Polynomial Singular Value Decompositions of a Family of Source-Channel Models

Anuran Makur, *Student Member, IEEE*, and Lizhong Zheng, *Fellow, IEEE*

***Abstract*— In this paper, we show that the conditional expectation operators corresponding to a family of source-channel models, defined by natural exponential families with quadratic variance functions and their conjugate priors, have orthonormal polynomials as singular vectors. These models include the Gaussian channel with Gaussian source, the Poisson channel with gamma source, and the binomial channel with beta source. To derive the singular vectors of these models, we prove and employ the equivalent condition that their conditional moments are strictly degree preserving polynomials.**

***Index Terms*— Singular value decomposition, natural exponential family, conjugate prior, orthogonal polynomials.**

I. Introduction

SPECTRAL and singular value decompositions (SVDs) of conditional expectation operators have many uses in information theory and statistics [1]–[3]. As a result, it is valuable to analytically determine the singular vectors corresponding to some widely studied toy models. In this paper, we illustrate that a certain simple family of source-channel models always has corresponding conditional expectation operators with orthogonal polynomial singular vectors. We commence by presenting this family of models and formally defining conditional expectation operators in the next two subsections.

A. Natural Exponential Families With Quadratic Variance Functions and Their Conjugate Priors

Since we will study source-channel models that have exponential family and conjugate prior structure, we briefly introduce these notions. Exponential families form an important class of distributions in statistics because they are analytically tractable and intimately tied to several theoretical phenomena [4], [5]. For instance, they have sufficient statistics with bounded dimension after i.i.d. sampling (Pitman-Koopman-Darmois theorem) [6], they have conjugate priors [7], they admit efficient estimators that achieve the Cramér-Rao bound under a mean parametrization [5], they are maximum entropy distributions under moment constraints [8], and they are used in tilting arguments in large deviations theory [5]. We are interested in a particular subclass of one-parameter exponential families known as *natural exponential families with quadratic variance functions* (NEFQVF). So, we define natural exponential families next.

Definition 1 (Natural Exponential Family): Given a measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ with a σ-finite measure μ, where $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ-algebra on $\mathcal{Y}$, the parametrized family of probability densities $\{P_Y(\cdot; x) : x \in \mathcal{X}\}$ with respect to μ that have support $\mathcal{Y}$ (independent of x) is called a *natural exponential family* when each density has the form:

$$\forall x \in \mathcal{X}, \quad \forall y \in \mathcal{Y}, \quad P_Y(y; x) = \exp\left(xy - \alpha(x)\right) P_Y(y; 0)$$

where $P_Y(\cdot; 0)$ is a *base density function*, and:

$$\forall x \in \mathcal{X}, \quad \alpha(x) = \log\left(\int_{\mathcal{Y}} \exp(xy)\, P_Y(y; 0)\, d\mu(y)\right)$$

is known as the *log-partition function* which satisfies $\alpha(0) = 0$ without loss of generality. The parameter x is called the *natural parameter*, and the *natural parameter space* $\mathcal{X} \triangleq \{x \in \mathbb{R} : |\alpha(x)| < +\infty\} \subseteq \mathbb{R}$ is defined as the largest interval where the log-partition function is finite. We usually assume without loss of generality that $0 \in \mathcal{X}$. (Here, and throughout this paper, $\exp(\cdot)$ and $\log(\cdot)$ refer to the natural exponential and the natural logarithm with the base e, respectively.)

In [9] and [10], Morris specialized Definition 1 further in an effort to justify why certain natural exponential families like the Gaussian, Poisson, and binomial enjoy "many useful mathematical properties" [9]. He asserted that the tractability of these distributions stemmed from their quadratic variance functions. To define this, observe that $\alpha(\cdot)$ is infinitely differentiable on $\mathcal{X}^\circ$ (the interior of $\mathcal{X}$) [4], and satisfies:

$$\forall x \in \mathcal{X}, \quad \alpha(x) = \log\left(\mathbb{E}_{P_Y(\cdot;0)}\left[\exp(xY)\right]\right) \tag{1}$$

$$\forall x \in \mathcal{X}^\circ, \quad \alpha'(x) = \mathbb{E}_{P_Y(\cdot;x)}[Y] \tag{2}$$

$$\forall x \in \mathcal{X}^\circ, \quad \alpha''(x) = \mathbb{VAR}_{P_Y(\cdot;x)}(Y) \tag{3}$$

where Y denotes a random variable taking values in $\mathcal{Y}$, (1) is the cumulant generating function of Y, and (3) is the Fisher information Y carries about x [5]. Following the exposition in [9], we may define the *variance function* $V : \mathsf{image}(\alpha') \to \mathbb{R}^+$ as the variance of Y written as a function of the mean of Y:

$$\forall \gamma \in \mathsf{image}(\alpha'), \quad V(\gamma) \triangleq \alpha''(\alpha'^{-1}(\gamma)) \tag{4}$$

Manuscript received December 9, 2015; revised July 16, 2017; accepted September 21, 2017. Date of publication October 6, 2017; date of current version November 20, 2017. This work was supported in part by the National Science Foundation under Award 1216476 and in part by the Hewlett-Packard Fellowship. This work was presented at the 2016 54th Annual Allerton Conference on Communication, Control, and Computing.

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: a_makur@mit.edu; lizhong@mit.edu).

Communicated by C. Nair, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2017.2760626

where $\alpha'(\cdot)$ is injective because $\alpha''(\cdot)$ is strictly positive in non-degenerate scenarios. We can now define NEFQVFs.

Definition 2 (NEFQVF): An NEFQVF is a natural exponential family whose variance function $V(\gamma)$ is a polynomial in γ with degree at most 2.

We will only analyze channel conditional distribution models $P_{Y|X}(y|x) = P_Y(y;x)$ that are NEFQVFs (although we will not explicitly use the NEFQVF parametrization in our calculations). There are six possible NEFQVFs [9]:

1) Gaussian pdfs with mean parameter and fixed variance
2) Poisson pmfs with rate parameter
3) binomial pmfs with success probability parameter and fixed number of Bernoulli trials
4) gamma pdfs with rate parameter and fixed "shape"
5) negative binomial pmfs with success probability parameter and fixed "number of failures"
6) generalized hyperbolic secant pdfs (see [9] for details regarding this family)

and only the first three will lead to non-degenerate situations.

Given a channel $P_{Y|X}(y|x) = P_Y(y;x)$ defined by an NEFQVF, we will only analyze source distributions that belong to the corresponding conjugate prior family. For any natural exponential family, we may define a conjugate prior family as shown next [4], [5].

Definition 3 (Conjugate Prior): Suppose we are given a natural exponential family from Definition 1 such that $\mathcal{X}$ is a non-empty open interval defining the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with σ-finite measure λ. The corresponding *conjugate prior* family is the parametrized family of probability densities $\{P_X(\cdot; z, n) : (z, n) \in \Xi\}$ with respect to λ that have support $\mathcal{X}$ (independent of (z, n)), and are of the form:

$$\forall x \in \mathcal{X}, \quad P_X(x; z, n) = \exp\left(zx - n\alpha(x) - \tau(z, n)\right)$$

for any $(z, n) \in \Xi$, where the *log-partition function* $\tau : \Xi \to \mathbb{R}$ is given by:

$$\forall\ (z, n) \in \Xi, \quad \tau(z, n) = \log\left(\int_{\mathcal{X}} \exp\left(zx - n\alpha(x)\right)\, d\lambda(x)\right)$$

and (z, n) are *hyper-parameters* that belong to the *hyper-parameter space* $\Xi \triangleq \{(z, n) \in \mathbb{R} \times \mathbb{R} : |\tau(z, n)| < +\infty\}$.

When channels are given by natural exponential families, if we use a conjugate prior source, then posterior distributions also belong to the conjugate family. This structure allows computationally efficient updating of beliefs in Bayesian inference problems [5]. A comprehensive list of different conjugate prior families can be compiled from [5], [10], [11], and we will present the conjugate prior families for the first three NEFQVFs listed above (without tediously referring back to the aforementioned sources) in section II.

B. Conditional Expectation Operators

We next formally define conditional expectation operators. We fix a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and define an input random variable $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}$ with *source* probability density P_X with respect to a σ-finite measure λ on the standard measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Likewise, we define an output random variable $Y : \Omega \to \mathcal{Y} \subseteq \mathbb{R}$, and *channel* conditional probability densities $\left\{P_{Y|X=x} : x \in \mathcal{X}\right\}$ with respect to a σ-finite measure μ on the standard measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. We will use the notation $\mathbb{P}_X$ and $\mathbb{P}_Y$ to denote the marginal probability laws of X and Y, and we will assume that $\mathbb{P}_X$ and $\mathbb{P}_Y$ have (measure theoretic) supports $\overline{\mathcal{X}}$ and $\overline{\mathcal{Y}}$ (which are the closures of $\mathcal{X}$ and $\mathcal{Y}$), respectively. Finally, we note that this *source-channel model* defines a joint probability density $P_{X,Y}$ on the product measure space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}), \lambda \times \mu)$ such that $P_{X,Y}(x, y) = P_{Y|X}(y|x)P_X(x)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In section II, our channels will be NEFQVFs and our sources will be the corresponding conjugate priors.

We next define the Hilbert spaces and linear operators pertinent to our discussion. Corresponding to the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$, we define the separable Hilbert space $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ over the field $\mathbb{R}$:

$$\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \triangleq \left\{f : \mathcal{X} \to \mathbb{R} \,\middle|\, \mathbb{E}\left[f^2(X)\right] < +\infty\right\} \tag{5}$$

which is the space of all Borel measurable and $\mathbb{P}_X$-square integrable functions, with correlation as the inner product:

$$\forall f, g \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X), \quad \langle f, g\rangle_{\mathbb{P}_X} \triangleq \mathbb{E}[f(X)g(X)] \tag{6}$$

and induced norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X), \ \|f\|_{\mathbb{P}_X} \triangleq \langle f, f\rangle_{\mathbb{P}_X}^{\frac{1}{2}} = \mathbb{E}\left[f^2(X)\right]^{\frac{1}{2}}. \tag{7}$$

Likewise, we define the separable Hilbert space $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ corresponding to the measure space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathbb{P}_Y)$. The *conditional expectation operators* are maps that are defined between these Hilbert spaces. The "forward" conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is defined as:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X), \quad (C(f))(y) \triangleq \mathbb{E}[f(X)|Y = y], \tag{8}$$

and the "reverse" conditional expectation operator $C^* : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ is defined as:

$$\forall g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y), \quad \left(C^*(g)\right)(x) \triangleq \mathbb{E}[g(Y)|X = x]. \tag{9}$$

It is straightforward to verify from (8) and (9) that the codomains of C and C^* are indeed Hilbert spaces. The next proposition collects some simple properties of these operators.

Proposition 1 (Conditional Expectation Operators): C and C^* are bounded linear operators with operator norms $\|C\|_{\mathsf{op}} = \|C^*\|_{\mathsf{op}} = 1$. Moreover, C^* is the adjoint operator of C.

Proof: See Appendix A. ∎

Given NEFQVF channels and conjugate prior sources, we will prove that the corresponding operators C and C^* have singular vectors that are orthonormal polynomials under the regularity condition that the input and output Hilbert spaces have orthonormal polynomial bases. The ensuing two subsections provide some illustrations from the literature where such SVDs can be useful.

C. Maximal Correlation Functions

In statistics, one utility of singular vectors of conditional expectation operators is that they can be construed as "maximal correlation functions." To explain this, we first recall the Hirschfeld-Gebelein-Rényi *maximal correlation*, which is

a variational generalization of the well-known Pearson correlation coefficient. Given two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, the maximal correlation between them is:

$$\rho(X;Y) \triangleq \sup_{f,\,g} \mathbb{E}\left[f(X)g(Y)\right] \tag{10}$$

where the supremum is over all functions $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$ [1]. Furthermore, if X or Y is a constant almost surely, then $\rho(X;Y) = 0$. $\rho(X;Y)$ was originally introduced as a normalized measure of the statistical dependence between X and Y that satisfies seven "reasonable" axioms [1]. Indeed, $0 \le \rho(X;Y) \le 1$, and $\rho(X;Y) = 0$ if and only if X and Y are independent random variables.

Maximal correlation turns out to have an elegant spectral characterization. Notice that the everywhere unity functions $\mathbf{1}_{\mathcal{X}} \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $\mathbf{1}_{\mathcal{Y}} \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ (which are defined in Appendix A) are the right and left singular vectors of the conditional expectation operator C corresponding to its largest singular value of $\|C\|_{\mathsf{op}} = 1$:

$$C(\mathbf{1}_{\mathcal{X}}) = \mathbf{1}_{\mathcal{Y}} \quad \text{and} \quad C^*\big(\mathbf{1}_{\mathcal{Y}}\big) = \mathbf{1}_{\mathcal{X}}. \tag{11}$$

The orthogonal complement of the span of this right singular vector, $\mathsf{span}(\mathbf{1}_{\mathcal{X}})^{\perp} = \big\{f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_{\mathcal{X}}) : \mathbb{E}[f(X)] = 0\big\}$, is a sub-Hilbert space of $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$. Indeed, it is clearly a linear subspace of $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ that inherits the same inner product (6), and its completeness follows from the continuity of the inner product. It is proved in [1, Th. 1] that maximal correlation can be written as the Courant-Fischer-Weyl variational characterization of the second largest singular value of C:

$$\rho(X;Y) = \sup_{f \in \mathsf{span}(\mathbf{1}_{\mathcal{X}})^{\perp} \setminus \{\mathbf{0}\}} \frac{\|C(f)\|_{\mathbb{P}_Y}}{\|f\|_{\mathbb{P}_X}} \tag{12}$$

where $\mathbf{0}$ denotes the zero function. If C is a compact operator, the supremum in (12) is actually achieved by some right singular vector $f^\star \in \mathsf{span}(\mathbf{1}_{\mathcal{X}})^{\perp}$. Furthermore, $f^\star$ and the corresponding left singular vector $g^\star = C(f^\star)/\|C(f^\star)\|_{\mathbb{P}_Y}$ are precisely the maximal correlation functions achieving the supremum in (10).

Maximal correlation functions can also be construed as the solutions to a general version of *non-linear regression* studied in [2]:

$$\min_{f \in \mathcal{F},\, g \in \mathcal{G}} \mathbb{E}\Big[\big(f(X) - g(Y)\big)^2\Big] \tag{13}$$

where $\mathcal{F} = \big\{f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : \mathbb{E}[f(X)] = 0,\ \mathbb{E}[f^2(X)] = 1\big\}$ and $\mathcal{G} = \big\{g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) : \mathbb{E}[g(Y)] = 0,\ \mathbb{E}[g^2(Y)] = 1\big\}$ are collections of arbitrary (non-linear) Borel measurable functions, and we assume the minimum exists. Note that when real data (that is assumed to be drawn i.i.d. from $P_{X,Y}$) is given, the idealized problem in (13) can be modified by replacing the theoretical (population) expectations with empirical (sample) expectations. Breiman and Friedman proposed the *alternating conditional expectations* (ACE) algorithm in [2] to solve (13) and find the optimal $f^\star \in \mathcal{F}$ and $g^\star \in \mathcal{G}$ that provide the best linear relationship between $f^\star(X)$ and $g^\star(Y)$. Moreover, these $f^\star$ and $g^\star$ are also the maximal correlation functions that achieve (10), because $\mathbb{E}[(f(X) - g(Y))^2] = \mathbb{E}[f^2(X)] - 2\mathbb{E}[f(X)g(Y)] + \mathbb{E}[g^2(Y)] = 2 - 2\mathbb{E}[f(X)g(Y)]$. Hence, the non-linear regression problem (13) studied in [2] is equivalent to the maximal correlation problem (10).

While the singular vectors $f^\star$ and $g^\star$ of C have palpable significance in the contexts of regression and maximal correlation, we may impart other singular vectors of C with similar operational interpretations. The pair of singular vectors corresponding to the kth largest singular value of C (for $k \in \{2, 3, 4, \dots\}$) are the functions that are maximally correlated and orthogonal to all previous pairs of singular vectors. Hence, we refer to all such singular vectors as "maximal correlation functions." Maximal correlation functions associated with larger singular values of C can be interpreted as more informative score functions, and are useful in decomposing information into several mutually orthogonal parts. Indeed, such functions are used to perform inference on hidden Markov models in an image processing context in [12], and algorithms based on the ACE algorithm to learn such functions are presented in [13]. These algorithms are essentially power iteration methods to compute singular vectors of C. Our main results in section II provide explicit characterizations of maximal correlation functions for conditional expectation operators defined by NEFQVFs and their conjugate priors.

D. Local Perturbation Arguments in Information Theory

SVDs of conditional expectation operators are also useful when performing perturbation arguments in network information theory. For instance, SVDs of Gaussian conditional expectation operators are used to demonstrate that non-Gaussian codes can achieve higher rates than Gaussian codes for various Gaussian networks in [3] (where in particular, the strong Shamai-Laroia conjecture for the Gaussian ISI channel is disproved). As another example, we briefly delineate the linear information coupling problem studied in [14].

Suppose $\mathcal{X}$ and $\mathcal{Y}$ are finite sets (and λ and μ are counting measures), $P_{X,Y}$ is a joint pmf such that $P_X(x) > 0$ for every $x \in \mathcal{X}$ and $P_Y(y) > 0$ for every $y \in \mathcal{Y}$, and $U \in \mathcal{U}$ (with $|\mathcal{U}| < \infty$) is an arbitrary random variable that is conditionally independent of Y given X so that $U \to X \to Y$ is a Markov chain. For any fixed $\epsilon \neq 0$, we first consider the extremal problem that maximizes $I(U;Y)$ with the constraint that only a thin layer of information can pass through X:

$$\sup_{\substack{P_U,\, P_{X|U} : U \to X \to Y \\ I(U;X) \le \frac{1}{2}\epsilon^2}} I(U;Y) \tag{14}$$

where the supremum is over all P_U and $P_{X|U}$ such that $P_{X,Y}$ is fixed (or equivalently, over all $P_{U|X}$). Note that problem (14) and some of its variants have also been considered in the contexts of investment portfolio theory [15], the information bottleneck method [16], and strong data processing inequalities [17]–[19]. Then, we assume each conditional pmf $P_{X|U=u}$ for $u \in \mathcal{U}$ is a (multiplicative) local perturbation of P_X by $\phi_u \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$:

$$\forall u \in \mathcal{U}, \quad P_{X|U=u} = P_X(\mathbf{1}_{\mathcal{X}} + \epsilon\, \phi_u) \tag{15}$$

where the sums and products in (15) hold pointwise, and for every $u \in \mathcal{U}$, $\mathbb{E}[\phi_u(X)] = 0$ so that $P_{X|U=u}$ is a valid pmf.

From the Markov relation $U \to X \to Y$, we have:

$$\forall u \in \mathcal{U}, \quad P_{Y|U=u} = P_Y(\mathbf{1}_{\mathcal{Y}} + \epsilon\, C(\phi_u)) \tag{16}$$

where the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ was defined in (8). Using (15) and (16), we can locally approximate the mutual information terms in (14) via the result in Corollary 1 of Appendix B. Neglecting all $o(\epsilon^2)$ terms, this produces the *linear information coupling* problem:

$$\max_{P_U,\, \{\phi_u : u \in \mathcal{U}\}} \sum_{u \in \mathcal{U}} P_U(u)\, \|C(\phi_u)\|^2_{\mathbb{P}_Y} \tag{17}$$

where we maximize over all P_U and $\{\phi_u \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : u \in \mathcal{U}\}$ subject to the constraints: $\mathbb{E}[\phi_u(X)] = 0$ for every $u \in \mathcal{U}$, $\sum_{u \in \mathcal{U}} P_U(u)\, \|\phi_u\|^2_{\mathbb{P}_X} \le 1$, and $\sum_{u \in \mathcal{U}} P_U(u) P_X \phi_u = \mathbf{0}$ (which ensures that the marginal pmf of X is fixed at P_X).

It is straightforward to verify that problem (17) can be solved by setting $\mathcal{U} = \{-1, 1\}$ with $P_U(-1) = P_U(1) = \frac{1}{2}$ (i.e. $U \sim$ Rademacher), letting ϕ_1 be a unit norm right singular vector of C corresponding to its second largest singular value, and setting $\phi_{-1} = -\phi_1$. Hence, the SVD of a conditional expectation operator C solves the linear information coupling problem by identifying the optimal perturbations as right singular vectors of C. Moreover, problem (17) is actually a single letter case of a more general multi-letter problem, which can be solved via single letterization using tensorization properties of the SVD [14]. Huang and Zheng [14] exploit this tensorization to study questions in network information theory.

E. Outline

Having illustrated the utility of SVDs of conditional expectation operators, we briefly outline the remaining discussion. In section II, we will state the polynomial SVDs of three source-channel models in key theorems. In section III, we will present the proofs of these results via a useful lemma.

II. Main Results

In this section, we present our main results. Informally, we show that:

Conditional expectation operators corresponding to every NEFQVF channel and its conjugate prior source, such that all moments of the marginal distributions exist and are finite, have orthonormal polynomial singular vectors.

It is straightforward to verify that the moments of the output marginal distributions corresponding to the gamma, negative binomial, and generalized hyperbolic secant NEFQVFs and their conjugate priors do not always exist, and are sometimes infinite. So, the Hilbert space $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ does not have an orthonormal basis of polynomials for these joint distributions, and we cannot hope for the singular vectors of C to be orthonormal polynomials. Hence, we will establish three main results in this paper corresponding to the Poisson, binomial, and Gaussian NEFQVFs. These results are outlined in the ensuing subsections.

A. The Laguerre SVD

Recalling the setup in subsection I-B, let $\mathcal{X} = (0, \infty)$ and $\mathcal{Y} = \mathbb{N} \triangleq \{0, 1, 2, \dots\}$, and let λ be the Lebesgue measure and μ be the counting measure. For our first result, the channel conditional pmfs $\{P_{Y|X=x} \sim \mathsf{Poisson}(x) : x \in (0, \infty)\}$ are the NEFQVF of Poisson distributions:

$$\forall x \in (0, \infty), \quad \forall y \in \mathbb{N}, \quad P_{Y|X}(y|x) = \frac{x^y e^{-x}}{y!} \tag{18}$$

where $x \in (0, \infty)$ is the rate (or expectation) parameter of the Poisson distribution. We remark that the Poisson channel is a widely used model in optical communications where X represents the intensity of transmitted light and Y represents the number of photons hitting a direct-detection receiver; see [20] and the references therein. The corresponding conjugate prior family consists of gamma distributions, and we assume that the source pdf is $P_X \sim \mathsf{gamma}(\alpha, \beta)$:

$$\forall x \in (0, \infty), \quad P_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{19}$$

where $\alpha \in (0, \infty)$ is the shape parameter, $\beta \in (0, \infty)$ is the rate parameter, and the gamma function, $\Gamma : (0, \infty) \to \mathbb{R}$, is:

$$\Gamma(z) \triangleq \int_0^\infty x^{z-1} e^{-x}\, dx. \tag{20}$$

Note that when $\alpha \in \mathbb{Z}^+ \triangleq \{1, 2, 3, \dots\}$, the gamma distribution specializes to an Erlang distribution. The posterior pdfs $\{P_{X|Y=y} \sim \mathsf{gamma}(\alpha + y, \beta + 1) : y \in \mathbb{N}\}$ are also gamma distributions as we used a conjugate prior. Finally, the output marginal pmf $P_Y \sim \mathsf{negative\text{-}binomial}\big(\alpha, p = \frac{1}{\beta+1}\big)$ is a negative binomial distribution:

$$\forall y \in \mathbb{N}, \quad P_Y(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha) y!} \left(\frac{1}{\beta+1}\right)^y \left(\frac{\beta}{\beta+1}\right)^\alpha \tag{21}$$

where $p = \frac{1}{\beta+1} \in (0, 1)$ is the success probability parameter and $\alpha \in (0, \infty)$ is the number of failures parameter. When $\alpha \in \mathbb{Z}^+$, the negative binomial random variable is the sum of α independent geometric random variables, and models the number of successes in a Bernoulli process until α failures.

The Hilbert space $\mathcal{L}^2((0, \infty), \mathbb{P}_X)$ has an orthonormal basis of *generalized Laguerre polynomials*. In particular, the generalized Laguerre polynomial with degree $k \in \mathbb{N}$, denoted $L_k^{(\alpha,\beta)} : (0, \infty) \to \mathbb{R}$, is defined by the Rodrigues formula:

$$L_k^{(\alpha,\beta)}(x) \triangleq \sqrt{\frac{\Gamma(\alpha)}{\Gamma(k+\alpha)\, k!}}\; x^{1-\alpha} e^{\beta x}\, \frac{d^k}{dx^k}\Big(x^{k+\alpha-1} e^{-\beta x}\Big) \tag{22}$$

with the parameters $\alpha, \beta \in (0, \infty)$. These polynomials satisfy the orthogonality relation:

$$\forall j, k \in \mathbb{N}, \quad \mathbb{E}\Big[L_j^{(\alpha,\beta)}(X) L_k^{(\alpha,\beta)}(X)\Big] = \delta_{jk} \tag{23}$$

with respect to the gamma pdf, where δ_{jk} is the Kronecker delta function that equals 1 if $j = k$ and equals 0 otherwise.

The Hilbert space $\mathcal{L}^2(\mathbb{N}, \mathbb{P}_Y)$ has a unique (up to arbitrary sign changes) orthonormal polynomial basis of *Meixner polynomials*. The Meixner polynomial with degree $k \in \mathbb{N}$, denoted $M_k^{(s,p)} : \mathbb{N} \to \mathbb{R}$, is parametrized by $s \in (0, \infty)$ and

$p \in (0,1)$. These polynomials satisfy the orthogonality relation:

$$\sum_{y=0}^{\infty} M_j^{(s,p)}(y) M_k^{(s,p)}(y) \frac{\Gamma(s+y)}{\Gamma(s)y!} p^y (1-p)^s = \delta_{jk} \tag{24}$$

for every $j,k \in \mathbb{N}$, with respect to the negative binomial distribution with parameters $s \in (0,\infty)$ and $p \in (0,1)$. All our definitions of orthogonal polynomials are derived from [21]–[23], and we use to these sources in subsequent sections without tediously referring back to them.

The next theorem presents the orthogonal polynomial SVD of the conditional expectation operator C corresponding to the gamma source and Poisson channel model.

Theorem 1 (Laguerre SVD): For the Poisson channel with gamma source, as presented in (18) *and* (19)*, the conditional expectation operator, $C : \mathcal{L}^2((0,\infty), \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{N}, \mathbb{P}_Y)$, has SVD:*

$$\forall k \in \mathbb{N}, \quad C\left(L_k^{(\alpha,\beta)}\right) = \sigma_k M_k^{\left(\alpha, \frac{1}{\beta+1}\right)}$$

where $\{\sigma_k \in (0,1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\to\infty} \sigma_k = 0$.

The $\alpha = 1$ case of Theorem 1, where P_X is an exponential distribution and P_Y is a geometric distribution, is also presented in [12], and the corresponding singular values are calculated to be:

$$\forall k \in \mathbb{N}, \quad \sigma_k = \left(\frac{1}{\beta+1}\right)^{\frac{k}{2}}. \tag{25}$$

Note that when $\alpha = 1$, the right singular vectors of C are known as *Laguerre polynomials*. Although the left singular vectors of C are Meixner polynomials, we refer to this result as the "Laguerre SVD" because Meixner polynomials behave like discrete Laguerre polynomials. Indeed, the negative binomial distribution is the discrete analog of the gamma distribution (much like the geometric distribution is the discrete analog of the exponential distribution).

B. The Jacobi SVD

For our second result, let $\mathcal{X} = (0,1)$ and $\mathcal{Y} = [n] \triangleq \{0,\dots,n\}$, and let λ be the Lebesgue measure and μ be the counting measure in subsection I-B. The channel conditional pmfs $\{P_{Y|X=x} \sim \mathsf{binomial}(n,x) : x \in (0,1)\}$ are the NEFQVF of binomial distributions:

$$\forall x \in (0,1),\ \forall y \in [n],\ P_{Y|X}(y|x) = \binom{n}{y} x^y (1-x)^{n-y} \tag{26}$$

where $x \in (0,1)$ is the success probability parameter and $n \in \mathbb{Z}^+$ is the fixed number of Bernoulli trials of the binomial distribution. The capacity of this "biased coin channel" model has been studied in the literature [24]. The corresponding conjugate prior family consists of beta distributions, and we assume that the source pdf is $P_X \sim \mathsf{beta}(\alpha,\beta)$:

$$\forall x \in (0,1), \quad P_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \tag{27}$$

where $\alpha,\beta \in (0,\infty)$ are shape parameters, and the beta function, $\mathrm{B} : (0,\infty)^2 \to \mathbb{R}$, is defined as:

$$\mathrm{B}(z_1,z_2) \triangleq \int_0^1 x^{z_1-1}(1-x)^{z_2-1}\,dx = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}. \tag{28}$$

The posterior pdfs $\{P_{X|Y=y} \sim \mathsf{beta}(\alpha+y, \beta+n-y) : y \in [n]\}$ are also beta distributions since we used a conjugate prior. Lastly, the output marginal pmf $P_Y \sim \mathsf{beta\text{-}binomial}(n,\alpha,\beta)$ is the beta-binomial distribution:

$$\forall y \in [n], \quad P_Y(y) = \binom{n}{y} \frac{\mathrm{B}(\alpha+y, \beta+n-y)}{\mathrm{B}(\alpha,\beta)} \tag{29}$$

with parameters $n \in \mathbb{Z}^+$, $\alpha \in (0,\infty)$, and $\beta \in (0,\infty)$.

The Hilbert space $\mathcal{L}^2((0,1),\mathbb{P}_X)$ has an orthonormal basis of *Jacobi polynomials*. In particular, the Jacobi polynomial with degree $k \in \mathbb{N}$, denoted $J_k^{(\alpha,\beta)} : (0,1) \to \mathbb{R}$, is defined by the Rodrigues formula:

$$J_k^{(\alpha,\beta)}(x) \triangleq x^{1-\alpha}(1-x)^{1-\beta} \frac{d^k}{dx^k}\left(x^{k+\alpha-1}(1-x)^{k+\beta-1}\right) \cdot (-1)^k \sqrt{\frac{(2k+\alpha+\beta-1)\mathrm{B}(\alpha,\beta)\Gamma(k+\alpha+\beta-1)}{\Gamma(k+\alpha)\Gamma(k+\beta)\,k!}} \tag{30}$$

with the parameters $\alpha,\beta \in (0,\infty)$. These polynomials satisfy the orthogonality relation:

$$\forall j,k \in \mathbb{N}, \quad \mathbb{E}\left[J_j^{(\alpha,\beta)}(X) J_k^{(\alpha,\beta)}(X)\right] = \delta_{jk} \tag{31}$$

with respect to the beta distribution. They also generalize several other orthogonal polynomial families such as the Legendre and Chebyshev polynomials.

The Hilbert space $\mathcal{L}^2([n],\mathbb{P}_Y)$ has a unique (up to arbitrary sign changes) orthonormal polynomial basis of *Hahn polynomials*. The Hahn polynomial with degree $k \in [n]$, denoted $Q_k^{(\alpha,\beta)} : [n] \to \mathbb{R}$, is parametrized by $\alpha,\beta \in (0,\infty)$. These polynomials satisfy the orthogonality relation:

$$\forall j,k \in [n], \quad \mathbb{E}\left[Q_j^{(\alpha,\beta)}(Y) Q_k^{(\alpha,\beta)}(Y)\right] = \delta_{jk} \tag{32}$$

with respect to the beta-binomial distribution. The Hahn polynomials also generalize several other families of orthogonal polynomials in the limit, including the Jacobi and Meixner polynomials defined earlier, and the Krawtchouk and Charlier polynomials which are orthogonal with respect to the binomial and Poisson distributions, respectively [21].

The following theorem presents the orthogonal polynomial SVD of the conditional expectation operator C corresponding to the beta source and binomial channel model.

Theorem 2 (Jacobi SVD): For the binomial channel with beta source, as presented in (26) *and* (27)*, the conditional expectation operator, $C : \mathcal{L}^2((0,1),\mathbb{P}_X) \to \mathcal{L}^2([n],\mathbb{P}_Y)$, has SVD:*

$$\forall k \in [n],\ C\left(J_k^{(\alpha,\beta)}\right) = \sigma_k Q_k^{(\alpha,\beta)}$$
$$\forall k \in \mathbb{N}\backslash[n],\ C\left(J_k^{(\alpha,\beta)}\right) = \mathbf{0}$$

where $\{\sigma_k \in (0,1] : k \in [n]\}$ are the singular values such that $\sigma_0 = 1$.

When $\alpha = \beta = 1$, P_X is the uniform pdf and P_Y is the uniform pmf. The corresponding orthonormal polynomials

are known as *Legendre polynomials* and *discrete Chebyshev or Gram polynomials* respectively, and are analogs of each other. For this reason, and the fact that Jacobi polynomials can be obtained as limits of Hahn polynomials, we refer to the SVD in Theorem 2 as the "Jacobi SVD."

C. The Hermite SVD

For our final result, let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, and let λ and μ be the Lebesgue measure in subsection I-B. The channel conditional pdfs $\{P_{Y|X=x} \sim \mathcal{N}(x, \nu) : x \in \mathbb{R}\}$ are the NEFQVF of Gaussian distributions:

$$\forall x, y \in \mathbb{R}, \quad P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) \tag{33}$$

where $x \in \mathbb{R}$ is the expectation parameter of the Gaussian distribution and $\nu \in (0, \infty)$ is some fixed variance. We can construe (33) as the well-known single letter additive white Gaussian noise (AWGN) channel:

$$Y = X + W, \quad X \perp\!\!\!\perp W \sim \mathcal{N}(0, \nu) \tag{34}$$

where the input X is independent of the Gaussian noise W. The corresponding conjugate prior family consists of Gaussian distributions, and we assume that the source pdf is $P_X \sim \mathcal{N}(r, p)$:

$$\forall x \in \mathbb{R}, \quad P_X(x) = \frac{1}{\sqrt{2\pi p}} \exp\left(-\frac{(x-r)^2}{2p}\right) \tag{35}$$

where $r \in \mathbb{R}$ is the expectation parameter, and $p \in (0, \infty)$ is the variance parameter. The posterior pdfs $\{P_{X|Y=y} \sim \mathcal{N}((py+\nu r)/(p+\nu), p\nu/(p+\nu)) : y \in \mathbb{R}\}$ are also Gaussian distributions as we used a conjugate prior. Finally, the output marginal pdf $P_Y \sim \mathcal{N}(r, p+\nu)$ is also a Gaussian distribution.

The Hilbert spaces $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_X)$ and $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ have orthonormal bases of *Hermite polynomials*. In particular, the Hermite polynomial with degree $k \in \mathbb{N}$, denoted $H_k^{(r,\tau)} : \mathbb{R} \to \mathbb{R}$, is defined by the Rodrigues formula:

$$H_k^{(r,\tau)}(x) \triangleq \sqrt{\frac{\tau^k}{k!}} (-1)^k e^{\frac{(x-r)^2}{2\tau}} \frac{d^k}{dx^k}\left(e^{-\frac{(x-r)^2}{2\tau}}\right) \tag{36}$$

with the parameters $r \in \mathbb{R}$ and $\tau \in (0, \infty)$. These polynomials satisfy the orthogonality relation:

$$\forall j, k \in \mathbb{N}, \quad \int_{-\infty}^{\infty} H_j^{(r,\tau)}(x) H_k^{(r,\tau)}(x) \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(x-r)^2}{2\tau}} dx = \delta_{jk} \tag{37}$$

with respect to the Gaussian distribution $\mathcal{N}(r, \tau)$.

The ensuing theorem presents the orthogonal polynomial SVD of the conditional expectation operator C corresponding to the Gaussian source-channel model.

Theorem 3 (Hermite SVD): For the Gaussian channel with Gaussian source, as presented in (33) *and* (35)*, the conditional expectation operator, $C : \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$, has SVD:*

$$\forall k \in \mathbb{N}, \quad C\left(H_k^{(r,p)}\right) = \sigma_k H_k^{(r,p+\nu)}$$

where $\{\sigma_k \in (0, 1] : k \in \mathbb{N}\}$ are the singular values such that $\sigma_0 = 1$ and $\lim_{k\to\infty} \sigma_k = 0$.

This result is known in the literature. For example, Theorem 3 was derived in [3] for the $r = 0$ case, where the authors also computed the singular values to be:

$$\forall k \in \mathbb{N}, \quad \sigma_k = \left(\frac{p}{p+\nu}\right)^{\frac{k}{2}}. \tag{38}$$

Furthermore, the authors of [3] also remarked upon the possible relation between Theorem 3 and the classical theory of the Ornstein-Uhlenbeck process. We include Theorem 3 here for completeness, and provide an alternative proof of it. Theorems 1 and 2 generalize Theorem 3 by establishing a "nice" class of source-channel models whose conditional expectation operators have orthogonal polynomial singular vectors.

D. Related Results in the Literature

The general problem of analyzing when the singular vectors (or eigenvectors) of certain linear operators are orthogonal polynomials has been widely studied in mathematics. Comprehensive resources on the general theory of orthogonal polynomials include [21]–[23]. In particular, it is well-known that the classical orthogonal polynomials (Hermite, Laguerre, and Jacobi) arise as eigenfunctions of certain second order (*Sturm-Liouville* type of) differential operators. Both [23] and [25] meticulously expound various relationships between orthogonal polynomials and differential or integral linear operators.

In the setting of probability theory, there are deep ties between orthogonal polynomials and certain *Markov semigroups*. Under regularity conditions, the conditional expectation operators of a semigroup are completely characterized by an *infinitesimal generator*, because they form the unique solution to the *heat equation* defined by their generator (due to the Hille-Yosida theorem) [26]. When the generator is a *diffusion operator* (which is a kind of second order differential operator), the orthogonal polynomials with respect to the invariant measure of the semigroup turn out to be eigenfunctions of the generator, or equivalently, the conditional expectation operators. Moreover, there are only three families of orthogonal polynomials (up to scaling and translations) that are eigenfunctions of diffusion operators: the Hermite, Laguerre, and Jacobi polynomials [27]. The three corresponding diffusion operators are precisely the aforementioned second order differential operators with classical orthogonal polynomial eigenfunctions. In particular, the Markov semigroup in the Hermite case is the well-known *Ornstein-Uhlenbeck semigroup*. We refer readers to [26], [27], and the references therein for detailed expositions of these ideas.

Our results are closer in spirit to a line of work in statistics initiated by Lancaster [28], [29]. Given marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$, and sequences of orthonormal functions, $\{f_j \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)\}$ and $\{g_k \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)\}$, a bivariate distribution $\mathbb{P}_{X,Y}$ is called a *Lancaster distribution* (with respect to $\mathbb{P}_X$, $\mathbb{P}_Y$, $\{f_j\}$, and $\{g_k\}$) if for every j, k:

$$\mathbb{E}\left[f_j(X) g_k(Y)\right] = \sigma_k \delta_{jk} \tag{39}$$

for some *Lancaster sequence* of non-negative correlations $\{\sigma_k\}$. In [28], Lancaster proved that if $\mathbb{P}_{X,Y}$ is

absolutely continuous with respect to the product distribution $\mathbb{P}_X \times \mathbb{P}_Y$, and has finite "mean square contingency" (i.e. the χ^2-divergence $\chi^2(\mathbb{P}_{X,Y}||\mathbb{P}_X \times \mathbb{P}_Y) = \mathbb{E}_{\mathbb{P}_X \times \mathbb{P}_Y}\left[\tau(X,Y)^2\right] - 1$ is finite), then there exist orthonormal bases, $\{f_j \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)\}$ and $\{g_k \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)\}$, such that $\mathbb{P}_{X,Y}$ is a Lancaster distribution, and the following expansion holds:

$$\tau(x,y) = \sum_k \sigma_k f_k(x) g_k(y) \tag{40}$$

where $\tau(x,y)$ denotes the Radon-Nikodym derivative of $\mathbb{P}_{X,Y}$ with respect to $\mathbb{P}_X \times \mathbb{P}_Y$. It is straightforward to see that an expansion of the form (40) captures the SVD structure of the conditional expectation operators associated with $\mathbb{P}_{X,Y}$. Explicit expansions of the form (40) in terms of orthogonal polynomials have since been derived for various bivariate distributions. For instance, orthogonal polynomial expansions for bivariate distributions that are generated additively from three independent NEFQVF random variables were established in [30]. We refer readers to [30], [31], and the references therein for further details on such classical work. More contemporary results on Lancaster distributions are presented in in [32], [33], and the references therein. As explained in [33], one direction of research is to find the extremal Lancaster sequences corresponding to the extremal points of the compact, convex set of Lancaster distributions corresponding to certain marginal distributions and their orthogonal polynomial sequences.

In contrast to the aforementioned classical examples of orthogonal polynomial eigenfunctions, the conditional expectation operators that we derive SVDs for are defined by NEFQVF channels and conjugate prior sources. As we mentioned earlier, the Hermite SVD result in Theorem 3 can be related to results on the Ornstein-Uhlenbeck semigroup since a Gaussian NEFQVF has a Gaussian conjugate prior family. However, we emphasize that the Laguerre and Jacobi SVDs in Theorems 1 and 2 are distinct from classical results (in the contexts of differential equations, integral equations, Markov semigroups, or Lancaster distributions). To our knowledge, these classical results do not analyze the setting of NEFQVF channels and conjugate prior sources. On the other hand, we would like to acknowledge that results similar to ours on spectral decompositions of Markov chains have been independently derived in [34] to analyze the convergence rate of Gibbs sampling.

Finally, it is worth mentioning that although we refer to Morris' unified theory of NEFQVFs in [9] and [10] in section I, the importance of NEFQVFs was recognized much earlier by Meixner. Indeed, Meixner characterized the orthogonal polynomial families corresponding to NEFQVFs as precisely those that have generating functions with a certain tractable form in [35]. (Since we only use orthogonal polynomials corresponding to $\mathbb{P}_X$ and $\mathbb{P}_Y$ rather than those corresponding to NEFQVFs, Meixner's results are not directly of relevance to us.)

III. Proofs of Main Results

In this section, we will prove our main results under the conditions stated in subsection I-B. To this end, we will first derive an auxiliary result that provides simple necessary and sufficient conditions for conditional expectation operators to have orthonormal polynomial singular vectors. We refer readers to [36]–[38] for the relevant functional analysis background.

Recall that we are given the Hilbert space $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ with dimension:

$$\dim\left(\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)\right) = |\mathcal{X}| \in \mathbb{Z}^+ \cup \{+\infty\} \tag{41}$$

i.e. $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ is infinite dimensional when $|\mathcal{X}| = +\infty$, and finite dimensional when $|\mathcal{X}| < +\infty$. Since $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ is separable, it equivalently has a countable complete orthonormal (Schauder) basis. We will assume that $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ has a unique (up to arbitrary sign changes) orthonormal basis of polynomials $\left\{p_k \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : k < |\mathcal{X}|\right\}$, where p_k is an orthonormal polynomial with degree $k \in \mathbb{N}$. Typically, such orthonormal polynomials can be constructed by applying the Gram-Schmidt algorithm to the monomials $\left\{1, x, x^2, \ldots\right\}$. Note that the finiteness of the moment generating function (MGF) of X on an interval containing zero guarantees the existence of an orthonormal polynomial basis of $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ (since it ensures that all moments of X exist and are finite), and the positive definiteness of $\langle \cdot, \cdot \rangle_{\mathbb{P}_X}$ with respect to the subspace of all polynomials guarantees the uniqueness of this basis. Furthermore, this discussion holds mutatis mutandis for the Hilbert space $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, and we also assume that it has a unique orthonormal basis of polynomials $\left\{q_k \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) : k < |\mathcal{Y}|\right\}$, where q_k is an orthonormal polynomial with degree k. The next definition presents a pertinent property of bounded linear operators between the Hilbert spaces $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$.

Definition 4 (Degree Preservation): A bounded linear operator $T : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is *degree preserving* if for any polynomial $p \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ with degree $k \in \mathbb{N}$, $T(p) \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is also a polynomial with degree at most k. T is *strictly degree preserving* if:

- Case $|\mathcal{X}| \leq |\mathcal{Y}|$: For any polynomial $p \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ with degree $k < |\mathcal{X}|$, $T(p) \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is also a polynomial with degree exactly k.
- Case $|\mathcal{X}| > |\mathcal{Y}|$ ($\Rightarrow |\mathcal{Y}| < +\infty$): For any polynomial $p \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ with degree $k < |\mathcal{Y}|$, $T(p) \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is also a polynomial with degree exactly k, and for any polynomial $p \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ with degree $|\mathcal{Y}| \leq k < |\mathcal{X}|$, $T(p) \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is also a polynomial with degree at most $|\mathcal{Y}| - 1$.

In Definition 4, we use the convention that $\infty \leq \infty$ is true, and $\infty < \infty$ is false. We also remark that when $\mathcal{X} = \mathcal{Y}$, this definition implies that polynomials form an invariant subspace of a degree preserving operator T. The next proposition presents our auxiliary result using this definition.

Proposition 2 (Orthogonal Polynomial SVD):

Let $T : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ be a compact linear operator, and $T^* : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ be its unique adjoint operator. Then, T and T^* are strictly degree preserving if and only if T has SVD:

$$\forall k < \min\{|\mathcal{X}|, |\mathcal{Y}|\},\ T(p_k) = \beta_k q_k$$
$$|\mathcal{X}| > |\mathcal{Y}| \ \Rightarrow\ \forall |\mathcal{Y}| \leq k < |\mathcal{X}|,\ T(p_k) = \mathbf{0}$$

where $\{\beta_k \in (0,\infty) : k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ *are singular values such that* $\lim_{k\to\infty} \beta_k = 0$ *when* $\min\{|\mathcal{X}|, |\mathcal{Y}|\} = +\infty$.

Proof: We first prove the forward direction. Since T is a compact linear operator, its adjoint T^* is also compact by Schauder's theorem [36], [37]. Hence, the (Gramian) operator T^*T is self-adjoint, positive, and compact since the composition of compact operators is compact. Moreover, since T and T^* are strictly degree preserving, T^*T is degree preserving; in fact, T^*T is strictly degree preserving when $|\mathcal{X}| \leq |\mathcal{Y}|$.

Using the spectral theorem for compact self-adjoint operators [36], T^*T has a countable orthonormal eigenbasis $\{r_i \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : i \in \mathbb{N}, i < |\mathcal{X}|\}$:

$$\forall i < |\mathcal{X}|, \quad T^*T(r_i) = \alpha_i r_i$$

where $\{\alpha_i \in \mathbb{R}^+ : i < |\mathcal{X}|\}$ are the non-negative eigenvalues (since T^*T is positive) such that $\lim_{i\to\infty} \alpha_i = 0$ when $|\mathcal{X}| = +\infty$. We will prove by strong induction that these eigenfunctions are orthonormal polynomials.

The first eigenfunction of T^*T must be the constant function $r_0 = p_0 = \mathbf{1}_{\mathcal{X}}$ since T^*T is degree preserving. Assume that the first $k+1$ eigenfunctions are orthonormal polynomials: $r_i = p_i$ for $i \in \{0, \ldots, k\}$ (inductive hypothesis). Then, since p_{k+1} is orthogonal to $\mathsf{span}(r_0, \ldots, r_k) = \mathsf{span}(p_0, \ldots, p_k)$, we have:

$$p_{k+1} = \sum_{k+1 \leq j < |\mathcal{X}|} \langle p_{k+1}, r_j \rangle_{\mathbb{P}_X} r_j.$$

When $|\mathcal{X}| = +\infty$, this equality holds in the sense that the partial sums converge to p_{k+1} in $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$-norm. Applying T^*T to both sides and using the continuity (or equivalently, the boundedness) of T^*T, we get:

$$T^*T(p_{k+1}) = \sum_{k+1 \leq j < |\mathcal{X}|} \alpha_j \langle p_{k+1}, r_j \rangle_{\mathbb{P}_X} r_j$$

which also holds in the $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$-norm sense when $|\mathcal{X}| = +\infty$. Hence, $T^*T(p_{k+1})$ is orthogonal to $\mathsf{span}(p_0, \ldots, p_k)$ using the continuity of the inner product, and it is a polynomial with degree at most $k+1$ as T^*T is degree preserving. This implies that:

$$T^*T(p_{k+1}) = \alpha_{k+1} p_{k+1}$$

where α_{k+1} is possibly zero, which means that $r_{k+1} = p_{k+1}$ (without loss of generality). By strong induction, $\{p_k \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : k < |\mathcal{X}|\}$ are the eigenfunctions of T^*T:

$$\forall k < |\mathcal{X}|, \quad T^*T(p_k) = \alpha_k p_k \tag{42}$$

where for all $k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, $\alpha_k > 0$ because both T and T^* do not reduce the degrees of input polynomials with degrees less than $\min\{|\mathcal{X}|, |\mathcal{Y}|\}$.

Now observe (by definition of the adjoint operator) that:

$$\begin{aligned} \forall j, k < |\mathcal{X}|, \quad \langle T(p_j), T(p_k) \rangle_{\mathbb{P}_Y} &= \langle p_j, T^*T(p_k) \rangle_{\mathbb{P}_X} \\ &= \alpha_k \langle p_j, p_k \rangle_{\mathbb{P}_X} \\ &= \alpha_k \delta_{jk}. \end{aligned}$$

This means that $\{T(p_k) : k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ are scaled versions of the orthonormal polynomials in $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ since T is strictly degree preserving:

$$\forall k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}, \quad T(p_k) = \sqrt{\alpha_k} q_k$$

where the sign of each orthonormal polynomial q_k is chosen to keep $\sqrt{\alpha_k} > 0$. On the other hand, if $|\mathcal{X}| > |\mathcal{Y}|$, then for any $|\mathcal{Y}| \leq k < |\mathcal{X}|$, $T(p_k)$ is a polynomial with degree at most $|\mathcal{Y}| - 1$ and is orthogonal to every q_j with $j < |\mathcal{Y}|$. This implies that $T(p_k) = \mathbf{0}$ as $\{q_j \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) : j < |\mathcal{Y}|\}$ is an orthonormal basis of $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$. Moreover, $\alpha_k = 0$ for every $|\mathcal{Y}| \leq k < |\mathcal{X}|$. Therefore, we have:

$$\begin{aligned} &\forall k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}, \quad T(p_k) = \sqrt{\alpha_k} q_k \\ |\mathcal{X}| > |\mathcal{Y}| \;\Rightarrow\; &\forall |\mathcal{Y}| \leq k < |\mathcal{X}|, \quad T(p_k) = \mathbf{0} \end{aligned}$$

which is the SVD of T with singular values, $\beta_k = \sqrt{\alpha_k} > 0$ for every $k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, that satisfy $\lim_{k\to\infty} \beta_k = 0$ when $\min\{|\mathcal{X}|, |\mathcal{Y}|\} = +\infty$ (since $\lim_{k\to\infty} \alpha_k = 0$). This completes the proof of the forward direction.

To prove the converse direction, notice that T having SVD:

$$\begin{aligned} &\forall k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}, \quad T(p_k) = \beta_k q_k \\ |\mathcal{X}| > |\mathcal{Y}| \;\Rightarrow\; &\forall |\mathcal{Y}| \leq k < |\mathcal{X}|, \quad T(p_k) = \mathbf{0} \end{aligned}$$

implies that T^* has SVD:

$$\begin{aligned} &\forall k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}, \quad T^*(q_k) = \beta_k p_k \\ |\mathcal{Y}| > |\mathcal{X}| \;\Rightarrow\; &\forall |\mathcal{X}| \leq k < |\mathcal{Y}|, \quad T^*(q_k) = \mathbf{0}. \end{aligned}$$

This is an exercise in functional analysis. Since $\beta_k > 0$ for any $k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, and any polynomial can be decomposed into a weighted sum of orthonormal polynomials, these SVDs imply that T and T^* are strictly degree preserving. This completes the proof. ∎

We briefly make some remarks regarding Proposition 2. Firstly, the result continues to hold if $\mathbb{P}_X$ or $\mathbb{P}_Y$ are relaxed to be non-probability measures that have unique orthonormal polynomial bases. Secondly, the singular values $\{\beta_k \in (0,\infty) : k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ of T must be computed on a case by case basis if desired. Thirdly, as mentioned in the proof, the signs of $\{p_k \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) : k < |\mathcal{X}|\}$ and $\{q_k \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) : k < |\mathcal{Y}|\}$ are chosen to ensure that the singular values are non-negative. This convention also applies to Theorems 1, 2, 3, and Lemma 1 below. Fourthly, it is worth considering Proposition 2 when $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ are finite dimensional, and isomorphic to $\mathbb{R}^{|\mathcal{X}|}$ and $\mathbb{R}^{|\mathcal{Y}|}$, respectively. In this scenario, T and T^* have finite rank, and are trivially compact operators that have SVDs. Moreover, every basis of a Euclidean space $\mathbb{R}^n$ ($n = |\mathcal{X}|$ or $n = |\mathcal{Y}|$) corresponds to a basis of polynomials, where each polynomial has degree at most $n-1$, by the unisolvence theorem. So, the singular vectors of T and T^* will always be polynomials. The non-trivial aspect of Proposition 2 in this finite dimensional setting is that T and T^* have orthonormal polynomial singular vector bases if and only if T and T^* are strictly degree preserving. Lastly, it is worth noting that although the SVD result in Proposition 2 requires strict degree preservation, the spectral decomposition result in (42) only requires degree preservation as the proof illustrates.

The ensuing lemma is a straightforward corollary of Proposition 2 specializing it for conditional expectation operators.

Lemma 1 (Conditional Moment Condition):

Suppose the conditional expectation operator $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ *is compact, and suppose* $|\mathcal{X}| \geq |\mathcal{Y}|$ *without loss of generality. Then, for every* $n \in \mathbb{N}$, $n < |\mathcal{Y}|$, $\mathbb{E}\left[Y^n|X\right]$ *is a polynomial in* X *with degree* n *and* $\mathbb{E}\left[X^n|Y\right]$ *is a polynomial in* Y *with degree* n, *and for every* $n \in \mathbb{N}$, $|\mathcal{Y}| \leq n < |\mathcal{X}|$, $\mathbb{E}\left[X^n|Y\right]$ *is a polynomial in* Y *with degree at most* $|\mathcal{Y}| - 1$ *if and only if* C *has SVD:*

$$\forall k < |\mathcal{Y}|, \quad C(p_k) = \beta_k q_k$$
$$\forall |\mathcal{Y}| \leq k < |\mathcal{X}|, \quad C(p_k) = \mathbf{0}$$

where $\{\beta_k \in (0, 1] : k < |\mathcal{Y}|\}$ *are singular values such that* $\beta_0 = 1$, *and* $\lim_{k\to\infty} \beta_k = 0$ *when* $|\mathcal{Y}| = +\infty$.

Proof: The conditional moment conditions of the lemma are equivalent to C and C^* being strictly degree preserving. The SVD of C then follows from Proposition 2, where as before, the signs of the orthonormal polynomials are selected to keep the singular values positive. (When $|\mathcal{X}| = |\mathcal{Y}|$, no value of k satisfies the second line of the SVD, and it is vacuously true.) Furthermore, $\beta_k \leq 1$ for all $k < |\mathcal{Y}|$ because $\|C\|_{\mathsf{op}} = 1$ by Proposition 1, and $\beta_0 = 1$ since $C(\mathbf{1}_{\mathcal{X}}) = \mathbf{1}_{\mathcal{Y}}$. ∎

Lemma 1 provides an easily testable equivalent condition for a conditional expectation operator to have an orthonormal polynomial SVD; it holds with natural modifications when $|\mathcal{X}| \leq |\mathcal{Y}|$. We must also verify that C is compact when using this lemma. A well-known sufficient condition that ensures that C (and C^*) are compact is the *Hilbert-Schmidt condition*:

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{P_{X,Y}^2(x, y)}{P_X(x) P_Y(y)} \, d\mu(y) \, d\lambda(x) < +\infty \tag{43}$$

where $P_{X,Y}$, P_X, and P_Y are the probability densities defined in subsection I-B. In functional analysis, this condition arises from the compactness of *Hilbert-Schmidt operators*, which are integral operators with square integrable kernels [36]. In statistics, it corresponds to the finite "mean square contingency" condition mentioned in subsection II-D; for example, it is mentioned after Assumption 5.2 in [2], and in the premise of Theorem 2 in [1]. In our ensuing proofs, we will not explicitly check for compactness of the operators for brevity.

A. Finite Alphabet Examples

Before proving our main results, we briefly provide two basic examples of polynomial SVDs in the finite alphabet case.

Example 1 (Uniform Source and Binary Symmetric Channel): Suppose $X \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right)$, and $Y \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right)$ is the output of passing X through a binary symmetric channel with crossover probability $\delta \in \left(0, \frac{1}{2}\right)$. The orthonormal polynomials in $\mathcal{L}^2\left(\{0, 1\}, \mathsf{Bernoulli}\left(\frac{1}{2}\right)\right)$ are $p_0 = (1, 1)$ and $p_1 = (1, -1)$, where $p_k = (p_k(0), p_k(1))$ for $k = 0, 1$. It is straightforward to directly verify that the SVD of C is:

$$C(p_0) = p_0 \quad \text{and} \quad C(p_1) = (1 - 2\delta) \, p_1. \tag{44}$$

So, C and C^* are strictly degree preserving by Lemma 1.

Example 2 (Uniform Source and Binary Erasure Channel): Suppose $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \mathcal{X} \cup \{\mathsf{e}\}$ (where e is the erasure symbol), $X \sim \mathsf{Bernoulli}\left(\frac{1}{2}\right)$, and Y is the output of passing X through a binary erasure channel with erasure probability $\epsilon \in (0, 1)$. In this case, the SVD of C is:

$$C(p_0) = g_0 \quad \text{and} \quad C(p_1) = \sqrt{1 - \epsilon} \, g_1 \tag{45}$$

where the right singular vectors are given in Example 1, and the left singular vectors are the orthonormal vectors $g_0 = (1, 1, 1)$ and $g_1 = \left(1/\sqrt{1-\epsilon}, -1/\sqrt{1-\epsilon}, 0\right)$ in $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, where $g_k = (g_k(0), g_k(1), g_k(\mathsf{e}))$ for $k = 0, 1$. Note that C has an orthonormal polynomial SVD if and only if g_1 is a linear polynomial, which is true if and only if $\mathsf{e} = \frac{1}{2}$. Therefore, when $\mathsf{e} \in \mathbb{R}$ and $\mathsf{e} \neq \frac{1}{2}$, C and C^* are not strictly degree preserving by Lemma 1 (although g_1 is a non-linear polynomial).

B. Proof of Theorem 1: Laguerre SVD

Proof: First notice that given $X = x \in (0, \infty)$, Y is Poisson distributed with rate x as shown in (18). This means that the *cumulants* of $P_{Y|X=x}$ are all equal to x. Since the nth moment $\mathbb{E}\left[Y^n|X = x\right]$ for $n \in \mathbb{N}$ is a polynomial in the first n cumulants with degree n [39], $\mathbb{E}\left[Y^n|X\right]$ is a polynomial in X with degree n for every $n \in \mathbb{N}$. (Note that this can also be proved directly by using induction on the derivatives of the MGF of a Poisson random variable.)

Next, we prove that the moments of a gamma distribution $P_X \sim \mathsf{gamma}(\alpha, \beta)$ with $\alpha, \beta \in (0, \infty)$, shown in (19), are polynomials in α with the same degree. The MGF of X is:

$$M_X(s) \triangleq \mathbb{E}\left[e^{sX}\right] = \begin{cases} \left(\frac{\beta}{\beta - s}\right)^{\alpha}, & s < \beta \\ +\infty, & s \geq \beta \end{cases}$$

and as $\beta > 0$, the MGF is finite on an open interval around $s = 0$. This means the moments of X are given by:

$$\begin{aligned} \mathbb{E}\left[X^n\right] &= \frac{d^n}{ds^n} M_X(s)\Big|_{s=0} \\ &= \frac{d^n}{ds^n} \left(\frac{\beta}{\beta - s}\right)^{\alpha}\Big|_{s=0} \\ &= \left(\frac{\beta}{\beta - s}\right)^{\alpha} \frac{1}{(\beta - s)^n} \prod_{i=0}^{n-1} (\alpha + i)\Big|_{s=0} \\ &= \frac{1}{\beta^n} \prod_{i=0}^{n-1} (\alpha + i) \end{aligned}$$

for every $n \in \mathbb{N}\backslash\{0\}$. Thus, for every $n \in \mathbb{N}$, $\mathbb{E}\left[X^n\right]$ is a polynomial in α with degree n.

As mentioned earlier in subsection II-A, the posterior pdfs $\{P_{X|Y=y} \sim \mathsf{gamma}(\alpha+y, \beta+1) : y \in \mathbb{N}\}$ are also gamma pdfs with updated parameters. Hence, for every $n \in \mathbb{N}$, $\mathbb{E}\left[X^n|Y\right]$ is a polynomial in Y with degree n. Applying Lemma 1 completes the proof. ∎

C. Proof of Theorem 2: Jacobi SVD

Proof: First observe that given $X = x \in (0, 1)$, $P_{Y|X=x} \sim \mathsf{binomial}(n, x)$, which means that $Y = Z_1 + \cdots + Z_n$ where $Z_1, \ldots, Z_n$ are conditionally i.i.d. $\mathsf{Bernoulli}(x)$ random

variables (i.e. $\mathbb{P}(Z_i = 1) = x$ and $\mathbb{P}(Z_i = 0) = 1 - x$ for $i = 1, \ldots, n$). Hence, we have for any $m \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}\left[Y^m | X = x\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i\right)^m \middle| X = x\right] \\ &= \sum_{\substack{0 \le k_1, \ldots, k_n \le m \\ k_1 + \cdots + k_n = m}} \frac{m!}{k_1! \cdots k_n!} \prod_{i=1}^{n} \mathbb{E}\left[Z_i^{k_i} \middle| X = x\right] \\ &= \sum_{\substack{0 \le k_1, \ldots, k_n \le m \\ k_1 + \cdots + k_n = m}} \frac{m!}{k_1! \cdots k_n!} x^{N(k_1, \ldots, k_n)}\end{aligned}$$

where the second equality follows from the multinomial theorem, the third equality follows from the fact that the moments of the Bernoulli random variables are $\mathbb{E}\left[Z_i^0 | X = x\right] = 1$ and for every $m \in \mathbb{N}\backslash\{0\}$, $\mathbb{E}\left[Z_i^m | X = x\right] = x$, and $N(k_1, \ldots, k_n)$ denotes the number of non-zero k_i. Since $N(k_1, \ldots, k_n) \le \min\{m, n\}$ and $N(k_1, \ldots, k_n) = \min\{m, n\}$ for at least one of the terms, we have that for every $m \in [n]$, $\mathbb{E}\left[Y^m | X\right]$ is a polynomial in X with degree m.

Next, as mentioned in subsection II-B, we note that the posterior pdfs $\left\{P_{X|Y=y} \sim \mathsf{beta}(\alpha + y, \beta + n - y) : y \in [n]\right\}$ are also beta pdfs with updated parameters. For any fixed $Y = y \in [n]$ and any $m \in \mathbb{N}$, we have:

$$\begin{aligned}\mathbb{E}\left[X^m | Y = y\right] &= \int_{(0,1)} x^m \frac{x^{\alpha+y-1}(1-x)^{\beta+n-y-1}}{\mathrm{B}(\alpha + y, \beta + n - y)}\, d\lambda(x) \\ &= \frac{\mathrm{B}(\alpha + y + m, \beta + n - y)}{\mathrm{B}(\alpha + y, \beta + n - y)} \\ &= \frac{\Gamma(\alpha + y + m)\,\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + y)\,\Gamma(\alpha + \beta + n + m)} \\ &= \frac{\prod_{k=0}^{m-1} (\alpha + y + k)}{\prod_{k=0}^{m-1} (\alpha + \beta + n + k)} \qquad (46)\end{aligned}$$

where the first equality uses (27) with the updated parameters. Therefore, for every $m \in [n]$, $\mathbb{E}\left[X^m | Y\right]$ is a polynomial in Y with degree m, and for every $m \in \mathbb{N}\backslash[n]$, $\mathbb{E}\left[X^m | Y\right]$ is a polynomial in Y with degree at most n. The latter deduction seems counter-intuitive in light of (46), which seems to suggest that $\mathbb{E}\left[X^m | Y\right]$ is always a polynomial in Y with degree m. However, since the function $y \mapsto \mathbb{E}\left[X^m | Y = y\right]$ is supported on a set of size $n + 1$, it can only be uniquely represented as a polynomial with degree at most n by the unisolvence theorem.

Finally, employing Lemma 1 completes the proof. ∎

D. Proof of Theorem 3: Hermite SVD

As mentioned earlier, Theorem 3 was proved in [3] using the Appell sequence recurrence relation of Hermite polynomials. We now provide another proof using Lemma 1 here. Our proof uses the following lemma, cf. the line below equation (51) in [40, Ch. 7].

Lemma 2 (Translation Invariant Kernels): Fix $u, v \in \mathbb{R}\backslash\{0\}$, and a Borel measurable λ-integrable function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} \phi\, d\lambda = 1$, where λ is the Lebesgue measure. If $T : \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X) \to \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ is a bounded integral operator with translation invariant kernel *ϕ:*

$$\forall f \in \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X), \quad (T(f))(y) = \int_{\mathbb{R}} \phi(uy + vx) f(x)\, d\lambda(x),$$

then T is strictly degree preserving.

Proof: We include a proof of this known result in Appendix C for completeness. ∎

Note that $u = 1$ and $v = -1$ corresponds to a *difference kernel* setting where T represents *convolution* with the function ϕ. We also remark that although $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_X)$ and $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$ are the Hilbert spaces defined in subsection II-C, Lemma 2 also holds for other (appropriately generalized) Hilbert spaces.

Proof of Theorem 3: Observe that when $P_{X,Y}$ is defined by (33) and (35), both C and C^* are integral operators with translation invariant kernels satisfying the conditions of Lemma 2. Indeed, for any $f \in \mathcal{L}^2(\mathbb{R}, \mathbb{P}_X)$ and any $g \in \mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$:

$$\begin{aligned}(C(f))(y) &= \int_{\mathbb{R}} f(x) P_{X|Y}(x|y)\, d\lambda(x) \\ &= \int_{\mathbb{R}} \frac{f(x)}{\sqrt{2\pi\left(\frac{pv}{p+v}\right)}} \exp\left(-\frac{\left(x - \frac{py+vr}{p+v}\right)^2}{2\left(\frac{pv}{p+v}\right)}\right) d\lambda(x), \\ \left(C^*(g)\right)(x) &= \int_{\mathbb{R}} g(y) P_{Y|X}(y|x)\, d\lambda(y) \\ &= \int_{\mathbb{R}} \frac{g(y)}{\sqrt{2\pi v}} \exp\left(-\frac{(y-x)^2}{2v}\right) d\lambda(y).\end{aligned}$$

Hence, C and C^* are strictly degree preserving by Lemma 2. Finally, applying Lemma 1 completes the proof. ∎

IV. Conclusion

In this paper, we first illustrated the utility of SVDs of conditional expectation operators by citing examples from the literature such as maximal correlation functions (which are themselves singular vectors of conditional expectation operators) and linear information coupling problems (which are solved by SVDs of conditional expectation operators). We then proved that conditional expectation operators corresponding to NEFQVF channels and conjugate prior sources, where all marginal moments exist and are finite, have orthonormal polynomial SVDs. In particular, the Gaussian source and Gaussian channel produce Hermite polynomial singular vectors, the gamma source and Poisson channel produce generalized Laguerre and Meixner polynomial singular vectors, and the beta source and binomial channel produce Jacobi and Hahn polynomial singular vectors. To establish these results, we verified that the corresponding conditional expectation operators and their adjoint operators are strictly degree preserving, which we showed is equivalent to having orthonormal polynomial SVDs. This equivalence between strict degree preservation and orthonormal polynomial SVDs may also be useful in future for deriving orthonormal polynomial SVDs for other source-channel models.

APPENDIX A
PROOF OF PROPOSITION 1

Proof: We prove this for completeness. The maps C and C^* are clearly linear since they are defined using expectations. To show that they are bounded, it suffices to prove that they have operator norms equal to unity (and this also verifies that their codomains are indeed Hilbert spaces). We now prove that:

$$\|C\|_{\mathsf{op}} \triangleq \sup_{h \in \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)\setminus\{\mathbf{0}\}} \frac{\|C(h)\|_{\mathbb{P}_Y}}{\|h\|_{\mathbb{P}_X}} = 1$$

where $\mathbf{0}$ denotes the zero function. Let $\mathbf{1}_S$ denote the everywhere unity function on the set $S \subseteq \mathbb{R}$: $\mathbf{1}_S : S \to \mathbb{R}$, $\mathbf{1}_S(x) = 1$. Observe that $\mathbf{1}_{\mathcal{X}} \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ (because $\|\mathbf{1}_{\mathcal{X}}\|^2_{\mathbb{P}_X} = 1$), and using (8), $C(\mathbf{1}_{\mathcal{X}}) = \mathbf{1}_{\mathcal{Y}} \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ (because $\|\mathbf{1}_{\mathcal{Y}}\|^2_{\mathbb{P}_Y} = 1$). Hence, $\|C\|_{\mathsf{op}} \geq 1$. Now, for any $h \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, we have:

$$\|C(h)\|^2_{\mathbb{P}_Y} = \mathbb{E}\left[\mathbb{E}\left[h(X)\middle|Y\right]^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[h^2(X)\middle|Y\right]\right] = \|h\|^2_{\mathbb{P}_X}$$

using (8), conditional Jensen's inequality, and the tower property and (7), respectively. Thus, $\|C\|_{\mathsf{op}} = 1$ and $\|C^*\|_{\mathsf{op}} = 1$, where the latter follows from an analogous argument.

Since C and C^* are bounded, they have unique adjoint operators using the Riesz representation theorem [36], [38]. Note that for every $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ and every $g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, we have:

$$\begin{aligned}\langle C(f), g\rangle_{\mathbb{P}_Y} &= \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right] g(Y)\right] \\ &= \mathbb{E}\left[f(X)\mathbb{E}\left[g(Y)|X\right]\right] = \left\langle f, C^*(g)\right\rangle_{\mathbb{P}_X}\end{aligned}$$

where the second equality follows from applying the tower property to $\mathbb{E}[f(X)g(Y)]$. Hence, C^* is the adjoint operator of C (as is implicitly suggested by the notation). ∎

APPENDIX B
LOCAL APPROXIMATIONS OF f-DIVERGENCES

In this appendix, we locally approximate f-divergences satisfying some regularity conditions, and specialize the approximation for Kullback-Leibler (KL) divergence. The f-divergences are a class of divergence measures over distributions that were independently introduced by Csiszár in [41], [42], and by Ali and Silvey in [43]. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Given two probability densities Q_X and P_X with respect to λ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ (recall the setup in subsection I-B), the *f-divergence* between Q_X and P_X is defined as:

$$D_f(Q_X||P_X) \triangleq \int_{\mathcal{X}} P_X(x) f\left(\frac{Q_X(x)}{P_X(x)}\right) d\lambda(x) \tag{47}$$

where we assume that $f(0) = \lim_{t\to 0^+} f(t)$, $0f\left(\frac{0}{0}\right) = 0$, and for all $r > 0$, $0f\left(\frac{r}{0}\right) = \lim_{s\to 0^+} sf\left(\frac{r}{s}\right) = r\lim_{s\to 0^+} sf\left(\frac{1}{s}\right)$ based on continuity arguments. With appropriate choices of the function $f : (0, \infty) \to \mathbb{R}$, f-divergences generalize many known divergence measures including KL divergence, total variation distance, χ^2-divergence, squared Hellinger distance, Jensen-Shannon divergence, and Jeffreys divergence [44], [45].

Proposition 3 (Local f-Divergence): Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$, $f(t)$ is thrice differentiable over some open interval around $t = 1$, $f'''(t)$ is locally bounded at $t = 1$, and $f''(1) > 0$. Suppose we are given a family of probability densities that are local perturbations of the reference probability density $P_X > 0$ λ-a.e.:

$$Q_\epsilon = P_X(\mathbf{1}_{\mathcal{X}} + \epsilon\,\phi)$$

where $\phi \in \mathcal{L}^\infty(\mathcal{X}, \mathbb{P}_X)$ is a fixed perturbation function such that $\mathbb{E}[\phi(X)] = 0$, and Q_ϵ is a valid probability density for all $\epsilon \neq 0$ of sufficiently small magnitude. Then, the f-divergence between Q_ϵ and P_X can be locally approximated as:

$$D_f(Q_\epsilon||P_X) = \frac{f''(1)}{2}\epsilon^2\,\mathbb{E}\Big[\phi(X)^2\Big] + o\big(\epsilon^2\big)$$

where $o\big(\epsilon^2\big)$ is the Bachmann-Landau asymptotic notation denoting a function which satisfies $\lim_{\epsilon\to 0} o\big(\epsilon^2\big)/\epsilon^2 = 0$.

Proof: First observe that using $f(1) = 0$, the thrice differentiability of $f(t)$ over some open interval around $t = 1$, and Taylor's theorem about the point $t = 1$, we have:

$$f(t) = f'(1)(t-1) + \frac{1}{2}f''(1)(t-1)^2 + \frac{1}{6}f'''(s)(t-1)^3$$

for every $t \in (0, \infty)$ sufficiently close to unity, and some corresponding $\min\{1, t\} \leq s \leq \max\{1, t\}$, where we have used the Lagrange form of the remainder. Using this Taylor approximation of $f(t)$, we have for every $x \in \widehat{\mathcal{X}}$:

$$\begin{aligned} f\left(\frac{Q_\epsilon(x)}{P_X(x)}\right) = \;& f'(1)\,\epsilon\,\phi(x) + \frac{f''(1)}{2}\,\epsilon^2\phi(x)^2 \\ &+ \frac{f'''(s(x))}{6}\,\epsilon^3\phi(x)^3 \end{aligned} \tag{48}$$

where $\widehat{\mathcal{X}} \subseteq \mathcal{X}$ is a Borel measurable set such that $P_X(x) > 0$ for all $x \in \widehat{\mathcal{X}}$ and $\lambda(\mathcal{X}\setminus\widehat{\mathcal{X}}) = 0$, and where for every $x \in \widehat{\mathcal{X}}$, $\min\{1, Q_\epsilon(x)/P_X(x)\} \leq s(x) \leq \max\{1, Q_\epsilon(x)/P_X(x)\}$. Let $h(x, \epsilon)$ denote the Lagrange remainder term:

$$h(x, \epsilon) \triangleq \frac{f'''(s(x))}{6}\,\epsilon^3\phi(x)^3. \tag{49}$$

Since $\phi \in \mathcal{L}^\infty(\mathcal{X}, \mathbb{P}_X)$, $\|\phi\|_\infty \triangleq \operatorname{ess\,sup}_{x\in\mathcal{X}} |\phi(x)| < +\infty$. So, the remainder term is bounded λ-a.e. on $\mathcal{X}$:

$$|h(x,\epsilon)| = \frac{|f'''(s(x))|}{6}|\epsilon|^3|\phi(x)|^3 \leq \frac{B}{6}|\epsilon|^3\,\|\phi\|^3_\infty < +\infty$$

where $|f'''(s(x))| \leq B < +\infty$ λ-a.e. for sufficiently small $\epsilon \neq 0$ because $f'''(t)$ is locally bounded at $t = 1$, and we have:

$$\begin{aligned} &\min\{1, 1+\epsilon\,\phi(x)\} \leq s(x) \leq \max\{1, 1+\epsilon\,\phi(x)\} \;\; \lambda\text{-a.e.} \\ &\Rightarrow \;\; 1 - |\epsilon|\,\|\phi\|_\infty \leq s(x) \leq 1 + |\epsilon|\,\|\phi\|_\infty \;\; \lambda\text{-a.e.} \end{aligned}$$

which means for sufficiently small $\epsilon \neq 0$, $s(x)$ will be in the neighborhood of $t = 1$ around which $f'''(t)$ is bounded. In fact, we also require $Q_\epsilon(x)/P_X(x)$ to be sufficiently close to unity λ-a.e. for (48) to be valid, and this also holds for sufficiently small $\epsilon \neq 0$ because:

$$\left|\frac{Q_\epsilon(x)}{P_X(x)} - 1\right| = |\epsilon||\phi(x)| \leq |\epsilon|\,\|\phi\|_\infty \;\; \lambda\text{-a.e.} \;.$$

We now take the expectation of both sides of (48) with respect to $\mathbb{P}_X$ and use $\mathbb{E}[\phi(X)] = 0$ to obtain:

$$D_f(Q_\epsilon||P_X) = \frac{f''(1)}{2}\,\epsilon^2\,\mathbb{E}\Big[\phi(X)^2\Big] + \mathbb{E}[h(X,\epsilon)]. \quad (50)$$

(Note that $\mathbb{E}\big[\phi(X)^2\big]$ is finite because $\phi \in \mathcal{L}^\infty(\mathcal{X},\mathbb{P}_X) \subseteq \mathcal{L}^2(\mathcal{X},\mathbb{P}_X)$, and the inclusion holds as $\mathbb{P}_X$ is a finite measure.) Since $|h(x,\epsilon)|/\epsilon^2 \leq B\,\|\phi\|_\infty^3\,|\epsilon|/6$ λ-$a.e.$ for sufficiently small $\epsilon \neq 0$, we have:

$$0 \leq \frac{|\mathbb{E}[h(X,\epsilon)]|}{\epsilon^2} \leq \mathbb{E}\left[\frac{|h(X,\epsilon)|}{\epsilon^2}\right] \leq \frac{B}{6}\,\|\phi\|_\infty^3\,|\epsilon|$$

which implies that $\mathbb{E}[h(X,\epsilon)] = o\big(\epsilon^2\big)$. Therefore, (50) produces the desired approximation. ∎

Proposition 3 asserts that if we consider the stochastic manifold of probability densities on $\mathcal{X}$ in the "neighborhood" of P_X, the f-divergence between any two densities is a squared weighted $\mathcal{L}^2$-norm. This intuition is known in the literature; the discrete case of Proposition 3 is proved in [45] where f-divergences are locally shown to be χ^2-divergences. Our proof uses additional regularity conditions to provide a more general version of the result. The specialization of Proposition 3 for KL divergence is provided in the next corollary.

Corollary 1 (Local KL Divergence): Suppose we are given a family of probability densities that are local perturbations of the reference probability density $P_X > 0$ λ-$a.e.$:

$$Q_\epsilon = P_X(\mathbf{1}_\mathcal{X} + \epsilon\,\phi)$$

where $\phi \in \mathcal{L}^\infty(\mathcal{X},\mathbb{P}_X)$ is a fixed perturbation function such that $\mathbb{E}[\phi(X)] = 0$, and Q_ϵ is a valid probability density for all $\epsilon \neq 0$ of sufficiently small magnitude. Then, the KL divergence between Q_ϵ and P_X can be locally approximated as:

$$D(Q_\epsilon||P_X) = \frac{1}{2}\epsilon^2\,\mathbb{E}\Big[\phi(X)^2\Big] + o\big(\epsilon^2\big).$$

Proof: It is straightforward to verify the conditions of Proposition 3 for $f(t) = t\log(t)$ (where $\log(\cdot)$ denotes the natural logarithm), and this choice of f generates KL divergence. ∎

Corollary 1 is well-known in the literature. The discrete version can be found in [8], and the general case in [46].

Appendix C
Proof of Lemma 2

Proof: For any polynomial $p : \mathbb{R} \to \mathbb{R}$ with degree $n \in \mathbb{N}$, $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ such that $a_n \neq 0$, we have:

$$\begin{aligned}
\forall y \in \mathbb{R}, \quad (T(p))(y) &= \int_\mathbb{R} \phi(uy + \upsilon x)\,p(x)\,d\lambda(x) \\
&= \int_\mathbb{R} \frac{1}{|\upsilon|}\phi(z)\,p\left(\frac{z-uy}{\upsilon}\right)\,d\lambda(z) \\
&= \int_\mathbb{R} \frac{1}{|\upsilon|}\phi(z)\sum_{i=0}^{n} f_i(z)\,y^i\,d\lambda(z) \\
&= \sum_{i=0}^{n} y^i \int_\mathbb{R} \frac{1}{|\upsilon|}\phi(z)\,f_i(z)\,d\lambda(z)
\end{aligned}$$

where the second equality follows from a change of variables $z = uy + \upsilon x$, and the third equality holds because $p((z-uy)/\upsilon) = a_0 + a_1(z-uy)/\upsilon + \cdots + a_n\,((z-uy)/\upsilon)^n$ is a polynomial in y with degree n, coefficients $f_i(z)$ of y^i for $i = 0,\ldots,n$ (these depend on u,υ as well), and leading coefficient $f_n(z) = a_n(-u/\upsilon)^n \neq 0$. The non-zero leading coefficient of $T(p)$ is:

$$\int_\mathbb{R} \frac{1}{|\upsilon|}\phi(z)a_n\left(\frac{-u}{\upsilon}\right)^n d\lambda(z) = \frac{a_n}{|\upsilon|}\left(\frac{-u}{\upsilon}\right)^n$$

since $\int_\mathbb{R} \phi\,d\lambda = 1$. Hence, $T(p)$ is a polynomial with degree n, which implies that T is strictly degree preserving. ∎

References

[1] A. Rényi, "On measures of dependence," *Acta Math. Acad. Sci. Hungarica*, vol. 10, nos. 3–4, pp. 441–451, 1959.

[2] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 580–598, Sep. 1985.

[3] E. Abbe and L. Zheng, "A coordinate system for Gaussian networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 721–733, Feb. 2012.

[4] R. W. Keener, *Theoretical Statistics: Topics for a Core Course* (Springer Texts in Statistics). New York, NY, USA: Springer, 2010.

[5] G. W. Wornell, "Inference and information," Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, USA, Lecture Notes 6.437, May 2017.

[6] E. J. G. Pitman, "Sufficient statistics and intrinsic accuracy," *Math. Proc. Cambridge Philos. Soc.*, vol. 32, no. 4, pp. 567–579, 1936.

[7] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," *Ann. Statist.*, vol. 7, no. 2, pp. 269–281, 1979.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[9] C. N. Morris, "Natural exponential families with quadratic variance functions," *Ann. Statist.*, vol. 10, no. 1, pp. 65–80, 1982.

[10] C. N. Morris, "Natural exponential families with quadratic variance functions: Statistical theory," *Ann. Statist.*, vol. 11, no. 2, pp. 515–529, 1983.

[11] D. Fink, "A compendium of conjugate priors," Environ. Statist. Group, Dept. Biol., Montana State Univ., Bozeman, MT, USA, Tech. Rep., May 1997.

[12] S.-L. Huang, A. Makur, F. Kozynski, and L. Zheng, "Efficient statistics: Extracting information from iid observations," in *Proc. 52nd Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep./Oct. 2014, pp. 699–706.

[13] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep./Oct. 2015, pp. 972–979.

[14] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, USA, Jul. 2012, pp. 1029–1033.

[15] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[16] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep. 1999, pp. 368–377.

[17] V. Anantharam, A. Gohari, S. Kamath, and C. Nair. (Apr. 2013). "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover." [Online]. Available: https://arxiv.org/abs/1304.6133

[18] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan. 2016.

[19] A. Makur and L. Zheng, "Bounds between contraction coefficients," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Sep./Oct. 2015, pp. 1422–1429.

[20] A. Lapidoth and S. Shamai (Shitz), "The Poisson multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 488–501, Mar. 1998.

[21] G. E. Andrews and R. Askey, "Classical orthogonal polynomials," in *Polynômes Orthogonaux et Applications* (Lecture Notes in Mathematics), vol. 1171, C. Brezinski, A. Draux, A. P. Magnus, P. Maroni, and A. Ronveaux, Eds. Berlin, Heidelberg, Germany: Springer, 1985, pp. 36–62.

[22] T. S. Chihara, *An Introduction to Orthogonal Polynomials*. New York, NY, USA: Dover, 2011.
[23] G. Szegö, *Orthogonal Polynomials* (Colloquium Publications), vol. 23, 4th ed. Providence, RI, USA: AMS, 1975.
[24] C. Komninakis, L. Vandenberghe, and R. D. Wesel, "Capacity of the binomial channel, or minimax redundancy for memoryless sources," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Washington, DC, USA, Jun. 2001, p. 127.
[25] A. M. Krall, *Hilbert Space, Boundary Value Problems and Orthogonal Polynomials* (Operator Theory: Advances and Applications), vol. 133, 1st ed. Boston, MA, USA: Birkhäuser, 2002.
[26] D. Bakry, "Functional inequalities for Markov semigroups," in *Probability Measures on Groups: Recent Directions and Trends* (Proceedings of the CIMPA-TIFR School, Tata Institute of Fundamental Research, Mumbai, India, 2002), S. G. Dani and P. Graczyk, Eds. New Delhi, India: Narosa Publishing House, 2006, pp. 91–147.
[27] O. Mazet, "Classification des semi-groupes de diffusion sur $\mathbb{R}$ associés à une famille de polynômes orthogonaux," in *Séminaire de Probabilités XXXI* (Lecture Notes in Mathematics), vol. 1655, J. Azéma, M. Yor, and M. Emery, Eds. Berlin, Heidelberg, Germany: Springer, 1997, pp. 40–53.
[28] H. O. Lancaster, "The structure of bivariate distributions," *Ann. Math. Statist.*, vol. 29, no. 3, pp. 719–736, 1958.
[29] H. O. Lancaster, *The Chi-Squared Distribution*. New York, NY, USA: Wiley, 1969.
[30] G. K. Eagleson, "Polynomial expansions of bivariate distributions," *Ann. Math. Statist.*, vol. 35, no. 3, pp. 1208–1215, 1964.
[31] R. C. Griffiths, "The canonical correlation coefficients of bivariate gamma distributions," *Ann. Math. Statist.*, vol. 40, no. 4, pp. 1401–1408, 1969.
[32] A. E. Koudou, "Probabilités de Lancaster," *Expos. Math.*, vol. 14, no. 3, pp. 247–275, 1996.
[33] A. E. Koudou, "Lancaster bivariate probability distributions with Poisson, negative binomial and gamma margins," *Test*, vol. 7, no. 1, pp. 95–110, 1998.
[34] P. Diaconis, K. Khare, and L. Saloff-Coste, "Gibbs sampling, exponential families and orthogonal polynomials," *Statist. Sci.*, vol. 23, no. 2, pp. 151–178, 2008.
[35] J. Meixner, "Orthogonale polynomsysteme mit einer besonderen gestalt der erzeugenden funktion," *J. London Math. Soc.*, vol. s1-9, pp. 6–13, Jan. 1934.
[36] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces* (Princeton Lectures in Analysis), vol. 3. Princeton, NJ, USA: Princeton Univ. Press, 2005.
[37] R. E. Megginson, *An Introduction to Banach Space Theory* (Graduate Texts in Mathematics), vol. 183. New York, NY, USA: Springer, Oct. 1998.
[38] R. Melrose, "Functional analysis," Dept. Math., MIT, Cambridge, MA, USA, Lecture Notes 18.102, May 2017.
[39] M. G. Kendall, *The Advanced Theory of Statistics*, vol. 1, 2nd ed. London, U.K.: Griffin, 1945.
[40] W. Rudin, *Principles of Mathematical Analysis* (International Series in Pure and Applied Mathematics), 3rd ed. New York, NY, USA: McGraw-Hill, 1976.
[41] I. Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von Markoffschen ketten," *Pub. Math. Inst. Hungarian Acad. Sci. A*, vol. 8, pp. 85–108, 1963.
[42] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Stud. Sci. Math. Hungarica*, vol. 2, pp. 299–318, 1967.
[43] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 28, no. 1, pp. 131–142, 1966.
[44] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jerusalem, Israel, Apr./May 2015, pp. 1–5.
[45] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial* (Foundations and Trends in Communications and Information Theory), vol. 1, no. 4, S. Verdú, Ed. Hanover, MA, USA: now Publishers Inc., 2004.
[46] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, USA, Lecture Notes 6.441, Aug. 2017.

Anuran Makur (S'16) received a B.S. degree with highest honors (*summa cum laude*) from the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley (UC Berkeley), USA, in 2013, and an S.M. degree from the Department of Electrical Engineering and Computer Science (EECS) at the Massachusetts Institute of Technology (MIT), Cambridge, USA, in 2015. He is currently pursuing a Ph.D. degree from the Department of EECS at MIT. His research interests include information theory, statistics, and other areas in applied probability. He was a recipient of the Edward Frank Kraft Award from UC Berkeley in 2011, the Leman, Rubin Merit Scholarship from UC Berkeley in 2012, the Arthur M. Hopkin Award from UC Berkeley in 2013, the Irwin Mark Jacobs and Joan Klein Jacobs Presidential Fellowship from MIT in 2013, the Hewlett-Packard Fellowship from MIT in 2015, and the Ernst A. Guillemin Master's Thesis Award from MIT in 2015.

Lizhong Zheng (S'00–M'02–F'16) received the B.S. and M.S. degrees from the Department of Electronic Engineering at Tsinghua University, Beijing, China, in 1994 and 1997, respectively, and a Ph.D. degree from the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley (UC Berkeley), USA, in 2002. Since 2002, he has been working in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, Cambridge, USA, where he is currently a Professor of Electrical Engineering. His research interests include information theory, statistical inference, and wireless communications and networks. He was a recipient of the Eli Jury Award from UC Berkeley in 2002, the IEEE Information Theory Society Paper Award in 2003, the NSF CAREER Award in 2004, and the AFOSR Young Investigator Award in 2007. He became an IEEE Fellow in 2016.