324}$$

We have that

$$\begin{aligned}&\frac{\nu f_R(\nu t) - g_R(t)}{u_R(t) - v_R(\nu t)} \\ &\quad = \varphi_R(t;\nu)\left(\frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} + \frac{\nu f_R(\nu t) - f_R(t)}{u_R(t) - v_R(t)}\right).\end{aligned} \tag{325}$$

By the inequalities in (284), we have that

$$0 \le \varphi_R(t;\nu) - 1 \le s. \tag{326}$$

Hence, nonnegativity of f_R and g_R yields that

$$(\varphi_R(t;\nu) - 1) \cdot \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} \le s \cdot \frac{f_R(t)}{u_R(t) - v_R(t)}, \tag{327}$$

so

$$\varphi_R(t;\nu) \cdot \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} \le \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} + s \cdot \frac{f_R(t)}{u_R(t) - v_R(t)}, \tag{328}$$

By the upper bound in (284) and (312)

$$\varphi_R(t;\nu) \cdot \frac{\nu f_R(\nu t) - f_R(t)}{u_R(t) - v_R(t)} \le \frac{(1+s)(\nu^{\binom{n+1}{2}-1} - 1) f_R(t)}{u_R(t) - v_R(t)}.$$

Hence, inequality (322) yields the upper bound

$$\begin{aligned}&\int_0^\infty \Theta_n(W_m;t)\,dt - \int_0^\infty \Theta_n(R;t)\,dt \\ &\le \left((1+s)\nu^{\binom{n+1}{2}-1} - 1\right)\int_0^\infty \frac{f_R(t)}{u_R(t) - v_R(t)}\,dt.\end{aligned} \tag{329}$$

On the other hand, we have that

$$\begin{aligned}&\frac{f_R(t) - \nu g_R(\nu t)}{u_R(\nu t) - v_R(t)} \\ &\quad = \psi_R(t;\nu)\left(\frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} + \frac{g_R(t) - \nu g_R(\nu t)}{u_R(t) - v_R(t)}\right).\end{aligned} \tag{330}$$

From (284),

$$s \ge 1 - \psi_R(t;\nu) \ge 0. \tag{331}$$

Hence, nonnegativity of f_R and g_R yields that

$$s \cdot \frac{f_R(t)}{u_R(t) - v_R(t)} \ge (1 - \psi_R(t;\nu)) \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)}, \tag{332}$$

so

$$\psi_R(t;\nu)\frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} \ge \frac{f_R(t) - g_R(t)}{u_R(t) - v_R(t)} - s \cdot \frac{f_R(t)}{u_R(t) - v_R(t)}. \tag{333}$$

From $\psi_R(t;\nu) \le 1$ and (312),

$$\psi_R(t;\nu)\frac{g_R(t) - \nu g_R(\nu t)}{u_R(t) - v_R(t)} \ge \left(1 - \nu^{\binom{n+1}{2}-1}\right)\frac{g_R(t)}{u_R(t) - v_R(t)}.$$

Hence, inequality (321) yields the lower bound

$$\begin{aligned}&\int_0^\infty \Theta_n(W_m;t)\,dt - \int_0^\infty \Theta_n(R;t)\,dt \\ &\quad \ge -s\int_0^\infty \frac{f_R(t)}{u_R(t) - v_R(t)}\,dt \\ &\qquad - \left(\nu^{\binom{n+1}{2}-1} - 1\right)\int_0^\infty \frac{g_R(t)}{u_R(t) - v_R(t)}\,dt.\end{aligned} \tag{334}$$

Combining inequalities (329) and (334), we obtain the bound (277), namely,

$$\begin{aligned}&\left|\int_0^\infty \Theta_n(W_m;t)\,dt - \int_0^\infty \Theta_n(R;t)\,dt\right| \\ &\le \left((1+s)\nu^{\binom{n+1}{2}-1} - 1\right)\int_0^\infty \frac{f_R(t) + g_R(t)}{u_R(t) - v_R(t)}\,dt.\end{aligned} \tag{335}$$

D. Error in Estimating $\det M_{R,n}$

We will use equation (315), namely,

$$\det M_{R,n} = \beta^{(+)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) - \beta^{(-)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{336}$$

to bound the error when estimating $\det M_{R,n}$ from the samples $\{R_i\}_{1 \le i \le m}$. We define the random variables

$$\beta^{(+,m)} := \beta^{(+)}_{\binom{n+1}{2}}\left(\frac{\sum_{i=1}^m R_i}{m}, \cdots, \frac{\sum_{i=1}^m R_i^{2n}}{m}\right) \tag{337}$$

$$\beta^{(-,m)} := \beta^{(-)}_{\binom{n+1}{2}}\left(\frac{\sum_{i=1}^m R_i}{m}, \cdots, \frac{\sum_{i=1}^m R_i^{2n}}{m}\right), \tag{338}$$

and as in (279) and (280) we write

$$\beta^{(+)}_{R,n} = \beta^{(+)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{339}$$

$$\beta^{(-)}_{R,n} = \beta^{(-)}_{\binom{n+1}{2}}(\mathcal{R}_1, \cdots, \mathcal{R}_{2n}) \tag{340}$$

for short. Then,

$$\det M_{R,n} = \beta^{(+)}_{R,n} - \beta^{(-)}_{R,n} \tag{341}$$

$$\det M_{W_m,n} = \beta^{(+,m)} - \beta^{(-,m)}. \tag{342}$$

We assume that $m > n$ and

$$0 < \eta < \min\left(\frac{1}{2}, \frac{1}{1 + \frac{2}{\left(\frac{1}{2}\left(\frac{\beta^{(+)}_{R,n}}{\beta^{(-)}_{R,n}} + 1\right)\right)^{1/(n+1)} - 1}}\right). \tag{343}$$

Then we show that under $E_{n,\eta}(\{R_i\}_{i\in[m]})$

$$\left|\frac{1}{n(n+1)}\log\frac{\det M_{W_m,n}}{\det M_{R,n}}\right| \le \frac{6\left(\beta^{(+)}_{R,n} + \beta^{(-)}_{R,n}\right)\eta}{\left(\beta^{(+)}_{R,n} - \beta^{(-)}_{R,n}\right)n} \tag{344}$$

In the proof of (344), we will denote $\beta^{(\pm)}_{R,n}$ by $\beta^{(\pm)}$ for convenience.

First, we note that each term in the polynomials

$$\beta^{(+)}_{\binom{n+1}{2}} \quad \text{and} \quad \beta^{(-)}_{\binom{n+1}{2}} \tag{345}$$

is a product of at most $n+1$ monomials (in fact, also at least n monomials). Thus,

$$(1-\eta)^{n+1}\beta^{(+)} \le \beta^{(+,m)} \le (1+\eta)^{n+1}\beta^{(+)} \tag{346}$$

and

$$(1-\eta)^{n+1}\beta^{(-)} \le \beta^{(-,m)} \le (1+\eta)^{n+1}\beta^{(-)}. \tag{347}$$

We may assume $\beta^{(-)}_{\binom{n+1}{2}}$ is not the zero polynomial, for if it is the zero polynomial then we have a stronger bound than (344), e.g.,

$$\left|\frac{1}{n(n+1)}\log\frac{\det M_{W_m,n}}{\det M_{R,n}}\right| \le \frac{-\log(1-\eta)}{n} < \frac{2\eta}{n} \tag{348}$$

where the last inequality follow because $-\log(1-z) < 2z$ for $z \in (0,1/2)$, which can be verified by checking the derivative. From (346) and (347), we have that

$$\log\frac{\beta^{(+)} - \nu^{\frac{n+1}{2}}\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}} + (n+1)\log(1-\eta) \le \log\frac{\det M_{W_m,n}}{\det M_{R,n}} \tag{349}$$

and

$$\log\frac{\det M_{W_m,n}}{\det M_{R,n}} \le \log\frac{\beta^{(+)} - \nu^{\frac{n+1}{-2}}\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}} + (n+1)\log(1+\eta) \tag{350}$$

where we used the fact that

$$\nu^{\frac{n+1}{2}} < \frac{\beta^{(+)}}{\beta^{(-)}}. \tag{351}$$

We note that

$$|\log(1+\eta)| < |\log(1-\eta)|. \tag{352}$$

Further, 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

$$\frac{w - z/r}{w-z} < \frac{w-z}{w-rz}. \tag{353}$$

Setting $(w,z,r) = (\beta^{(+)}, \beta^{(-)}, \nu^{(n+1)/2})$, we obtain that

$$\left|\log\frac{\beta^{(+)} - \nu^{\frac{n+1}{-2}}\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}}\right| < \left|\log\frac{\beta^{(+)} - \nu^{\frac{n+1}{2}}\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}}\right|. \tag{354}$$

Hence,

$$\left|\log\frac{\det M_{W_m,n}}{\det M_{R,n}}\right| \le \log\frac{\beta^{(+)}-\beta^{(-)}}{\beta^{(+)} - \nu^{\frac{n+1}{2}}\beta^{(-)}} + (n+1)\log\frac{1}{1-\eta}. \tag{355}$$

Now, we may write

$$\frac{\beta^{(+)}-\beta^{(-)}}{\beta^{(+)} - \nu^{\frac{n+1}{2}}\beta^{(-)}} = \left(1 - \frac{\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}}\left(\nu^{\frac{n+1}{2}} - 1\right)\right)^{-1}. \tag{356}$$

The proof is complete once we show that for any $(w,z,r) \in \mathbb{R}^3_{>0}$ such that $(1+z)^r < 1 + \frac{1}{2w}$ we have that

$$-\log\left(1 - w\left((1+z)^r - 1\right)\right) \le (2w+1)rz. \tag{357}$$

Before showing that (357) holds, we note how it completes the proof. Setting

$$(w,z,r) = \left(\frac{\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}}, \frac{2\eta}{1-\eta}, n+1\right), \tag{358}$$

we obtain that

$$\log\frac{\beta^{(+)}-\beta^{(-)}}{\beta^{(+)} - \nu^{\frac{n+1}{2}}\beta^{(-)}} \le \frac{\beta^{(+)}+\beta^{(-)}}{\beta^{(+)}-\beta^{(-)}} \cdot (n+1) \cdot \frac{2\eta}{1-\eta} \tag{359}$$

since

$$\nu^{\frac{n+1}{2}} < \frac{1}{2}\left(\frac{\beta^{(+)}}{\beta^{(-)}} + 1\right). \tag{360}$$

Then $-\log(1-\eta) < 2\eta$ yields inequality (344).

Finally, to see that (357) holds, we consider for fixed $w, r > 0$

$$f(z) := (2w+1)rz + \log\left(1 - w\left((1+z)^r - 1\right)\right) \tag{361}$$

over

$$z \in \left(0, \left(1 + \frac{1}{2w}\right)^{1/r} - 1\right). \tag{362}$$

Inequality (357) follows since f is continuous, $f(0) = 0$, and

$$f'(z) = (2w+1)r - \frac{wr(1+z)^{r-1}}{1 - w((1+z)^r - 1)} \tag{363}$$

$$> (2w+1)r - \frac{wr(1+z)^r}{1 - w((1+z)^r - 1)} \tag{364}$$

$$> (2w+1)r - \frac{wr(1+1/(2w))}{1 - w((1+1/(2w)) - 1)} = 0 \tag{365}$$

for every z in the domain of f.

E. Empirical Measure Error

By Hoeffding's inequality we have that for any $\zeta > 0$

$$\begin{aligned}
&\Pr\left\{\max_{x\in\mathcal{X}}\left|\widehat{P_X}^{(m)}(x) - P_X(x)\right| \ge \zeta\right\} \\
&= \Pr\left\{\bigcup_{x\in\mathcal{X}}\left\{\left|\widehat{P_X}^{(m)}(x) - P_X(x)\right| \ge \zeta\right\}\right\} \\
&\le \sum_{x\in\mathcal{X}} \Pr\left\{\left|\widehat{P_X}^{(m)}(x) - P_X(x)\right| \ge \zeta\right\} \qquad (366) \\
&\le 2|\mathcal{X}|e^{-2m\zeta^2}. \qquad (367)
\end{aligned}$$

Therefore, for any $\zeta, \delta > 0$, if

$$m > \frac{1}{2\zeta^2}\log\frac{6|\mathcal{X}|}{\delta} \tag{368}$$

then

$$\Pr\left\{\max_{x\in\mathcal{X}}\left|\widehat{P_X}^{(m)}(x) - P_X(x)\right| < \zeta\right\} > 1 - \frac{\delta}{3}. \tag{369}$$

We denote

$$F_\zeta(\{X_j\}_{1\le j\le m}) := \left\{\max_{x\in\mathcal{X}}\left|\widehat{P_X}^{(m)}(x) - P_X(x)\right| < \zeta\right\}. \tag{370}$$

F. Proof of Proposition 2

We define the following constants

$$\mu_{X,Y,n} := \max\left(\mu_Y,\ \max_{x\in\mathcal{X}} \mu_{Y_x}\right) \tag{371}$$

$$\rho_{X,Y,n} := \min\left(\frac{\beta^{(+)}_{Y,n}}{\beta^{(-)}_{Y,n}}, \min_{x\in\mathcal{X}} \frac{\beta^{(+)}_{Y_x,n}}{\beta^{(-)}_{Y_x,n}}\right) \tag{372}$$

$$A^{(1)}_{X,Y,n} := \frac{1}{2}\int_0^\infty \frac{f_Y(t)+g_Y(t)}{u_Y(t)-v_Y(t)}\,dt \tag{373}$$

$$+\frac{1}{2}\sum_{x\in\mathcal{X}}\int_0^\infty \frac{f_{Y_x}(t)+g_{Y_x}(t)}{u_{Y_x}(t)-v_{Y_x}(t)}\,dt \tag{374}$$

$$A^{(2)}_{X,Y,n} := \frac{6}{n}\left(\frac{\beta^{(+)}_{Y,n}+\beta^{(-)}_{Y,n}}{\beta^{(+)}_{Y,n}-\beta^{(-)}_{Y,n}} + \sum_{x\in\mathcal{X}}\frac{\beta^{(+)}_{Y_x,n}+\beta^{(-)}_{Y_x,n}}{\beta^{(+)}_{Y_x,n}-\beta^{(-)}_{Y_x,n}}\right) \tag{375}$$

$$A^{(3)}_{X,Y,n} := \sum_{x\in\mathcal{X}}\left|\frac{\log\det M_{Y_x,n}}{n(n+1)} + \frac{1}{2}\int_0^\infty \Theta_n(Y_x;t)\,dt\right|. \tag{376}$$

We note that

$$\mu_{X,Y,n} < 1 < \rho_{X,Y,n}. \tag{377}$$

For each $\eta,\zeta>0$, we define the event

$$G_{n,\eta,\zeta}(\mathcal{S}^{(m)}) := D_n(\mathcal{S}^{(m)}) \cap E_{n,\eta}(\{Y_j\}_{j\in[m]}) \cap \bigcap_{x\in\mathcal{X}} E_{n,\eta}(\{Y_j\}_{j\in\mathcal{T}_x^{(m)}}) \cap F_\zeta(\{X_j\}_{j\in[m]}). \tag{378}$$

For any strictly positive reals s,η, and ζ such that

$$s < \frac{1-\mu_{X,Y,n}}{1+\mu_{X,Y,n}} \tag{379}$$

and

$$\eta < \min\left(\frac{1}{2}, \frac{1}{1+\frac{2}{\min\left(\rho^{1/(n+1)}_{X,Y,n}, \left(\frac{1-s\mu_{X,Y,n}}{1-s}\right)^{1/(n(n+1))}\right)-1}}\right), \tag{380}$$

if $G_{n,\eta,\zeta}(\mathcal{S}^{(m)})$ occurs, then we have the estimate

$$\left|\widehat{I}_n\left(\mathcal{S}^{(m)}\right) - I_n(X;Y)\right| \le A^{(1)}_{X,Y,n}\left((1+s)\nu^{\binom{n+1}{2}}-1\right) + A^{(2)}_{X,Y,n}\eta + A^{(3)}_{X,Y,n}\zeta. \tag{381}$$

We utilize (381) to find a sufficient condition on m so that the estimate $\widehat{I}_n\left(\mathcal{S}^{(m)}\right)$ is within ε, for every small enough ε, of $I_n(X;Y)$ with high probability.

We define

$$\omega_{X,Y,n} := \frac{\min(1-\mu_{X,Y,n}, 2/3)}{6A^{(1)}_{X,Y,n}} \tag{382}$$

$$\varepsilon_{X,Y,n} := \min\left(6A^{(1)}_{X,Y,n}\min\left(\frac{1}{2}, \frac{1-\mu_{X,Y,n}}{1+\mu_{X,Y,n}}\right), \frac{\rho^n_{X,Y,n}-1}{\omega_{X,Y,n}}, \frac{3A^{(2)}_{X,Y,n}}{2}\right), \tag{383}$$

and fix $\varepsilon\in(0,\varepsilon_{X,Y,n})$ and $\delta\in(0,1/2)$. Setting

$$s = \frac{\varepsilon}{6A^{(1)}_{X,Y,n}}, \tag{384}$$

we get that (379) is satisfied. We will show next that setting

$$\eta = \varepsilon\cdot\min\left(\frac{1}{3A^{(2)}_{X,Y,n}}, \frac{3\omega_{X,Y,n}\cdot\left((4/3)^{1/(n(n+1))}-1\right)}{(4/3)^{1/(n(n+1))}+1}\right) \tag{385}$$

yields that

$$\eta < \frac{1}{1+\frac{2}{(1+\varepsilon\omega_{X,Y,n})^{1/(n(n+1))}-1}}. \tag{386}$$

Since

$$\varepsilon_{X,Y,n}\omega_{X,Y,n} \in \left(0,\frac{1}{3}\right), \tag{387}$$

and since for any $\theta>1$ the condition $\nu<\theta$ is equivalent to

$$\eta < \frac{1}{1+\frac{2}{\sqrt{\theta}-1}}, \tag{388}$$

we see that inequality (386) would imply that both inequalities (380) and

$$\nu^{\binom{n+1}{2}} < 1 + \frac{\varepsilon}{9A^{(1)}_{X,Y,n}} \tag{389}$$

hold. Then, setting

$$\zeta = \frac{\varepsilon}{3A^{(3)}_{X,Y,n}} \tag{390}$$

would give that each of the three terms in the upper bound in (381) is at most $\varepsilon/3$, thereby making

$$\left|\widehat{I}_n(\mathcal{S}^{(m)}) - I_n(X;Y)\right| \le \varepsilon \tag{391}$$

whenever $G_{n,\eta,\zeta}(\mathcal{S}^{(m)})$ occurs. We next show that (386) holds, then we give a threshold on m making $G_{n,\eta,\zeta}(\mathcal{S}^{(m)})$ occur with probability higher than $1-\delta$.

For (386), it suffices to show that

$$\frac{3\left((4/3)^{1/(n(n+1))}-1\right)}{(4/3)^{1/(n(n+1))}+1} < \frac{(1+\varepsilon\omega_{X,Y,n})^{1/(n(n+1))}-1}{\varepsilon\omega_{X,Y,n}((1+\varepsilon\omega_{X,Y,n})^{1/(n(n+1))}+1)}. \tag{392}$$

We show that (392) holds by showing that, for any $\theta\in(0,1)$, the function $f:(1,\infty)\to(0,\infty)$ defined by

$$f(z) = \frac{z^\theta-1}{(z-1)(z^\theta+1)} \tag{393}$$

is strictly decreasing. As $\varepsilon\omega_{X,Y,n} < 1/3$, inequality (392) would follow. Now, to show that f is strictly decreasing, it suffices to check that $z\mapsto(z^\theta+1)f(z)$ is strictly decreasing. But

$$\frac{\partial}{\partial z}((z^\theta+1)f(z)) = \frac{(\theta-1)z+z^{1-\theta}-\theta}{z^{1-\theta}(z-1)^2}. \tag{394}$$

Let $g:(1,\infty)\to\mathbb{R}$ be defined by

$$g(z) := (\theta-1)z + z^{1-\theta} - \theta. \tag{395}$$

We have that

$$g'(z) = (1-\theta)(z^{-\theta} - 1) < 0 \tag{396}$$

for every z. Further, $g(0) = -\theta < 0$. Hence, g is strictly negative over its domain. Thus, $(z^\theta + 1)f(z)$ is strictly decreasing over its domain, as desired.

Finally, we give a sufficient condition on m that makes the probability of occurrence of $G_{n,\eta,\zeta}(\mathcal{S}^{(m)})$ larger than $1-\delta$. Let

$$\kappa_{X,Y,n} := \min\left(\frac{1}{3A^{(2)}_{X,Y,n}}, \frac{3\omega_{X,Y,n}\cdot\left((4/3)^{1/(n(n+1))}-1\right)}{(4/3)^{1/(n(n+1))}+1}\right) \tag{397}$$

so

$$\eta = \varepsilon \cdot \kappa_{X,Y,n}. \tag{398}$$

We set

$$C_{X,Y,n} := \max\left(\frac{2\varepsilon^2_{X,Y,n}\log_2(6|\mathcal{X}|)}{P_X(x_0)^2}, \frac{2n\varepsilon^2_{X,Y,n}}{P_X(x_0)\log 2}, \frac{9}{2}\left(A^{(3)}_{X,Y,n}\right)^2\log_2(12|\mathcal{X}|), \frac{\left((q/p)^{2n}-1\right)^2\log_2(48n(|\mathcal{X}|+1))}{P_X(x_0)\kappa^2_{X,Y,n}}\right), \tag{399}$$

and assume that

$$m > \frac{C_{X,Y,n}}{\varepsilon^2}\log\frac{1}{\delta}. \tag{400}$$

It suffices to show that

$$\Pr\left(G_{n,\eta,\zeta}(\mathcal{S}^{(m)}) \cap D_\ell(\mathcal{S}^{(m)})\right) > 1-\delta \tag{401}$$

for any $\ell > 0$.

We note that for any $z > 1$

$$\log_2(2z)\log\frac{1}{\delta} > \log\frac{z}{\delta}. \tag{402}$$

Then, we have that

$$m > \max\left(\frac{2n}{P_X(x_0)}, \frac{2}{P_X(x_0)^2}\log\frac{3|\mathcal{X}|}{\delta}\right). \tag{403}$$

Thus,

$$\Pr\left(D_n(\mathcal{S}^{(m)})\right) > 1 - \frac{\delta}{3}. \tag{404}$$

We have, with

$$\ell = \frac{\left((q/p)^{2n}-1\right)^2}{2\eta^2}\log\frac{24n(|\mathcal{X}|+1)}{\delta}, \tag{405}$$

that m satisfies

$$m > \frac{2\ell}{P_X(x_0)}. \tag{406}$$

Hence,

$$\Pr\left(D_\ell(\mathcal{S}^{(m)})\right) > 1 - \frac{\delta}{6}. \tag{407}$$

Furthermore,

$$\Pr\left(E_{n,\eta}(\{Y_j\}_{j\in[m]}) \cap \bigcap_{x\in\mathcal{X}} E_{n,\eta}(\{Y_j\}_{j\in\mathcal{T}_x^{(m)}}) \,\middle|\, D_\ell(\mathcal{S}^{(m)})\right) > 1 - \frac{\delta}{6}. \tag{408}$$

Thus,

$$\Pr\left(E_{n,\eta}(\{Y_j\}_{j\in[m]}) \cap \bigcap_{x\in\mathcal{X}} E_{n,\eta}(\{Y_j\}_{j\in\mathcal{T}_x^{(m)}}) \cap D_\ell(\mathcal{S}^{(m)})\right) > 1 - \frac{\delta}{3}. \tag{409}$$

Finally,

$$m > \frac{1}{2\zeta^2}\log\frac{6|\mathcal{X}|}{\delta}, \tag{410}$$

so

$$\Pr\left(F_\zeta(\{X_j\}_{j\in[m]})\right) > 1 - \frac{\delta}{3}. \tag{411}$$

Hence, (401) follows and the proof is complete.

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.