A Study of Local Approximations in Information Theory

by

Anuran Makur

B.S. Electrical Engineering and Computer Sciences
University of California, Berkeley, 2013

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Signature of Author: _____

Department of Electrical Engineering and Computer Science
April 27, 2015

Certified by: _____

Lizhong Zheng
Professor of Electrical Engineering
Thesis Supervisor

Accepted by: _____

Leslie A. Kolodziejski
Professor of Electrical Engineering
Chair, Department Committee on Graduate Students

A Study of Local Approximations in Information Theory

by

Anuran Makur

Submitted to the Department of Electrical Engineering and Computer Science
on April 27, 2015 in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Electrical Engineering and Computer Science.

ABSTRACT

The intractability of many information theoretic problems arises from the meaningful but non-linear definition of Kullback-Leibler (KL) divergence between two probability distributions. Local information theory addresses this issue by assuming all distributions of interest are perturbations of certain reference distributions, and then approximating KL divergence with a squared weighted Euclidean distance, thereby linearizing such problems. We show that large classes of statistical divergence measures, such as f-divergences and Bregman divergences, can be approximated in an analogous manner to local metrics which are very similar in form. We then capture the cost of making local approximations of KL divergence instead of using its global value. This is achieved by appropriately bounding the tightness of the Data Processing Inequality in the local and global scenarios. This task turns out to be equivalent to bounding the chordal slope of the hypercontractivity ribbon at infinity and the Hirschfeld-Gebelein-Rényi maximal correlation with each other. We derive such bounds for the discrete and finite, as well as the Gaussian regimes. An application of the local approximation technique is in understanding the large deviation behavior of sources and channels. We elucidate a source-channel decomposition of the large deviation characteristics of i.i.d. sources going through discrete memoryless channels. This is used to derive an additive Gaussian noise channel model for the local perturbations of probability distributions. We next shift our focus to infinite alphabet channels instead of discrete and finite channels. On this front, existing literature has demonstrated that the singular vectors of additive white Gaussian noise channels are Hermite polynomials, and the singular vectors of Poisson channels are Laguerre polynomials. We characterize the set of infinite alphabet channels whose singular value decompositions produce singular vectors that are orthogonal polynomials by providing equivalent conditions on the conditional moments. In doing so, we also unveil the elegant relationship between certain natural exponential families with quadratic variance functions, their conjugate priors, and their corresponding orthogonal polynomial singular vectors. Finally, we propose various related directions for future research in the hope that our work will beget more research concerning local approximation methods in information theory.

Thesis Supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering

Acknowledgments

I would first and foremost like to thank my advisor, Prof. Lizhong Zheng, for his unique guiding style. I remember being quite apprehensive of his "hands off" style when I joined the group two years ago. Having scant experience with information theory research, I felt like I was being plunged into the middle of an ocean without any knowledge of how to swim. Lizhong was extremely patient with me in these first few months. He emphasized the importance of a vision in research, focusing always on the elegance and symphony of ideas rather than the gritty details of my propositions. The visions he instilled in me served as little islands in that ocean of confusion, and I learned how to swim from one island to the next. The exposition in this thesis reflects this research style. It is a stream of diverse but connected ideas, where the diversity portrays my educational explorations, and the connections emerge from the motif of local approximations in information theory. I am truly grateful to Lizhong for emphasizing the importance of visions and big ideas to me, and offering me the freedom to pursue my sometimes over-ambitious hunches. The best way to learn to swim is indeed to jump in the water.

Next, I would like to acknowledge certain people for their help with particular sections of this thesis. The pivotal information projection idea in the proof of Theorem 4.2.1 was derived by Lizhong during one of my meetings with him. I would like to thank him separately for this. The introductory sections 3.1 and 3.2 in chapter 3 are partly based off my information theory course project (course 6.441 at MIT). So, I would like to thank my project partner Xuhong Zhang for working on the project with me.

I would like to thank several of my friends here at MIT for their support during the last two years. I am very grateful to Ganesh Ajjanagadde for reading through drafts of this thesis and providing numerous corrections and insightful suggestions. I am also thankful to Tuhin Sarkar for offering to read through parts of this thesis, and always lending me his patient ears when I need to clarify my doubts. I have had countless enriching conversations on information theory, and more generally on mathematics, with both Ganesh and Tuhin. I am grateful to Tarek Lahlou for his help in producing a beautiful figure for this thesis, and also for sharing with me his profound and maverick insights on signal processing, linear algebra, and optimization theory. I am indebted to James Noraky for being a trusted confidant, and my first point of contact when in trouble. Finally, I would like to thank Eren Kizildag, Govind Ramnarayan, Nirav Bhan, Lucas Nissenbaum, and Andy Shih for their wonderful friendship.

I would also like to express my gratitude to my lab-mates: David Qiu, Amir Salimi, Fabián Kozynski, Shao-Lun Huang, Hye Won Chung, and Mina Karzand. It is always heartening to know that we are all in this together. In particular, I would like to thank David Qiu for the comic relief he

has provided me on the most stressful of days, and his wolf-like ability to locate social events which we can attend for sustenance.

It would be foolish to try and thank my family: my mother Anindita Makur, my father Anamitra Makur, and my little sister Rhymee. A modest thank you note at the end of an acknowledgments section can hardly be considered proper appreciation for their endless love, support, and sacrifice. Nevertheless, I would like to express a small fraction of my gratitude to them in the next few lines. Ma, I am enormously grateful for the confidence, drive, and tenacity you fuel me with every time I speak to you. I can honestly say that I work tirelessly only to witness the happiness you derive from my otherwise meaningless accomplishments. Baba, I deeply admire you for your brilliant mind and carefree demeanor. Attaining your sharpness and wisdom are my lifelong goals. Thank you for your tacit belief in my abilities and your constant support for my dreams. Rhymee, to you I must ask not to worry! I know you feel that I have imprudently presented all my developed ideas in this thesis without saving any for my PhD. I have faith that I will find something new to work on over the next few years. Thank you for being a source of constant amusement and joy in my life.

Contents

List of Figures

Chapter 1

Introduction

Information theory is the mathematical discipline which classically analyzes the fundamental limits of communication. Since communication sources are conveniently modeled as stochastic processes and communication channels are modeled as probabilistic functions of these processes, hypothesis testing becomes a natural solution for decoding. As the likelihood ratio test is an optimal decoding rule in both Bayesian and non-Bayesian (Neyman-Pearson) formulations of hypothesis testing, it is unsurprising that the expected log-likelihood ratio is a fundamental quantity in information theory. Indeed, the log-likelihood ratio is a sufficient statistic for the source random variable. The expected log-likelihood ratio is defined as the Kullback-Leibler (KL) divergence or relative entropy, and several important information measures such as Shannon entropy and mutual information are known to emerge from it.

Definition 1.0.1 (KL Divergence)**.** Given a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and distributions P and Q on this space, the KL divergence between P and Q, denoted $D(P||Q)$, is given by:

$$D(P||Q) \triangleq \begin{cases} \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] & , \ P \ll Q \\ +\infty & , \ \text{otherwise} \end{cases}$$

where $P \ll Q$ denotes that P is absolutely continuous with respect to Q, $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative, and $\mathbb{E}_P[\cdot]$ denotes the abstract expectation (integration) with respect to the probability law corresponding to the distribution P.

(Discrete case) If Ω is finite or countably infinite, then given probability mass functions P and Q on Ω, the KL divergence between P and Q is given by:

$$D(P||Q) = \sum_{x \in \Omega} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

where we assume that $\forall q \geq 0, \ 0 \log\left(\frac{0}{q}\right) = 0$ and $\forall p > 0, \ p \log\left(\frac{p}{0}\right) = +\infty$ based on continuity arguments.

(Continuous case) If $\Omega = \mathbb{R}^n$ where $n \in \mathbb{Z}^+$, then given probability density functions f and g

on Ω corresponding to distributions P and Q respectively, the KL divergence between P and Q, or f and g, is given by:

$$D(P||Q) = D(f||g) = \begin{cases} \int_\Omega f(x) \log\left(\frac{f(x)}{g(x)}\right) d\lambda(x) & , \ \lambda\left(\{x \in \Omega : f(x) > 0, g(x) = 0\}\right) = 0 \\ +\infty & , \ \text{otherwise} \end{cases}$$

where λ denotes the Lebesgue measure, the integral over all $x \in \Omega$ is the Lebesgue integral, and we again assume that $\forall q \geq 0, \ 0 \log\left(\frac{0}{q}\right) = 0$ based on continuity arguments.

Although Definition 1.0.1 is fairly general, in the majority of this thesis we will be interested in discrete and finite sample spaces. So, only the discrete case of the definition will be pertinent there. We also point out that $\log(\cdot)$ refers to the natural logarithm and unless stated otherwise, this convention will hold for the entire thesis. To construe the definition of KL divergence, we first discuss some of its properties. It can be easily verified using Jensen's inequality that:

$$D(P||Q) \geq 0 \tag{1.1}$$

with equality if and only if $P = Q$ *a.e.* (almost everywhere with respect to an appropriate measure). This result on the non-negativity of KL divergence is known as Gibbs' inequality. Moreover, the KL divergence can be interpreted as a distance between distributions. This intuition becomes evident from a study of large deviation theory, but is also retrievable from the local approximation methods that follow. We will defer an explanation until we introduce the local approximation. As mentioned earlier, the KL divergence gives rise to two of the most fundamental quantities in information theory: Shannon entropy and mutual information. These are now defined in the discrete case.

Definition 1.0.2 (Shannon Entropy)**.** Given a discrete random variable X with pmf P_X on the countable set Ω, the Shannon entropy (or simply entropy) of X (or its pmf P_X) is given by:

$$H(X) = H\left(P_X\right) \triangleq -\mathbb{E}\left[\log\left(P_X(X)\right)\right] = -\sum_{x \in \Omega} P_X(x) \log\left(P_X(x)\right)$$

where we assume that $0 \log(0) = 0$ based on continuity arguments.

Definition 1.0.3 (Mutual Information)**.** Given discrete random variables X and Y defined on countable sets $\mathcal{X}$ and $\mathcal{Y}$ respectively, with joint pmf $P_{X,Y}$, the mutual information between X and Y is given by:

$$I(X;Y) = I\left(P_X; P_{Y|X}\right) \triangleq D\left(P_{X,Y}||P_X P_Y\right) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}\right).$$

Since we use natural logarithms in these definitions, entropy and mutual information are measured in nats (natural units). We note that the notation $H(X)$ and $I(X;Y)$ is more standard in the information theory literature, but the notation $H(P_X)$ and $I(P_X; P_{Y|X})$ (motivated by [1]) conveys the intuition that these information measures are properties of the distributions and do not change with values of the random variables.

Results in information theory operationally characterize the entropy as the ultimate compression of a nat of information in an appropriate asymptotic sense. Likewise, mutual information is related to the maximum number of nats that can be asymptotically sent through a communication channel per channel use. Hence, entropy and mutual information are of paramount significance in source and channel coding, respectively. It is trivial to see that for a discrete random variable X, the entropy is the self-information (or the mutual information with itself):

$$H(X) = I(X;X) = D\left(P_X||P_X^2\right). \tag{1.2}$$

Thus, equation 1.2 and Definition 1.0.3 illustrate that the KL divergence begets both these fundamental quantities.

Much that there is to know about KL divergence may be found in classical texts on information theory such as [2], which provides a lucid exposition of the material, and [1] or [3], which emphasize deeper intuition. Despite its intuitive elegance, KL divergence is analytically quite intractable even though the log function imparts useful convexity properties to it. It is asymmetric in its inputs and hence not a valid metric on the space of probability distributions, non-linear and hence algebraically challenging to manipulate, and difficult to estimate due to the non-compactness of the domain (asymptote at 0) of the log function. [4] and [5] attribute the difficulty of many unsolved problems in network information theory to these drawbacks in the definition of KL divergence.

The space of probability distributions is a manifold with the KL divergence as the distance measure. Since the neighborhood around any point in a manifold behaves like a vector space, the neighborhood around any distribution also behaves like a vector space. Compelled by this intuition, [4] and [5] propose a linearization technique where distributions of interest are assumed to be close to each other in the KL divergence sense. Under this assumption, second order Taylor approximations of the log function in the definition of KL divergence localize it into a squared weighted Euclidean norm. This transforms information theory problems into linear algebra problems, thereby providing an accessible geometric structure to such problems. It also makes these problems susceptible to the potent attack of single letterization which is often a difficult but essential step in determining the information capacity of various channels. The next sections in this chapter elaborate on this work in [4] and also introduce much of the notation that will be used in this thesis.

1.1 Local Approximation

The local approximation of KL divergence given in [4] and [5] is introduced next. We assume our sample space, $\Omega = \{1, \ldots, n\}$, is discrete and finite, and all probability mass functions (pmfs) on Ω that are of interest are close to each other (in a sense made precise below). In particular, we consider pmfs $P : \Omega \to [0,1]$ and $Q : \Omega \to [0,1]$. Since $|\Omega| = n$, we can represent any pmf on Ω as a column vector in $\mathbb{R}^n$. So, we let $P = [P(1) \cdots P(n)]^T$ and $Q = [Q(1) \cdots Q(n)]^T$ (with a slight abuse of notation for the sake of clarity). The assumption that P and Q are close to each other corresponds to Q being a perturbation of P, or vice versa. We arbitrarily choose P to be the reference pmf and precisely have:

$$Q = P + \epsilon J \tag{1.3}$$

for some small $\epsilon > 0$ and perturbation vector $J = [J(1) \cdots J(n)]^T$. Note that to be a valid perturbation, J must satisfy:

$$\sum_{x \in \Omega} J(x) = 0. \tag{1.4}$$

In general, when we define a pmf Q from the reference pmf P, we must also ensure that ϵJ satisfies:

$$\forall x \in \Omega, \;\; 0 \leq P(x) + \epsilon J(x) \leq 1, \tag{1.5}$$

but we will not impose these conditions explicitly and simply assume they hold in all future discussion since we can make ϵ arbitrarily small.

We now perform the local approximation of KL divergence. From Definition 1.0.1, we have:

$$\begin{aligned} D(P||Q) &= -\sum_{x \in \Omega} P(x) \log \left(\frac{Q(x)}{P(x)} \right) \\ D(P||Q) &= -\sum_{x \in \Omega} P(x) \log \left(1 + \epsilon \frac{J(x)}{P(x)} \right) \end{aligned} \tag{1.6}$$

Recall that the Maclaurin series (which is the Taylor series expansion around 0) of the natural logarithm, $\log(1+x)$, is:

$$\log(1+x) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{x^m}{m} \tag{1.7}$$

for all $|x| < 1$. From equation 1.7, we see that the second order Taylor approximation of the natural logarithm, $\log(1+x)$, is:

$$\log(1+x) = x - \frac{x^2}{2} + o\left(x^2\right) \tag{1.8}$$

where $o\left(x^2\right)$ denotes that $\lim_{x \to 0} \frac{o\left(x^2\right)}{x^2} = 0$. Inserting equation 1.8 into equation 1.6 and simplifying produces:

$$D(P||Q) = \frac{1}{2} \epsilon^2 \sum_{x \in \Omega} \frac{J^2(x)}{P(x)} + o\left(\epsilon^2\right). \tag{1.9}$$

We note that equation 1.9 implicitly assumes $\forall x \in \Omega, \;\; P(x) > 0$. This means that the reference pmf is not at the edge of the probability simplex. Indeed, as we are considering a neighborhood of pmfs around the reference pmf, we must require that the reference pmf is in the interior of the probability simplex. So, this implicit assumption is intuitively sound. Equation 1.9 expresses the KL divergence as a squared weighted Euclidean norm. To present it more elegantly, we introduce some new notation and definitions.

Definition 1.1.1 (Weighted Euclidean Norm)**.** Given a fixed vector $p = [p_1 \cdots p_n]^T \in \mathbb{R}^n$ such that $\forall 1 \leq i \leq n, \;\; p_i > 0$, and any vector $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$, the weighted Euclidean norm of x with respect to weights p is given by:

$$\|x\|_p \triangleq \sqrt{\sum_{i=1}^{n} \frac{x_i^2}{p_i}}.$$

Using Definition 1.1.1 and equation 1.9, we have the following definition of local KL divergence.

Definition 1.1.2 (Local KL Divergence)**.** Given a discrete and finite sample space, Ω, and a reference pmf P on Ω in the interior of the probability simplex, the local KL divergence between P and the perturbed pmf $Q = P + \epsilon J$, for some $\epsilon > 0$ and perturbation vector J, is given by:

$$D(P||Q) = \frac{1}{2}\epsilon^2 \|J\|_P^2 + o\left(\epsilon^2\right).$$

In statistics, the quantity, $\epsilon^2 \|J\|_P^2$, is known as the χ^2-divergence between Q and P. It is straightforward to derive using the second order Taylor approximation of $\log(1+x)$ given in equation 1.8 that:

$$D(P||Q) = \frac{1}{2}\epsilon^2 \|J\|_P^2 + o\left(\epsilon^2\right) = D(Q||P) \tag{1.10}$$

under the local approximation. Hence, the KL divergence becomes a symmetric weighted Euclidean metric within a neighborhood of distributions around a reference distribution in the interior of the probability simplex. In fact, any distribution in this neighborhood may be taken as the reference distribution without changing the KL divergences beyond the $o\left(\epsilon^2\right)$ term. This Euclidean characterization of local KL divergence elucidates why KL divergence is viewed as a distance measure between distributions.

1.2 Vector Space of Perturbations

We now provide two alternative ways to perceive perturbation vectors and explain their significance. Given the reference pmf P on $\Omega = \{1, \ldots, n\}$, in the interior of the probability simplex, we have already defined the additive perturbation vector, J. To obtain another pmf, $Q = P + \epsilon J$, we simply add P and ϵJ, where J provides the direction of perturbation and $\epsilon > 0$ is the parameter which controls how close P and Q are. This is encapsulated in the next definition.

Definition 1.2.1 (Additive Perturbation)**.** Given the reference pmf P on Ω, in the interior of the probability simplex, an additive perturbation, J, satisfies:

$$\sum_{x \in \Omega} J(x) = 0$$

such that $Q = P + \epsilon J$ is a valid pmf, because $\epsilon > 0$ is small enough so that $\forall x \in \Omega,\ 0 \leq Q(x) \leq 1$.

We briefly digress to introduce some notation which will be used consistently throughout this thesis. For any vector $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$, we let $[x]$ denote the $n \times n$ diagonal matrix with entries of x along its principal diagonal. So, we have:

$$[x] = \begin{bmatrix} x_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_n \end{bmatrix}. \tag{1.11}$$

Furthermore, given any function $f : \mathbb{R} \to \mathbb{R}$ which operates on scalars, for any vector $x = [x_1 \cdots x_n]^T \in \mathbb{R}^n$, the notation $f(x)$ denotes the element-wise application of the function:

$$f(x) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{1.12}$$

For example, $\sqrt{x} = \left[\sqrt{x_1} \cdots \sqrt{x_n}\right]^T$.

Returning to our discussion, since the local KL divergence between P and Q is of the form given in Definition 1.1.2, we may define an alternative normalized perturbation:

$$K = \left[\sqrt{P}\right]^{-1} J \tag{1.13}$$

which leads to the following formula for Q:

$$Q = P + \epsilon \left[\sqrt{P}\right] K. \tag{1.14}$$

Using the normalized perturbation K, we may recast Definition 1.1.2 of local KL divergence as:

$$D(P||Q) = \frac{1}{2}\epsilon^2 \|K\|^2 + o\left(\epsilon^2\right) \tag{1.15}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. Hence, the adjective "normalized" to describe perturbations K indicates that the KL divergence can be approximated by a standard Euclidean norm in terms of K instead of a weighted Euclidean norm in terms of J. The next definition summarizes the K notation.

Definition 1.2.2 (Normalized Perturbation)**.** Given the reference pmf P on Ω, in the interior of the probability simplex, a normalized perturbation, K, satisfies:

$$\sum_{x \in \Omega} \sqrt{P(x)} K(x) = 0$$

such that $Q = P + \epsilon \left[\sqrt{P}\right] K$ is a valid pmf, because $\epsilon > 0$ is small enough so that $\forall x \in \Omega,\ 0 \leq Q(x) \leq 1$.

Finally, we consider the multiplicative perturbation of the form:

$$L = [P]^{-1} J \tag{1.16}$$

which leads to the following formula for Q:

$$Q = P + \epsilon [P] L. \tag{1.17}$$

Indeed, L is a multiplicative perturbation because $\forall x \in \Omega,\ Q(x) = P(x)\left(1 + \epsilon L(x)\right)$. Moreover, it also represents the perturbation of the log-likelihood ratio between Q and P. The log-likelihood ratio is:

$$\forall x \in \Omega,\ \log\left(\frac{Q(x)}{P(x)}\right) = \log\left(\frac{P(x)\left(1 + \epsilon L(x)\right)}{P(x)}\right) = \log\left(1 + \epsilon L(x)\right) \tag{1.18}$$

using equation 1.17. From equation 1.7, we see that the first order Taylor approximation of the natural logarithm, $\log(1+x)$, is:

$$\log(1+x) = x + o(x) \tag{1.19}$$

where $o(x)$ denotes that $\lim_{x\to 0} \frac{o(x)}{x} = 0$. Combining equations 1.18 and 1.19 produces:

$$\forall x \in \Omega, \ \log\left(\frac{Q(x)}{P(x)}\right) = \epsilon L(x) + o(\epsilon). \tag{1.20}$$

In vector notation, we have:

$$\log\left([P]^{-1} Q\right) = \epsilon L + o(\epsilon)\, 1 \tag{1.21}$$

where 1 denotes the vector with all entries equal to 1. If Q is not perturbed from P, we have $Q = P$ which results in a log-likelihood ratio of 0 for each $x \in \Omega$. When Q is perturbed from P by the multiplicative perturbation L, the direction of perturbation of the log-likelihood ratio from 0 (zero vector) is also L. Hence, the multiplicative perturbation is also the perturbation of the log-likelihood ratio, and we will address L as the log-likelihood perturbation. This is presented in the next definition.

Definition 1.2.3 (Log-likelihood Perturbation)**.** Given the reference pmf P on Ω, in the interior of the probability simplex, a log-likelihood perturbation, L, satisfies:

$$\sum_{x\in\Omega} P(x)L(x) = 0$$

such that $Q = P + \epsilon [P] L$ is a valid pmf, because $\epsilon > 0$ is small enough so that $\forall x \in \Omega, \ 0 \leq Q(x) \leq 1$.

Equations 1.13 and 1.16 permit us to translate between these different representations of perturbations. As a mnemonic device, we note that the *L*og-likelihood perturbation is denoted L. Multiplying $\left[\sqrt{P}\right]$ with L produces the normalized perturbation K (shifting one letter back from L in the alphabet). Multiplying $\left[\sqrt{P}\right]$ with K produces the additive perturbation J (shifting one letter back from K in the alphabet). Throughout this thesis, we will conform to the J, K, and L notation to indicate the appropriate types of perturbations and use whichever notation simplifies our arguments. However, it is crucial to realize that J, K, and L all embody the same fundamental object. J is perhaps the most natural way to view perturbations, K leads to the most polished notation and algebra, and L is arguably the most meaningful.

As mentioned earlier, we study local perturbations of a reference distribution because the space of local perturbations is a vector space. The axioms of vector spaces are easily verified for the different perturbations we have defined. Informally, if we neglect the ϵ factor, the sum of two perturbation vectors (whether J, K, or L) is also a perturbation vector because it satisfies the zero sum condition (in Definition 1.2.1, 1.2.2, or 1.2.3). Furthermore, multiplying a perturbation vector by a scalar also gives a perturbation vector because the zero sum condition is satisfied. So, the space of perturbations is a vector space and this gives it intuitive linear structure. To add further geometric structure to it, we may also define a notion of inner product. For example, in

the additive perturbation vector space, two additive perturbation vectors, J_1 and J_2, define two different distributions on Ω, $Q_1 = P + \epsilon J_1$ and $Q_2 = P + \epsilon J_2$ for $\epsilon > 0$ small enough, both perturbed from the same reference pmf P in the interior of the probability simplex. We can then define the inner product:

$$\langle J_1, J_2 \rangle_P \triangleq \sum_{x \in \Omega} \frac{J_1(x) J_2(x)}{P(x)}. \tag{1.22}$$

This makes the additive perturbation vector space an inner product space, and the associated norm of this inner product is the local KL divergence (without the constant scaling in front). Equivalent inner products can be defined in the normalized and log-likelihood perturbation vector spaces. In fact, an equivalent inner product definition in the normalized perturbation vector space leads to the Euclidean space, $\mathbb{R}^{|\Omega|}$, with standard Euclidean inner product and norm.

Overall, such definitions of perturbations and their corresponding inner products allow us to perceive the neighborhood of a distribution as an Euclidean inner product space with notions of orthogonality and bases. This provides considerable geometric structure to an otherwise intricate mathematical system. In the subsequent section, we demonstrate the utility of applying this linearization technique.

1.3 Linear Information Coupling

Having developed the local approximation technique in the previous sections, we delineate the work in [4] to illustrate the advantage of viewing problems under this local lens. To this end, we state some definitions (based on [2]) and recall the well-known information capacity problem for discrete memoryless channels.

Definition 1.3.1 (Discrete Memoryless Channel)**.** A discrete channel consists of an input random variable X on a finite input alphabet $\mathcal{X}$, an output random variable Y on a finite output alphabet $\mathcal{Y}$, and conditional probability distributions $\forall x \in \mathcal{X},\ P_{Y|X}(\cdot|x)$ which can be written as a $|\mathcal{Y}| \times |\mathcal{X}|$ column stochastic transition probability matrix. The notation P_X and P_Y can be used to denote the marginal distributions of X and Y, respectively.

A discrete channel is said to be memoryless if the output probability distribution is conditionally independent of all previous channel inputs and outputs given the current channel input.

Definition 1.3.2 (Channel Capacity)**.** The channel capacity of a discrete memoryless channel with input random variable X on $\mathcal{X}$, output random variable Y on $\mathcal{Y}$, and conditional probability distributions $P_{Y|X}$, is defined as:

$$C = \max_{P_X} I(X;Y)$$

where the maximization is performed over all possible pmfs on $\mathcal{X}$.

In 1948, Shannon showed in his landmark paper [6] that the channel capacity represents the maximum rate at which nats of information can be sent through a discrete memoryless channel. Prompted by this classic problem (which does not lend itself to the local approximation approach because nothing guarantees the validity of such approximations), we consider a related problem called the linear information coupling problem [4].

Definition 1.3.3 (Linear Information Coupling Problem)**.** Suppose we are given an input random variable X on $\mathcal{X}$ with pmf P_X in the interior of the probability simplex, and a discrete memoryless channel with output random variable Y on $\mathcal{Y}$ and conditional probability distributions $P_{Y|X}$. Suppose further that X is dependent on another discrete random variable U, which is defined on $\mathcal{U}$ such that $|\mathcal{U}| < \infty$, and Y is conditionally independent of U given X so that $U \to X \to Y$ is a Markov chain. Assume that $\forall u \in \mathcal{U}$, the conditional pmfs $P_{X|U=u}$ are local perturbations of the reference pmf P_X:

$$\forall u \in \mathcal{U},\ \ P_{X|U=u} = P_X + \epsilon \left[\sqrt{P_X}\right] K_u$$

where $\forall u \in \mathcal{U},\ \ K_u$ are valid normalized perturbation vectors, and $\epsilon > 0$ is small enough so that $\forall u \in \mathcal{U},\ \ P_{X|U=u}$ are valid pmfs. Then, the linear information coupling problem is:

$$\max_{\substack{P_U, P_{X|U}: U \to X \to Y \\ I(U;X) \leq \frac{1}{2}\epsilon^2}} I(U;Y)$$

where we maximize over all possible pmfs P_U on $\mathcal{U}$ and all possible conditional distributions $P_{X|U}$, such that the marginal pmf of X is P_X.

We note that Definition 1.3.3 only presents the linear information coupling problem in the single letter case [4]. Intuitively, the agenda of this problem is to find the maximum amount of information that can be sent from U to Y given that only a thin layer of information can pass through X. Thus, X serves as a bottleneck between U and Y (and the problem very loosely resembles the information bottleneck method [7]).

The linear information coupling problem can be solved using basic tools from linear algebra when attacked using local approximations. We now present the solution given in [4]. Without loss of generality, let $\mathcal{X} = \{1, \ldots, n\}$ and $\mathcal{Y} = \{1, \ldots, m\}$. Moreover, let the channel conditional probabilities, $P_{Y|X}$, be denoted by the a $m \times n$ column stochastic matrix, W:

$$W = \begin{bmatrix} P_{Y|X}(1|1) & P_{Y|X}(1|2) & \cdots & P_{Y|X}(1|n) \\ P_{Y|X}(2|1) & P_{Y|X}(2|2) & \cdots & P_{Y|X}(2|n) \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y|X}(m|1) & P_{Y|X}(m|2) & \cdots & P_{Y|X}(m|n) \end{bmatrix}. \tag{1.23}$$

W takes a pmf of X as input and produces a pmf of Y as output, where the pmfs are represented as column vectors. So, $WP_X = P_Y$, where P_X and P_Y are the marginal pmfs of X and Y, respectively. From Definition 1.3.3, we know that:

$$\forall u \in \mathcal{U},\ \ P_{X|U=u} = P_X + \epsilon \left[\sqrt{P_X}\right] K_u \tag{1.24}$$

where $P_{X|U=u}$ are column vectors representing the conditional pmfs of X given $U = u$, and $\forall u \in \mathcal{U},\ \ K_u$ are valid normalized perturbation vectors. Furthermore, using the Markov property for $U \to X \to Y$, we have:

$$\forall u \in \mathcal{U},\ \ P_{Y|U=u} = WP_{X|U=u} \tag{1.25}$$

where $P_{Y|U=u}$ are column vectors representing the conditional pmfs of Y given $U = u$. Substituting equation 1.24 into equation 1.25, we have:

$$\forall u \in \mathcal{U},\ \ P_{Y|U=u} = W\left(P_X + \epsilon\left[\sqrt{P_X}\right]K_u\right)$$

$$\forall u \in \mathcal{U},\ \ P_{Y|U=u} = P_Y + \epsilon\left[\sqrt{P_Y}\right]\left(\left[\sqrt{P_Y}\right]^{-1} W\left[\sqrt{P_X}\right]\right)K_u. \tag{1.26}$$

Equation 1.26 makes the implicit assumption that P_Y is in the interior of the probability simplex. As we discussed in section 1.1, since we take P_Y as the reference pmf in the output distribution space, this assumption is reasonable. Hence, we assume that the conditional distributions, $P_{Y|X}$, satisfy all the regularity conditions necessary to ensure P_Y is in the interior of the probability simplex. For example, the condition that all entries of W are strictly positive is sufficient to conclude that $\forall y \in \mathcal{Y},\ P_Y(y) > 0$ because $\forall x \in \mathcal{X},\ P_X(x) > 0$ is assumed in Definition 1.3.3. Inspecting equation 1.26, we notice that the normalized perturbation in the output distribution space is a channel dependent linear transform applied to the normalized perturbation in the input distribution space. Thus, when analyzing the linear information coupling problem, we must recognize two different vector spaces of perturbations; the input perturbation vector space which corresponds to perturbations of P_X, and the output perturbation vector space which corresponds to perturbations of P_Y. The linear transform described by the matrix $\left[\sqrt{P_Y}\right]^{-1} W\left[\sqrt{P_X}\right]$ maps vectors in the input perturbation space to vectors in the output perturbation space. This matrix is defined as the divergence transition matrix (DTM) in Definition 1.3.4.

Definition 1.3.4 (Divergence Transition Matrix)**.** Suppose we are given an input random variable X on $\mathcal{X}$ with pmf P_X in the interior of the probability simplex, and a discrete memoryless channel with output random variable Y on $\mathcal{Y}$ and conditional probability distributions $P_{Y|X}$, such that the output pmf P_Y is also in the interior of the probability simplex. Let W be the $|\mathcal{Y}| \times |\mathcal{X}|$ column stochastic transition probability matrix, as shown in equation 1.23. The divergence transition matrix (DTM) of the channel is given by:

$$B \triangleq \left[\sqrt{P_Y}\right]^{-1} W\left[\sqrt{P_X}\right].$$

Using Definition 1.3.4, we may rewrite equation 1.26 as:

$$\forall u \in \mathcal{U},\ \ P_{Y|U=u} = P_Y + \epsilon\left[\sqrt{P_Y}\right]BK_u. \tag{1.27}$$

We now locally approximate the mutual information terms in the linear information coupling problem. To approximate $I(U;X)$, recall that mutual information can be written as an expectation of KL divergences:

$$I(U;X) = \mathbb{E}_{P_U}\left[D\left(P_{X|U}||P_X\right)\right] = \sum_{u\in\mathcal{U}} P_U(u) D\left(P_{X|U=u}||P_X\right). \tag{1.28}$$

Using equation 1.15, the local KL divergence between pmfs $P_{X|U=u}$ and P_X for every $u \in \mathcal{U}$ is:

$$D\left(P_{X|U=u}||P_X\right) = \frac{1}{2}\epsilon^2 \left\|K_u\right\|^2 + o\left(\epsilon^2\right) \tag{1.29}$$

and combining equations 1.28 and 1.29, we get:

$$I(U;X) = \frac{1}{2}\epsilon^2 \sum_{u\in\mathcal{U}} P_U(u) \left\| K_u \right\|^2 + o\left(\epsilon^2\right). \tag{1.30}$$

So, the constraint $I(U;X) \leq \frac{1}{2}\epsilon^2$ in Definition 1.3.3 becomes:

$$\sum_{u\in\mathcal{U}} P_U(u) \left\| K_u \right\|^2 \leq 1. \tag{1.31}$$

Correspondingly, to locally approximate $I(U;Y)$, we use equation 1.15 and 1.27 to give:

$$\forall u \in \mathcal{U},\ \ D\left(P_{Y|U=u} || P_Y\right) = \frac{1}{2}\epsilon^2 \left\| BK_u \right\|^2 + o\left(\epsilon^2\right) \tag{1.32}$$

which implies that:

$$I(U;Y) = \mathbb{E}_{P_U}\left[D\left(P_{Y|U} || P_Y\right)\right] = \sum_{u\in\mathcal{U}} P_U(u) D\left(P_{Y|U=u} || P_Y\right)$$

$$I(U;Y) = \frac{1}{2}\epsilon^2 \sum_{u\in\mathcal{U}} P_U(u) \left\| BK_u \right\|^2 + o\left(\epsilon^2\right). \tag{1.33}$$

Equation 1.33 contains the objective function of the maximization in Definition 1.3.3. Ignoring the $\frac{1}{2}\epsilon^2 > 0$ factor in this localized objective function and neglecting all $o\left(\epsilon^2\right)$ terms, the linear information coupling problem simplifies to:

$$\begin{aligned} \max_{P_U, \{K_u, u\in\mathcal{U}\}} \quad & \sum_{u\in\mathcal{U}} P_U(u) \left\| BK_u \right\|^2 \\ \text{subject to:} \quad & \sum_{u\in\mathcal{U}} P_U(u) \left\| K_u \right\|^2 = 1, \\ & \forall u \in \mathcal{U},\ \ \sqrt{P_X}^T K_u = 0, \\ \text{and} \quad & \sum_{u\in\mathcal{U}} P_U(u) \left[\sqrt{P_X}\right] K_u = 0. \end{aligned} \tag{1.34}$$

where the second constraint ensures that the normalized perturbations are valid, and the third constraint guarantees that the marginal pmf of X is fixed at P_X:

$$\sum_{u\in\mathcal{U}} P_U(u) P_{X|U=u} = P_X. \tag{1.35}$$

We note that the inequality constraint in equation 1.31 becomes an equality constraint in statement 1.34; this can be shown using a simple proof by contradiction. Furthermore, equations 1.29 and 1.32 bestow the DTM (which is used in the objective function of statement 1.34) with meaning. The KL divergence between the input marginal pmf P_X and the conditional pmf $P_{X|U=u}$ is given by the squared Euclidean norm of K_u, and B transforms K_u to BK_u, whose squared Euclidean norm is the KL divergence between the output marginal pmf P_Y and the conditional pmf $P_{Y|U=u}$. This explains why B is called the divergence transition matrix.

The linear information coupling problem in statement 1.34 can be solved using a singular value decomposition (SVD) of B [4]. It is readily seen that the largest singular value of B is 1 (as B originates from the column stochastic channel matrix W), and the corresponding right (input) singular vector and left (output) singular vector are $\sqrt{P_X}$ and $\sqrt{P_Y}$, respectively:

$$B\sqrt{P_X} = \sqrt{P_Y}. \tag{1.36}$$

Moreover, letting σ be the second largest singular value of B, it is well-known that:

$$\|BK_u\|^2 \leq \sigma^2 \|K_u\|^2 \tag{1.37}$$

for any valid normalized perturbation K_u, as K_u must be orthogonal to $\sqrt{P_X}$. Taking the expectation with respect to P_U on both sides of inequality 1.37 produces:

$$\sum_{u\in\mathcal{U}} P_U(u) \|BK_u\|^2 \leq \sigma^2 \sum_{u\in\mathcal{U}} P_U(u) \|K_u\|^2. \tag{1.38}$$

Then, employing the constraint: $\sum_{u\in\mathcal{U}} P_U(u) \|K_u\|^2 = 1$, on the right hand side of this inequality, we get:

$$\sum_{u\in\mathcal{U}} P_U(u) \|BK_u\|^2 \leq \sigma^2. \tag{1.39}$$

Unit norm right singular vectors of B which are orthogonal to $\sqrt{P_X}$ satisfy the first two constraints given in the optimization problem in statement 1.34, and are therefore, valid candidates for $\{K_u, u \in \mathcal{U}\}$. Without loss of generality, let $\mathcal{U} = \{1, \ldots, |\mathcal{U}|\}$, $2 \leq |\mathcal{U}| < \infty$. Observe that selecting K_1 as the unit norm right singular vector of B corresponding to σ, $K_2 = -K_1$, and $K_3 = \cdots = K_{|\mathcal{U}|} = 0$, we fulfill all the constraints of the optimization and maximize its objective function by achieving inequality 1.39 with equality. This shows that the pmf of U is irrelevant to the optimization, and we may simply assume U is a uniform Bernoulli random variable [4]. Hence, the SVD solves the linear information coupling problem because we can find $P_{X|U}$ from $\{K_u, u \in \mathcal{U}\}$.

There are several important observations that can be made at this point. Firstly, the solution to the linear information coupling problem produces a tighter data processing inequality which holds under local approximations [4]:

$$I(U;Y) \overset{\text{local}}{\leq} \sigma^2 I(U;X) \tag{1.40}$$

where $\sigma \leq 1$ is the second largest singular value of B. This inequality will be crucial in chapter 3, where we will identify σ as the Hirschfeld-Gebelein-Rényi maximal correlation. Chapter 3 will also elucidate many of the subtleties veiled by the "local" notation in equation 1.40, and bound the performance of these local approximations.

Secondly, as we mentioned earlier, the linear information coupling problem in Definition 1.3.3 is really the single letter case of a more general multi-letter problem. We briefly introduce this general problem. For a sequence of random variables, $X_1, \ldots, X_n$, we use the notation X_1^n to denote the random vector:

$$X_1^n = (X_1, \ldots, X_n). \tag{1.41}$$

Consider the Markov chain $U \to X_1^n \to Y_1^n$, where we assume (for simplicity) that X_1^n are independent identically distributed (i.i.d.) with pmf P_X. X_1^n are inputted into the single letter discrete memoryless channel (which is used n times). Then, the multi-letter case of the problem is:

$$\max_{\substack{P_U, P_{X_1^n|U}: U \to X_1^n \to Y_1^n \\ \frac{1}{n} I(U;X_1^n) \leq \frac{1}{2}\epsilon^2}} \frac{1}{n} I(U; Y_1^n) \tag{1.42}$$

where we maximize over all possible pmfs P_U and all possible conditional distributions $P_{X_1^n|U}$, such that the marginal pmf of X_1^n is the product pmf $P_{X_1^n}$. This problem can also be solved using the SVD of the corresponding DTM and some tensor algebra after performing local approximations. In fact, [4] illustrates that the simple tensor structure of the multi-letter problem after employing local approximations is what allows single letterization. We note the subtle point that problem statement 1.42 does not explicitly apply any local approximations on its conditional pmfs $P_{X_1^n|U}$. However, we implicitly assume that such local approximations are invoked as in the single letter version of the problem in Definition 1.3.3. Indeed intuitively, the constraint that the mutual information $\frac{1}{n} I(U; X_1^n)$ is small implies that the conditional pmfs $P_{X_1^n|U}$ are close to $P_{X_1^n}$ on average in the KL divergence sense by equation 1.28, and we know that locally perturbing $P_{X_1^n}$ to produce $P_{X_1^n|U}$ gives rise to small KL divergences.

Lastly, the multi-letter linear information coupling problem admits an appealing interpretation in terms of clustering. We may perceive the binary random variable U as indexing two clusters, and X_1^n as the pure data which originates (probabilistically) from either of these clusters. Suppose we observe the noisy data Y_1^n, and our objective is to identify the cluster it came from. Mathematically, we have the Markov chain $U \to X_1^n \to Y_1^n$, and we seek to maximize the mutual information between U and Y_1^n given that the mutual information between U and X_1^n is rather small (because X_1^n has a lot of redundancy in big data problems). This is precisely the multi-letter linear information coupling problem given in statement 1.42.

Therefore, we have provided an example of how the local approximation technique helps solve information theoretic problems with palpable statistical value. In doing so, we have illustrated how the technique transforms seemingly complex information theory problems into elementary problems in linear algebra. This offers an impetus to studying this local approximation technique further. The next section conveys the general philosophy of this technique and the subsequent section provides an outline of the thesis.

1.4 Philosophy of Approach

To complement the previous section which portrays how our local approximation technique is used to solve information theory problems, we now discuss the overarching philosophy of the approach. On the surface, the raison d'être of such an approach is to simplify intractable information theory problems so they become easier to solve. As mentioned earlier and evidenced in the previous section, local approximations transform information theory problems to linear algebra problems which are straightforward to solve using basic tools like the SVD. However, one may argue that this alone does not justify a study of the local approach. Although localized information theory problems admit painless solutions, these solutions often do not address the global problem. This is

evident from the discourse in section 1.3, where we solved the linear information coupling problem instead of the classical channel capacity problem. As another example, [8] suggests a novel model of communication by varying source empirical distributions based on the work in chapter 4 (which uses local approximations). In contrast, the source empirical distribution is kept (approximately) fixed in traditional channel codes such as capacity achieving fixed composition codes.

Thus, we champion the local approximation technique by addressing its value in engineering applications and theory pertaining to data processing. Real world communication or data processing systems are often much more complex than the mathematical models of theory. So, theoretical models serve only as toy problems to provide engineers with intuition with which they can approach their real world problems. This seemingly banal, but legitimate view has long been acknowledged by engineering theorists. For example, models in the theory of stochastic processes do not correspond to real world queuing problems but provide intuition about them. Likewise, the local approximation approach offers intuition on difficult (and possibly unsolved) information theory problems, even though it does not solve them.

From a purely theoretical standpoint, the local approximation method complies with a more profound understanding of the structure of data. When considering stochastic data processing problems, we must be cognizant of three interrelated spaces: the space of data, the space of distributions, and the space of random variables. The data originates from realizations of the random variables, distributions define the random variables, and empirical distributions of the data resemble the actual distributions as the amount of data increases. The spaces of data and distributions are both manifolds, and the space of random variables is often modeled as an inner product space with the covariance serving as an inner product. We restrict our attention to the spaces of data and distributions, because the actual values taken by random variables do not carry any additional information from an information theoretic perspective.

Many modern data processing methods, both simple such as principle component analysis (PCA) or more advanced such as compressive sensing methods, treat the space of data as a linear vector space instead of a more general manifold. This means that simple non-linear patterns of data, like clustering, cannot be captured by such methods. On the other hand, traditional algorithms like the Lloyd-Max algorithm or her sister, the k-means clustering algorithm, from the pattern recognition and machine learning literature address such non-linear data in a "data-blind" fashion. They simply try to find a single non-linear structure, like clusters, in the data. As a result, they perform poorly when the data does not have the non-linear structure they are looking for.

The local approximation method [4] leaves the data space as a manifold and instead imposes a linear assumption on the space of distributions. Local approximations transform the spaces of perturbations from reference distributions into vector spaces, and computing the SVD of the DTM effectively performs PCA in the space of distributions rather than the space of data (where it is conventionally employed). Since we can derive empirical distributions from the data, this is a viable method of processing data. Thus, the local approximation technique provides an effective method of data processing using elementary tools from linear algebra, while simultaneously respecting the possibly non-linear structure of data. Such considerations make this technique worthy of further study.

1.5 Outline of Thesis

This thesis delves into exploring the local approximation technique introduced in the earlier sections. It is split into four main chapters which address separate aspects of our study. Each chapter presents some pertinent background from the literature, and then proceeds to derive new results within its scope. We now provide a brief overview of the chapters.

Chapter 2 considers applying the local approximation technique of section 1.1 to approximate other statistical divergence measures. We derive that local approximations of f-divergences and Bregman divergences, which encompass large classes of other statistical divergences, produce local divergence measures which are remarkably similar in form to that in section 1.1 for KL divergence.

Since the local approximation has now been established, chapter 3 focuses on capturing the cost of making local approximations of KL divergence instead of using its global value. This is done by appropriately bounding the tightness of the Data Processing Inequality in local and global scenarios. We find such bounds for the discrete and finite, and Gaussian cases.

Chapter 4 then presents an application of the local approximation technique in understanding the large deviation behavior of sources and channels. Under the local lens, using well-known large deviation results such as Sanov's theorem, we elucidate a source-channel decomposition of the large deviation characteristics of i.i.d. sources going through memoryless channels in the discrete and finite regime. This has elegant consequences in modeling the perturbation vector channel.

Since the content of the aforementioned chapters is largely restricted to discrete and finite random variables, chapter 5 concentrates on spectral decompositions of infinite alphabet channels (channels whose input and output random variables have infinite ranges). Past literature provides examples where the singular value decompositions of such channels lead to singular vectors which are orthogonal polynomials (scaled by other functions). For example, singular vectors of Gaussian channels are Hermite polynomials, and singular vectors of Poisson channels are Laguerre polynomials. We characterize the set of infinite alphabet channels for which the singular vectors are indeed orthogonal polynomials. This unveils the elegant relationship between some natural exponential families with quadratic variance functions (namely the Gaussian, Poisson, and binomial distributions), their conjugate priors (namely the Gaussian, gamma, and beta distributions), and related orthogonal polynomials (namely the Hermite, generalized Laguerre, Meixner, Jacobi and Hahn polynomials).

Finally, chapter 6 recapitulates our main contributions and concludes the thesis by providing suggestions for future work. The thesis has been written so that basic knowledge of probability theory and linear algebra is sufficient to follow most of it. Familiarity with information theory can provide deeper intuition and understanding of the material, but is not essential as the thesis is fairly self-contained. At times, some real analysis and measure theory are used for rigor in definitions and arguments. It is also helpful to have some exposure to large deviations theory for chapter 4, and functional analysis for chapter 5. We hope that the reader finds the ensuing discussion illuminating.

Chapter 2

Locally Approximating Divergence Measures

The use of Taylor expansions to perform local approximations of functions is ubiquitous in mathematics. Indeed, many fundamental results of relevance in probability theory, including the Central Limit Theorem (CLT) and the Weak Law of Large Numbers (WLLN), are proven using Taylor approximations of characteristic functions (Fourier transforms). As we saw in chapter 1, [4] and [5] use Taylor approximations to locally approximate KL divergence in their work. We illustrate that when we locally approximate much larger classes of divergence measures using Taylor expansions, the approximations are analogous in form to the approximation of KL divergence in Definition 1.1.2. To this end, we consider two general classes of divergence measures used in statistics: the f-divergence and the Bregman divergence. Since these divergences generalize many other known divergences, finding local approximations for them is equivalent to finding local approximations for all the divergences they generalize. Thus, focusing on these two divergences allows us to make deductions about large classes of divergence measures with few calculations. The next two sections locally approximate the f-divergence and the Bregman divergence, respectively.

2.1 f-Divergence

Statistical divergences are used to define notions of distance between two probability distributions on a space of probability distributions with the same support. They satisfy certain axioms like non-negativity and vanishing when the two input distributions are equal (almost everywhere). The space of input distributions is usually a statistical manifold as discussed in section 1.4. This section is devoted to one such divergence measure, namely the f-divergence, which is also known as the Csiszár f-divergence or the Ali-Silvey distance in the literature. We now define the f-divergence.

Definition 2.1.1 (f-Divergence)**.** Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Given a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and distributions P and Q on this space such that $P \ll Q$, which denotes that P is absolutely continuous with respect to Q, the f-divergence between P and Q, denoted $D_f(P||Q)$, is given by:

$$D_f(P||Q) \triangleq \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right)\right]$$

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative, $\mathbb{E}_Q[\cdot]$ denotes the abstract expectation (integration) with respect to the probability law corresponding to the distribution Q, and we assume that $f(0) = \lim_{t\to 0^+} f(t)$.

(Discrete case) If Ω is finite or countably infinite, then given probability mass functions P and Q on Ω, the f-divergence between P and Q is given by:

$$D_f(P||Q) = \sum_{x\in\Omega} Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

where we assume that $0f\left(\frac{0}{0}\right) = 0$, $f(0) = \lim_{t\to 0^+} f(t)$, and $\forall p > 0$, $0f\left(\frac{p}{0}\right) = \lim_{q\to 0^+} qf\left(\frac{p}{q}\right)$.

(Continuous case) If $\Omega = \mathbb{R}^n$ where $n \in \mathbb{Z}^+$, then given probability density functions g and h on Ω corresponding to distributions P and Q respectively, the f-divergence between P and Q, or g and h, is given by:

$$D_f(P||Q) = D_f(g||h) = \int_\Omega h(x) f\left(\frac{g(x)}{h(x)}\right) d\lambda(x)$$

where λ denotes the Lebesgue measure, the integral over all $x \in \Omega$ is the Lebesgue integral, and we again assume that $0f\left(\frac{0}{0}\right) = 0$, $f(0) = \lim_{t\to 0^+} f(t)$, and $\forall p > 0$, $0f\left(\frac{p}{0}\right) = \lim_{q\to 0^+} qf\left(\frac{p}{q}\right)$.

Basic definitions and bounds concerning f-divergences can be found in [9]. With appropriate choices of the function f, they generalize many known divergence measures including KL divergence, total variation distance, χ^2-divergence, squared Hellinger distance, α-divergences, Jensen-Shannon divergence, and Jeffreys divergence. f-divergences satisfy many properties that are desirable for distance measures between distributions (which is natural since an f-divergence is a statistical divergence) and more generally for information measures. For example, given two probability distributions P and Q, the f-divergence is non-negative:

$$D_f(P||Q) \geq 0. \tag{2.1}$$

$D_f(P||Q)$ is also a convex function on the input pair (P, Q). Moreover, if WP and WQ are the output distributions corresponding to the input distributions P and Q after being passed through the channel (transition probabilities) W, then the f-divergence satisfies the Data Processing Inequality:

$$D_f(WP||WQ) \leq D_f(P||Q). \tag{2.2}$$

This captures the basic intuition of information loss along a Markov chain. The KL divergence inherits several such properties from the f-divergence. In fact, the Data Processing Inequality for KL divergence will be a vital component of our discussion in chapter 3.

We will now locally approximate the f-divergence by re-deriving the local approximations in section 1.1 in a more general setting. For simplicity, assume that all distributions of interest are either discrete or continuous. So, only the discrete and continuous cases of Definition 2.1.1 are pertinent. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is either countable or $\Omega = \mathbb{R}^n$ for some $n \in \mathbb{Z}^+$, and let P, Q, and R be some distributions (pmfs or pdfs) on Ω in this probability space.

Suppose P and Q are "close" to the reference distribution R, where in the discrete case we assume $\forall x \in \Omega,\ \ R(x) > 0$, and in the continuous case we assume, $R > 0\ \ a.e.$ (almost everywhere). These conditions on R are analogous to restricting the reference pmf to reside in the interior of the probability simplex in section 1.1. Precisely, we write P and Q as perturbations of R:

$$\forall x \in \Omega,\ \ P(x) = R(x) + \epsilon J_P(x) \tag{2.3}$$

$$\forall x \in \Omega,\ \ Q(x) = R(x) + \epsilon J_Q(x) \tag{2.4}$$

for some small $\epsilon > 0$ and additive perturbation functions J_P and J_Q. Note that a valid perturbation function, $J : \Omega \to \mathbb{R}$, must satisfy:

$$\begin{aligned} \sum_{x\in\Omega} J(x) &= 0 \qquad \text{(discrete case)} \\ \int_\Omega J(x)d\lambda(x) &= 0 \quad \text{(continuous case)} \end{aligned} \tag{2.5}$$

where λ denotes the Lebesgue measure, and the integral in the continuous case is the Lebesgue integral (and J must be a Borel measurable function). Furthermore, $\epsilon > 0$ is chosen small enough so that:

$$\begin{aligned} \forall x \in \Omega,\ \ 0 \le P(x), Q(x) \le 1 \quad &\text{(discrete case)} \\ P, Q \ge 0\ \ a.e. \qquad\qquad\quad &\text{(continuous case)} \end{aligned} \tag{2.6}$$

which ensures P and Q are valid pmfs or pdfs. The next theorem presents the local approximation of the f-divergence between P and Q with R taken as the reference distribution.

Theorem 2.1.1 (Local f-Divergence)**.** *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ such that $f(t)$ is twice differentiable at $t = 1$ and $f''(1) > 0$. Suppose we are given probability distributions P, Q, and R on the set Ω, such that P and Q are perturbations of the reference distribution R:*

$$\forall x \in \Omega,\ \ P(x) = R(x) + \epsilon J_P(x)$$

$$\forall x \in \Omega,\ \ Q(x) = R(x) + \epsilon J_Q(x)$$

where $\epsilon > 0$ and J_P, J_Q are valid additive perturbations. Then, the f-divergence between P and Q can be locally approximated as:

$$D_f(P||Q) = \frac{f''(1)}{2}\epsilon^2 \sum_{x\in\Omega} \frac{\left(J_P(x) - J_Q(x)\right)^2}{R(x)} + o\left(\epsilon^2\right) = \frac{f''(1)}{2} \sum_{x\in\Omega} \frac{\left(P(x) - Q(x)\right)^2}{R(x)} + o\left(\epsilon^2\right)$$

in the discrete case, and:

$$D_f(P||Q) = \frac{f''(1)}{2}\epsilon^2 \int_\Omega \frac{\left(J_P(x) - J_Q(x)\right)^2}{R(x)}\ d\lambda(x) + o\left(\epsilon^2\right) = \frac{f''(1)}{2} \int_\Omega \frac{\left(P(x) - Q(x)\right)^2}{R(x)}\ d\lambda(x) + o\left(\epsilon^2\right)$$

in the continuous case, where λ denotes the Lebesgue measure and the integral is the Lebesgue integral.

Proof.
We only prove the discrete case as the continuous case is identical if the summations are replaced with Lebesgue integrals. Using $f(1) = 0$ and Taylor's theorem about the point $t = 1$, we have:

$$f(t) = f'(1)(t-1) + \frac{1}{2}f''(1)(t-1)^2 + o\left((t-1)^2\right)$$

where $\lim_{t\to 1} \frac{o\left((t-1)^2\right)}{(t-1)^2} = 0$. This gives us:

$$\forall x \in \Omega,\ \ f\left(\frac{P(x)}{Q(x)}\right) = f'(1)\left(\frac{\epsilon(J_P(x) - J_Q(x))}{Q(x)}\right) + \frac{f''(1)}{2}\left(\frac{\epsilon(J_P(x) - J_Q(x))}{Q(x)}\right)^2 + o\left(\epsilon^2\right)$$

where $\lim_{\epsilon\to 0^+} \frac{o\left(\epsilon^2\right)}{\epsilon^2} = 0$. Taking the expectation with respect to Q produces:

$$D_f(P||Q) = \epsilon f'(1)\sum_{x\in\Omega}(J_P(x) - J_Q(x)) + \frac{f''(1)}{2}\epsilon^2\sum_{x\in\Omega}\frac{(J_P(x) - J_Q(x))^2}{R(x) + \epsilon J_Q(x)} + o\left(\epsilon^2\right)$$

and since J_P and J_Q are valid perturbations, by equation 2.5 we have:

$$D_f(P||Q) = \frac{f''(1)}{2}\epsilon^2\sum_{x\in\Omega}\frac{(J_P(x) - J_Q(x))^2}{R(x)\left(1 + \epsilon\frac{J_Q(x)}{R(x)}\right)} + o\left(\epsilon^2\right).$$

Using the Taylor approximation $(1+x)^{-1} = 1 - x + o(x)$, where $\lim_{x\to 0}\frac{o(x)}{x} = 0$, we get:

$$D_f(P||Q) = \frac{f''(1)}{2}\epsilon^2\sum_{x\in\Omega}\frac{(J_P(x) - J_Q(x))^2}{R(x)}\left(1 - \epsilon\frac{J_Q(x)}{R(x)} + o(\epsilon)\right) + o\left(\epsilon^2\right).$$

Collecting all $o\left(\epsilon^2\right)$ terms, we have the desired approximation:

$$D_f(P||Q) = \frac{f''(1)}{2}\epsilon^2\sum_{x\in\Omega}\frac{(J_P(x) - J_Q(x))^2}{R(x)} + o\left(\epsilon^2\right) = \frac{f''(1)}{2}\sum_{x\in\Omega}\frac{(P(x) - Q(x))^2}{R(x)} + o(\epsilon^2)$$

where the second equality follows from $\forall x \in \Omega,\ \ P(x) - Q(x) = \epsilon\left(J_P(x) - J_Q(x)\right)$. □

We remark that for the function f in Theorem 2.1.1, $f''(1) \geq 0$ by the convexity of f. The theorem statement requires $f''(1) > 0$ to avoid the case $f''(1) = 0$, which does not lead to any meaningful local approximation of $D_f(P||Q)$.

Theorem 2.1.1 is a rather powerful result. Intuitively, it asserts that if we zoom into the statistical manifold of distributions in the neighborhood of the reference distribution R, the f-divergence between any two distributions in the neighborhood is a squared weighted Euclidean norm regardless of the function f. Hence, all f-divergences (with f satisfying the conditions in Theorem 2.1.1) are locally equivalent to the same divergence measure to within a constant scale factor. Moreover, this

local f-divergence is symmetric in its inputs P and Q. It is also evident from the derivation of the theorem that the choice of reference distribution, R, is unimportant. Indeed, R can be any distribution as long as P and Q (and other distributions of interest) are local perturbations of it. So, the local f-divergence depends only on the neighborhood of interest and any distribution in this neighborhood is a valid reference distribution. In particular, if we set the the reference distribution $R = Q$, then the local f-divergence becomes:

$$D_f(P||Q) = \frac{f''(1)}{2} \sum_{x \in \Omega} \frac{(P(x) - Q(x))^2}{Q(x)} + o\left(\epsilon^2\right) \tag{2.7}$$

in the discrete case, and:

$$D_f(P||Q) = \frac{f''(1)}{2} \int_{\Omega} \frac{(P(x) - Q(x))^2}{Q(x)} \, d\lambda(x) + o\left(\epsilon^2\right) \tag{2.8}$$

in the continuous case. This version of local f-divergence can be recast using the χ^2-divergence, which we mentioned after Definition 1.1.2 of local KL divergence in section 1.1. We now formally define the χ^2-divergence in the discrete and continuous cases.

Definition 2.1.2 (χ^2-Divergence)**.** Given a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is either countable (discrete case) or $\Omega = \mathbb{R}^n$ for some $n \in \mathbb{Z}^+$ (continuous case), and distributions (pmfs or pdfs) P and Q on this space, the χ^2-divergence between P and Q, denoted $\chi^2(P, Q)$, is given by:

$$\chi^2(P, Q) \triangleq \sum_{x \in \Omega} \frac{(P(x) - Q(x))^2}{Q(x)}$$

in the discrete case, and:

$$\chi^2(P, Q) \triangleq \int_{\Omega} \frac{(P(x) - Q(x))^2}{Q(x)} \, d\lambda(x)$$

in the continuous case, where λ denotes the Lebesgue measure and the integral is the Lebesgue integral.

We note that the χ^2-divergence is in fact an f-divergence with $f(t) = (t-1)^2, \ t \geq 0$. So, it can be defined for general (neither discrete nor continuous) distributions using the abstract definition of f-divergence in Definition 2.1.1. However, its definition in the discrete and continuous cases suffices for our purposes. Using Definition 2.1.2 of χ^2-divergence, we can rewrite the local f-divergence (in equations 2.7 and 2.8) as:

$$D_f(P||Q) = \frac{f''(1)}{2} \chi^2(P, Q) + o\left(\epsilon^2\right) \tag{2.9}$$

in both discrete and continuous cases. This means that all f-divergences (with f satisfying the conditions in Theorem 2.1.1) are in fact locally equivalent to a particular type of f-divergence: the χ^2-divergence. It is worth mentioning that the local f-divergence can also be derived from an information geometric perspective. In fact, any f-divergence locally behaves like a Fisher information metric on the statistical manifold; the local f-divergence is exactly the Fisher information metric. We choose not to introduce local approximations in this manner, because we do not require the heavy machinery of differential geometry in our analysis. Finally, we note that [10] proves the result

in equation 2.9 for the discrete case, while our proof (which was derived independently) covers both the discrete and continuous cases.

The entire former discussion regarding local f-divergence also holds for KL divergence. This is because KL divergence is an f-divergence with $f(t) = t\log(t),\ t > 0$ where $f''(1) = 1$. The next corollary explicitly presents the local KL divergence, which trivially follows from Theorem 2.1.1.

Corollary 2.1.2 (Local KL Divergence)**.** *Suppose we are given probability distributions P, Q, and R on the set Ω, such that P and Q are perturbations of the reference distribution R:*

$$\forall x \in \Omega,\ \ P(x) = R(x) + \epsilon J_P(x)$$

$$\forall x \in \Omega,\ \ Q(x) = R(x) + \epsilon J_Q(x)$$

where $\epsilon > 0$ and J_P, J_Q are valid additive perturbations. Then, the KL divergence between P and Q can be locally approximated as:

$$D(P||Q) = \frac{1}{2}\epsilon^2 \sum_{x\in\Omega} \frac{\left(J_P(x) - J_Q(x)\right)^2}{R(x)} + o\left(\epsilon^2\right) = \frac{1}{2}\sum_{x\in\Omega} \frac{\left(P(x) - Q(x)\right)^2}{R(x)} + o\left(\epsilon^2\right)$$

in the discrete case, and:

$$D(P||Q) = \frac{1}{2}\epsilon^2 \int_\Omega \frac{\left(J_P(x) - J_Q(x)\right)^2}{R(x)}\ d\lambda(x) + o\left(\epsilon^2\right) = \frac{1}{2}\int_\Omega \frac{\left(P(x) - Q(x)\right)^2}{R(x)}\ d\lambda(x) + o\left(\epsilon^2\right)$$

in the continuous case, where λ denotes the Lebesgue measure and the integral is the Lebesgue integral.

Corollary 2.1.2 generalizes Definition 1.1.2 to infinite discrete and continuous cases. The local KL divergence in this corollary is consistent with the approximation derived in [4]; the only difference is that [4] chooses $R = Q$. Our analysis in the ensuing chapters will hinge upon the local KL divergence. Fortuitously, these local results will hold for f-divergences as well, as the local KL divergence is essentially equivalent to the local f-divergence.

2.2 Bregman Divergence

We now turn our attention to the Bregman divergences. Because of their geometrically meaningful definition, these divergences are attractive in algorithmic fields such as machine learning and computational geometry. We present their definition in the discrete and finite case below.

Definition 2.2.1 (Bregman Divergence)**.** Let $g : \mathcal{P} \to \mathbb{R}$ be a strictly convex function on the convex set $\mathcal{P} \subseteq \mathbb{R}^n$ where $n \in \mathbb{Z}^+$, such that g is differentiable on relint($\mathcal{P}$), the non-empty relative interior of $\mathcal{P}$. Given two points $p \in \mathcal{P}$ and $q \in$ relint($\mathcal{P}$), the Bregman divergence between them, denoted $B_g(p, q)$, is given by:

$$B_g(p, q) \triangleq g(p) - g(q) - \nabla g(q)^T (p - q)$$

where ∇g denotes the gradient of g, which is defined $\forall x = [x_1\ \cdots\ x_n]^T \in$ relint($\mathcal{P}$) as:

$$\nabla g(x) = \left[\frac{\partial g}{\partial x_1}\ \cdots\ \frac{\partial g}{\partial x_n}\right]^T.$$

The technical condition that g is differentiable on relint($\mathcal{P}$) is perhaps the only aspect of Definition 2.2.1 which requires clarification. Intuitively, we would like g to be differentiable on the entire set $\mathcal{P}$ with the possible exception of its boundary. So, it seems as though enforcing g to be differentiable on $\mathcal{P}^\circ$, the interior of the $\mathcal{P}$, suffices. However, when $\mathcal{P}$ lives in a subspace of $\mathbb{R}^n$, $\mathcal{P}^\circ$ is empty. The relative interior of $\mathcal{P}$, relint($\mathcal{P}$), refers to the interior of $\mathcal{P}$ with respect to the subspace on which $\mathcal{P}$ lives. Hence, letting g be differentiable on relint($\mathcal{P}$) is the right condition in this context.

From Definition 2.2.1, we see that the Bregman divergence is the error in the first order Taylor approximation of the function g around the point $q \in \text{relint}(\mathcal{P})$, evaluated at the point $p \in \mathcal{P}$. This provides an elegant geometric intuition for it. [11] provides a comprehensive list of the properties exhibited by Bregman divergences and illustrates their use in clustering. Bregman divergences also generalize many known divergence measures including squared Euclidean distance, squared Mahalanobis distance, Itakura-Saito distance, and KL divergence. Much like the f-divergence, the Bregman divergence is non-negative due to the convexity of g. Indeed, for any $p \in \mathcal{P}$ and $q \in \text{relint}(\mathcal{P})$:

$$B_g(p, q) \geq 0 \tag{2.10}$$

with equality if and only if $p = q$. $B_g(p, q)$ is also convex in its first argument p. Many notable results from information geometry such as Pythagoras' theorem for KL divergence can be generalized for Bregman divergences. [11] presents a Pythagoras' theorem for Bregman divergences which can be used to create concepts like Bregman projections in analogy with information projections (i-projections). A particularly useful property of Bregman divergences is their well-known affine equivalence property [11], which is presented in the next lemma.

Lemma 2.2.1 (Affine Equivalence of Bregman Divergence)**.** *Let $g : \mathcal{P} \to \mathbb{R}$ be a strictly convex function on the convex set $\mathcal{P} \subseteq \mathbb{R}^n$ where $n \in \mathbb{Z}^+$, such that g is differentiable on relint($\mathcal{P}$), the non-empty relative interior of $\mathcal{P}$. Let $f : \mathcal{P} \to \mathbb{R}$, $f(x) = a^T x + b$, be an affine function where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are fixed. For any two points $p \in \mathcal{P}$ and $q \in$ relint($\mathcal{P}$), we have:*

$$B_{g+f}(p, q) = B_g(p, q).$$

Proof.
By Definition 2.2.1, for any two points $p \in \mathcal{P}$ and $q \in \text{relint}(\mathcal{P})$:

$$B_{g+f}(p, q) = g(p) + f(p) - g(q) - f(q) - (\nabla g(q) + \nabla f(q))^T (p - q).$$

Since $\forall x \in \mathcal{P}$, $\nabla f(x) = a$, we get:

$$\begin{aligned} B_{g+f}(p, q) &= g(p) + a^T p + b - g(q) - a^T q - b - (\nabla g(q) + a)^T (p - q) \\ &= g(p) - g(q) - \nabla g(q)^T (p - q) \\ &= B_g(p, q) \end{aligned}$$

by Definition 2.2.1. This completes the proof. □

We now consider locally approximating Bregman divergences. This can be done using the multivariate version of Taylor's theorem. Lemma 2.2.1 will be useful in understanding the intuition behind this approximation. However, the proof will not use the lemma as it is not trivial to employ the

lemma rigorously in the proof. The next theorem presents the local Bregman divergence. We note that the notation for points in the set $\mathcal{P}$ is changed from lower case to upper case letters in Theorem 2.2.2 to permit a smooth transition to the subsequent discussion on probability distributions.

Theorem 2.2.2 (Local Bregman Divergence)**.** *Let $g : \mathcal{P} \to \mathbb{R}$ be a strictly convex function on the convex set $\mathcal{P} \subseteq \mathbb{R}^n$ where $n \in \mathbb{Z}^+$, such that g is twice continuously differentiable on relint($\mathcal{P}$), the non-empty relative interior of $\mathcal{P}$, and the Hessian matrix of g, $\nabla^2 g$, is symmetric and positive semidefinite:*

$$\nabla^2 g = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_n} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 g}{\partial x_2^2} & \cdots & \frac{\partial^2 g}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1} & \frac{\partial^2 g}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_n^2} \end{bmatrix} \succeq 0.$$

Suppose we are given the points $P \in \mathcal{P}$ and $Q, R \in$ relint($\mathcal{P}$), such that P and Q are perturbations of the reference point R:

$$P = R + \epsilon J_P$$

$$Q = R + \epsilon J_Q$$

for some small $\epsilon > 0$, and J_P and J_Q which do not violate $P \in \mathcal{P}$ and $Q \in$ relint($\mathcal{P}$), respectively. Then, the Bregman divergence between P and Q can be locally approximated as:

$$B_g(P, Q) = \frac{1}{2}\epsilon^2 (J_P - J_Q)^T \nabla^2 g(R)(J_P - J_Q) + o\left(\epsilon^2\right) = \frac{1}{2}(P - Q)^T \nabla^2 g(R)(P - Q) + o\left(\epsilon^2\right).$$

Proof.
By Taylor's theorem, we can perform a second order Taylor approximation of g around $R \in$ relint($\mathcal{P}$). Letting any $P = R + \epsilon J_P \in \mathcal{P}$, for some small $\epsilon > 0$, be the input to the Taylor approximation, we have:

$$g(P) = g(R) + \nabla g(R)^T (P - R) + \frac{1}{2} (P - R)^T \nabla^2 g(R) (P - R) + o\left(\epsilon^2\right)$$

where we express the error in the second order Taylor approximation as $o\left(\epsilon^2\right)$, which denotes $\lim_{\epsilon \to 0^+} \frac{o\left(\epsilon^2\right)}{\epsilon^2} = 0$, because $P - R = \epsilon J_P$. Expanding and collecting appropriate terms on the right hand side of this equation, we have:

$$\begin{aligned} g(P) &= \left(g(R) - \nabla g(R)^T R + \frac{1}{2} R^T \nabla^2 g(R) R\right) + \left(\nabla g(R) - \nabla^2 g(R) R\right)^T P \\ &\quad + \frac{1}{2} P^T \nabla^2 g(R) P + o\left(\epsilon^2\right). \end{aligned} \tag{2.11}$$

Thus, for any $P \in \mathcal{P}$ and $Q \in$ relint($\mathcal{P}$), such that P and Q are perturbations of R: $P = R + \epsilon J_P$ and $Q = R + \epsilon J_Q$ for some small $\epsilon > 0$, we have the Bregman divergence:

$$\begin{aligned} B_g(P, Q) &= g(P) - g(Q) - \nabla g(Q)^T (P - Q) \\ &= \frac{1}{2} P^T \nabla^2 g(R) P - \frac{1}{2} Q^T \nabla^2 g(R) Q + \left(\nabla g(R) - \nabla^2 g(R) R\right)^T (P - Q) \\ &\quad - \nabla g(Q)^T (P - Q) + o\left(\epsilon^2\right). \end{aligned} \tag{2.12}$$

To evaluate the $\nabla g(Q)$ term in equation 2.12, we find the first order Taylor approximation of $\nabla g : \text{relint}(\mathcal{P}) \to \mathbb{R}^n$ around R. Using Taylor's theorem:

$$\nabla g(Q) = \nabla g(R) + \nabla^2 g(R)(Q - R) + \vec{o}(\epsilon)$$

where $\vec{o}(\epsilon)$ denotes a vector in $\mathbb{R}^n$ whose entries are all $o(\epsilon)$, which stands for $\lim_{\epsilon \to 0^+} \frac{o(\epsilon)}{\epsilon} = 0$. The error in the Taylor approximation is expressed as $\vec{o}(\epsilon)$ because $Q - R = \epsilon J_Q$. We also note that the Hessian matrix $\nabla^2 g$ of g is used in the above equation because it is the Jacobian matrix of ∇g. Substituting the Taylor approximation of $\nabla g(Q)$ into equation 2.12, we have:

$$\begin{aligned}
B_g(P, Q) &= \frac{1}{2} P^T \nabla^2 g(R) P - \frac{1}{2} Q^T \nabla^2 g(R) Q + \left(\nabla g(R) - \nabla^2 g(R) R\right)^T (P - Q) \\
&\quad - \left(\nabla g(R) + \nabla^2 g(R)(Q - R) + \vec{o}(\epsilon)\right)^T (P - Q) + o\left(\epsilon^2\right) \\
&= \frac{1}{2} P^T \nabla^2 g(R) P - \frac{1}{2} Q^T \nabla^2 g(R) Q - Q^T \nabla^2 g(R)(P - Q) + \vec{o}(\epsilon)^T (P - Q) + o\left(\epsilon^2\right) \\
&= \frac{1}{2} \left(P^T \nabla^2 g(R) P - 2 Q^T \nabla^2 g(R) P + Q^T \nabla^2 g(R) Q\right) + \epsilon \vec{o}(\epsilon)^T (J_P - J_Q) + o\left(\epsilon^2\right) \\
&= \frac{1}{2} (P - Q)^T \nabla^2 g(R)(P - Q) + o\left(\epsilon^2\right) \\
&= \frac{1}{2} \epsilon^2 (J_P - J_Q)^T \nabla^2 g(R)(J_P - J_Q) + o\left(\epsilon^2\right)
\end{aligned}$$

as required. This completes the proof. □

Theorem 2.2.2. and its proof contain some subtle points which require further clarification. Firstly, while the local approximation of Bregman divergence makes it clear that g must be twice differentiable, the reason why g is twice continuously differentiable in the theorem statement is inconspicuous. We add the continuity assumption on the second partial derivatives of g because Clairaut's theorem states that this is a sufficient condition to conclude that the Hessian matrix $\nabla^2 g$ is symmetric (or that the partial derivative operators commute). Clairaut's theorem can be found in introductory texts on multivariate calculus like [12]. The proof of Theorem 2.2.2 uses the symmetry of $\nabla^2 g$ several times.

Secondly, Theorem 2.2.2 states that the Hessian matrix $\nabla^2 g$ is positive semidefinite, although g is strictly convex. Indeed, positive semidefiniteness of $\nabla^2 g$ is all we can deduce from the strict convexity of g; we cannot conclude $\nabla^2 g$ is positive definite [13]. However, the positive definiteness of $\nabla^2 g(R)$ is a very desirable condition even though it is not required for the proof of Theorem 2.2.2. When $\nabla^2 g(R)$ is positive semidefinite but not positive definite, it has an eigenvalue of 0 and hence a non-empty nullspace. This means that the local Bregman divergence between P and Q may be 0 if $P - Q \in \text{nullspace}\left(\nabla^2 g(R)\right)$ and $P \neq Q$. This contradicts our intuition of divergences which should vanish only when the inputs are equal. If $\nabla^2 g(R) \succ 0$ (positive definite), the local Bregman divergence is 0 if and only if its inputs are equal. Hence, the condition $\nabla^2 g(R) \succ 0$ causes the local Bregman divergence to conform to our intuition.

Finally, we provide some intuition on Theorem 2.2.2 by suggesting an alternative non-rigorous

proof. We first rewrite equation 2.11 in the proof of Theorem 2.2.2 for convenience:

$$\begin{aligned} g(P) &= \left(g(R) - \nabla g(R)^T R + \frac{1}{2} R^T \nabla^2 g(R) R\right) + \left(\nabla g(R) - \nabla^2 g(R) R\right)^T P \\ &\quad + \frac{1}{2} P^T \nabla^2 g(R) P + o\left(\epsilon^2\right). \end{aligned} \tag{2.13}$$

Notice that the first two terms on the right hand side of this equation form an affine function of P. Consider an alternative function, $h : \mathcal{P} \to \mathbb{R}$, which is the non-affine part of $g(P)$ in equation 2.13:

$$\forall P \in \mathcal{P}, \;\; h(P) = \frac{1}{2} P^T \nabla^2 g(R) P \tag{2.14}$$

where $\nabla^2 g(R)$ is a fixed matrix and we neglect the $o\left(\epsilon^2\right)$ term. Assuming that $\nabla^2 g(R)$ is symmetric and positive definite, the quadratic form h is strictly convex on $\mathcal{P}$. It is also differentiable on relint($\mathcal{P}$). So, for any $P \in \mathcal{P}$ and $Q \in$ relint($\mathcal{P}$), h can be used to define the Bregman divergence:

$$\begin{aligned} B_h(P, Q) &= h(P) - h(Q) - \nabla h(Q)^T (P - Q) \\ &= \frac{1}{2} P^T \nabla^2 g(R) P - \frac{1}{2} Q^T \nabla^2 g(R) Q - Q^T \nabla^2 g(R)(P - Q) \\ &= \frac{1}{2}\left(P^T \nabla^2 g(R) P - 2 Q^T \nabla^2 g(R) P + Q^T \nabla^2 g(R) Q\right) \\ &= \frac{1}{2}(P - Q)^T \nabla^2 g(R)(P - Q). \end{aligned} \tag{2.15}$$

By the affine equivalence property of Bregman divergences given in Lemma 2.2.1, we have that:

$$B_g(P, Q) = \frac{1}{2}(P - Q)^T \nabla^2 g(R)(P - Q) + o\left(\epsilon^2\right) \tag{2.16}$$

where we reinsert the $o\left(\epsilon^2\right)$ term. While equation 2.16 matches the statement of Theorem 2.2.2, this is clearly not a rigorous proof of the local Bregman divergence. The positive definite assumption on $\nabla^2 g(R)$ (which does not hold for all strictly convex g) was essential to ensure h is strictly convex, which in turn was essential in defining a Bregman divergence associated with h. Moreover, we neglected a thorough analysis of the $o\left(\epsilon^2\right)$ term. On the other hand, this calculation readily elucidates the intuition behind the local Bregman divergence. The second order Taylor approximation of g leads to a quadratic function whose affine part does not affect the associated Bregman divergence. Hence, we locally see the Bregman divergence associated to the quadratic term in the Taylor approximation of g. This is why the local Bregman divergence resembles a squared Euclidean distance with a symmetric weighting matrix.

In general, given the quadratic form $h : \mathcal{P} \to \mathbb{R}, \;\; h(x) = x^T A x$, where A is symmetric and positive definite to ensure h is strictly convex and h is differentiable on relint($\mathcal{P}$), the associated Bregman divergence between any $P \in \mathcal{P}$ and $Q \in$ relint($\mathcal{P}$) is given by the derivation preceding equation 2.15:

$$B_h(P, Q) = (P - Q)^T A (P - Q). \tag{2.17}$$

This particular Bregman divergence is known as the squared Mahalanobis distance, although the term is usually reserved for when A is the inverse of a covariance matrix [11]. The Mahalanobis distance has many applications in classification methods like linear discriminant analysis. Theorem

2.2.2 illustrates that much like f-divergences, all Bregman divergences (with g satisfying the conditions of the theorem and $\nabla^2 g$ being positive definite) are locally equivalent to a particular type of Bregman divergence: the squared Mahalanobis distance.

On the other hand, unlike the f-divergence, the Bregman divergence does not inherently operate on probability distributions. However, we may take $\mathcal{P}$ to be the probability simplex in $\mathbb{R}^n$. This defines Bregman divergences between pmfs in the discrete and finite case (which is the only case in which Bregman divergences are defined). We now compare the local approximations of Bregman and f-divergences. Restricting Theorem 2.1.1 to the discrete and finite case, the local f-divergence between the pmfs $P = R + \epsilon J_P \in \mathcal{P}$ and $Q = R + \epsilon J_Q \in \text{relint}(\mathcal{P})$, where $\epsilon > 0$, for some reference pmf $R \in \text{relint}(\mathcal{P})$ is:

$$D_f(P||Q) = \frac{f''(1)}{2}(P-Q)^T [R]^{-1} (P-Q) + o\left(\epsilon^2\right) \tag{2.18}$$

where we use the notation for diagonal matrix defined in section 1.2 in equation 1.11. Restating the local Bregman divergence in Theorem 2.2.2 for convenience, we have:

$$B_g(P,Q) = \frac{1}{2}(P-Q)^T \nabla^2 g(R)(P-Q) + o\left(\epsilon^2\right). \tag{2.19}$$

Since the Hessian matrix $\nabla^2 g(R)$ is symmetric and positive semidefinite, we can orthogonally diagonalize it by the spectral theorem:

$$\nabla^2 g(R) = UDU^T \tag{2.20}$$

where U is an orthogonal matrix of right eigenvectors of $\nabla^2 g(R)$, and D is a diagonal matrix of eigenvalues (which are all non-negative as $\nabla^2 g(R)$ is positive semidefinite). Substituting equation 2.20 into equation 2.19, we get:

$$B_g(P,Q) = \frac{1}{2}\left(U^T(P-Q)\right)^T D \left(U^T(P-Q)\right) + o\left(\epsilon^2\right). \tag{2.21}$$

Comparing equations 2.18 and 2.21, we see that both the local f-divergence and the local Bregman divergence are proportional to different squared weighted norms of $P - Q$. In the local f-divergence, the weights are given by the diagonal matrix $[R]^{-1}$. In the local Bregman divergence, we first change the basis of $P - Q$ using U^T, and then take its squared norm with respect to the diagonal weight matrix D. Although the local Bregman and f-divergences are similar, the general form of the local f-divergence matches that of the local KL divergence (Corollary 2.1.2) while that of the local Bregman divergence does not. Unfortunately, this means the results we will develop for local KL divergence in the ensuing chapters will not necessarily generalize for Bregman divergences.

To redirect our discussion to local KL divergence, which will be the focus of the remainder of this thesis, we verify that the local Bregman divergence formula agrees with that of the local KL divergence. The KL divergence is a Bregman divergence for the strictly convex function $H_- : \mathcal{P} \to \mathbb{R}$, where $\mathcal{P} \in \mathbb{R}^n$ is the probability simplex, and H_- is the negative Shannon entropy function [11] given in Definition 1.0.2:

$$\forall P = [p_1 \ \cdots \ p_n]^T \in \mathcal{P}, \ \ H_-(P) = \sum_{i=1}^{n} p_i \log(p_i). \tag{2.22}$$

Noting that H_- is twice continuously differentiable on relint($\mathcal{P}$), we compute the Hessian matrix of H_- for any $R \in$ relint($\mathcal{P}$):

$$\nabla^2 H_-(R) = [R]^{-1} \tag{2.23}$$

where we use the notation for diagonal matrix defined in section 1.2 in equation 1.11. Using Theorem 2.2.2, we have for a reference pmf $R \in$ relint($\mathcal{P}$), and any $P = R + \epsilon J_P \in \mathcal{P}$ and $Q = R + \epsilon J_Q \in$ relint($\mathcal{P}$) for some $\epsilon > 0$:

$$D(P||Q) = B_{H_-}(P, Q) = \frac{1}{2}(P - Q)^T [R]^{-1} (P - Q) + o\left(\epsilon^2\right). \tag{2.24}$$

This is consistent with the local KL divergence formula given in Corollary 2.1.2, and Definition 1.1.2 with $R = Q$. We note that $[R]^{-1}$ is well-defined as $R \in$ relint($\mathcal{P}$). In section 1.1 where we first introduced the reference pmf, we restricted it to the "interior of the probability simplex". This meant that all probability masses of the reference pmf were strictly positive. The discussion following Definition 2.2.1 explained why we should actually use the concept of relative interior to correctly define this notion. Although we avoided this additional rigor in the introductory chapter for simplicity, we will use it from hereon.

Before closing this chapter, we draw attention to the beautiful interpretation of KL divergence imparted by the geometric definition of Bregman divergence. Writing KL divergence using Definition 2.2.1 of Bregman divergence, we have for any $P \in \mathcal{P}$ and $Q \in$ relint($\mathcal{P}$):

$$\begin{aligned} D(P||Q) &= H_-(P) - H_-(Q) - \nabla H_-(Q)^T (P - Q) \\ &= H(Q) + \nabla H(Q)^T (P - Q) - H(P) \end{aligned} \tag{2.25}$$

where $H : \mathcal{P} \to \mathbb{R}$ denotes the Shannon entropy function from Definition 1.0.2. Equation 2.25 characterizes the KL divergence between P and Q as the non-negative error in the first order Taylor approximation of the Shannon entropy function around the pmf Q, evaluated at the pmf P. This elegant, albeit uncommon interpretation of KL divergence turns out to be useful in the next chapter.

Chapter 3

Bounds on Local Approximations

In this chapter, we analyze the performance of algorithms which exploit the local approximation framework introduced in chapter 1. We restrict our attention to algorithms developed from the study of the linear information coupling problem [4]. [8] provides an example of such an algorithm for inference on hidden Markov models and illustrates its use in image processing. To shed light on how we evaluate performance, we first recapitulate the linear information coupling problem given in Definition 1.3.3 in section 1.3. In this problem, we are given a Markov chain $U \to X \to Y$, where all alphabet sets are discrete and finite, and the marginal pmf P_X and channel conditional probabilities $P_{Y|X}$ are known. The objective is to maximize the mutual information between U and Y given that the mutual information between U and X is constrained. This is formally shown in the statement below:

$$\max_{\substack{P_U, P_{X|U}: U \to X \to Y \\ I(U;X) \leq \frac{1}{2}\epsilon^2}} I(U;Y) \tag{3.1}$$

where we maximize over all possible pmfs P_U and all possible conditional pmfs $P_{X|U}$. We assume that the conditional pmfs $P_{X|U}$ are perturbations of the marginal pmf P_X:

$$\forall u \in \mathcal{U}, \;\; P_{X|U=u} = P_X + \epsilon \left[\sqrt{P_X}\right] K_u \tag{3.2}$$

where $\mathcal{U}$ is the alphabet set of U, $\{K_u, u \in \mathcal{U}\}$ are normalized perturbation vectors, and $\epsilon > 0$ is small enough so that $\forall u \in \mathcal{U}, \;\; P_{X|U=u}$ are valid pmfs. As shown in statement 1.34 in section 1.3, applying such local approximations transforms the optimization problem in statement 3.1 into:

$$\begin{aligned} \max_{\{K_u, u \in \mathcal{U}\}} \quad & \sum_{u \in \mathcal{U}} P_U(u) \left\| B K_u \right\|^2 \\ \text{subject to:} \quad & \sum_{u \in \mathcal{U}} P_U(u) \left\| K_u \right\|^2 = 1, \\ & \forall u \in \mathcal{U}, \;\; \sqrt{P_X}^T K_u = 0, \\ \text{and} \quad & \sum_{u \in \mathcal{U}} P_U(u) \left[\sqrt{P_X}\right] K_u = 0. \end{aligned} \tag{3.3}$$

where B is the DTM, and we only maximize over the vectors $\{K_u, u \in \mathcal{U}\}$ because P_U does not affect the optimization.

For purposes which will soon become apparent, we transform the optimization problem in statement 3.3 into another equivalent optimization problem. Recall equation 1.38 from section 1.3:

$$\sum_{u\in\mathcal{U}} P_U(u)\left\|BK_u\right\|^2 \leq \sigma^2 \sum_{u\in\mathcal{U}} P_U(u)\left\|K_u\right\|^2 \tag{3.4}$$

where $0 \leq \sigma \leq 1$ is the second largest singular value of B. Rearranging this, we get:

$$\frac{\displaystyle\sum_{u\in\mathcal{U}} P_U(u)\left\|BK_u\right\|^2}{\displaystyle\sum_{u\in\mathcal{U}} P_U(u)\left\|K_u\right\|^2} \leq \sigma^2 \tag{3.5}$$

assuming the denominator on the left hand side is strictly positive. Without loss of generality, let the alphabet set $\mathcal{U} = \{1,\ldots,|\mathcal{U}|\}$, where $2 \leq |\mathcal{U}| < \infty$. Recalling from section 1.3 that choosing K_1 to be the unit norm right singular vector of B corresponding to σ, $K_2 = -K_1$, and $K_3 = \cdots = K_{|\mathcal{U}|} = 0$ solves the problem in statement 3.3, it is readily seen that this choice of $\{K_u, u \in \mathcal{U}\}$ also achieves inequality 3.5 with equality. Hence, this choice of $\{K_u, u \in \mathcal{U}\}$ solves the optimization problem given below:

$$\begin{aligned}
\sup_{\{K_u, u\in\mathcal{U}\}} \quad & \frac{\displaystyle\sum_{u\in\mathcal{U}} P_U(u)\left\|BK_u\right\|^2}{\displaystyle\sum_{u\in\mathcal{U}} P_U(u)\left\|K_u\right\|^2} \\
\text{subject to:} \quad & \sum_{u\in\mathcal{U}} P_U(u)\left\|K_u\right\|^2 > 0, \\
& \forall u \in \mathcal{U},\ \ \sqrt{P_X}^T K_u = 0, \\
\text{and} \quad & \sum_{u\in\mathcal{U}} P_U(u)\left[\sqrt{P_X}\right] K_u = 0.
\end{aligned} \tag{3.6}$$

Once again, note that P_U does not affect the optimization. By referring back to section 1.3, where we derived the different terms and constraints used in problem statement 3.6, we recognize that problem 3.6 is equivalent to:

$$\sup_{\substack{P_U, P_{X|U}: U\to X\to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \tag{3.7}$$

where P_X and $P_{Y|X}$ are fixed, and we try to find the optimizing $P_{X|U}$ by assuming they are local perturbations of P_X. Moreover, the preceding discussion reveals that:

$$\sup_{\substack{P_U, P_{X|U}: U\to X\to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \overset{\text{local}}{=} \sigma^2 \tag{3.8}$$

under the local approximations of $P_{X|U}$. This is proven rigorously in section 3.3 where $\overset{\text{local}}{=}$ is made precise. For now, it suffices to observe that equation 3.8 trivially implies the tighter Data Processing Inequality (DPI) presented in equation 1.40 in section 1.3:

$$I(U;Y) \overset{\text{local}}{\leq} \sigma^2 I(U;X) \tag{3.9}$$

which also holds under local approximations. We deduce from this discussion that the linear information coupling problem (statement 3.3) is equivalent to the problem in statement 3.6, which corresponds to the tightness of the DPI. Moreover, both problems are solved by computing the SVD of B.

Our interest in problem 3.7 is twofold. Under local approximations, it is equivalent to the linear information coupling problem. Furthermore, it is a recognized global problem (without any approximations) in the information theory literature. Therefore, by comparing the optimal value of this problem in the local and global scenarios, we can evaluate the performance of algorithms inspired by the local SVD solution to the linear information coupling problem. The global optimal value of problem 3.7 is intimately related to the concept of hypercontractivity, and the local optimal value equals the Hirschfeld-Gebelein-Rényi maximal correlation. We first introduce some of the literature surrounding hypercontractivity and the Hirschfeld-Gebelein-Rényi maximal correlation in sections 3.1 and 3.2, respectively. Then, we compare these values using bounds for the discrete and finite, and Gaussian cases, respectively, in sections 3.3 and 3.4.

3.1 Hypercontractivity

Hypercontractivity is a fundamental notion in statistics that has found applications in information theory, complexity theory, and quantum field theory. This is because hypercontractive inequalities are useful for bounding arguments in probabilistic theorems, and more generally in studying extremal problems in probabilistic spaces with distance measures. Hypercontractivity often finds use in information theory due to the tensorization properties it imparts on the quantities derived from it. Tensorization facilitates single letterization, which is a desirable property in many capacity determination problems. We introduce hypercontractivity by presenting a famous theorem in the context of Boolean functions. To this end, we first define some pertinent concepts such as p-norms of random variables and noise operators.

Definition 3.1.1 (p-Norm of Random Variables)**.** Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a random variable $X : \Omega \to \mathbb{R}$ on this space. For any $p \in \mathbb{R},\ p \geq 1$, the p-norm of X is given by:

$$\|X\|_p \triangleq \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}}.$$

Intuitively, Definition 3.1.1 [14] parallels the Euclidean notion of p-norm. To define the noise operator, consider a discrete memoryless binary symmetric channel (BSC) with flip-over probability $\frac{1-\rho}{2},\ \rho \in [-1, 1]$. Consider passing a Boolean random vector $X_1^n \in \{-1, 1\}^n$ (recalling the notation in equation 1.41 in section 1.3) through this BSC to get the output random vector Y_1^n. In such a scenario, we may define an entity known as the noise operator [15], which maps Boolean functions to other Boolean functions. The noise operator is characterized by the parameter ρ of the BSC.

Definition 3.1.2 (Noise Operator)**.** Suppose we are given a BSC with parameter $\rho \in [-1, 1]$, with input Boolean random vector X_1^n, and output Boolean random vector Y_1^n. Then, for any input Boolean function $g : \{-1, 1\}^n \to \mathbb{R}$, the noise operator, denoted T_ρ, is defined as:

$$\left(T_\rho g\right)\left(X_1^n\right) \triangleq \mathbb{E}_{P_{Y_1^n|X_1^n}}\left[g(Y_1^n)|X_1^n\right]$$

where T_ρ takes the Boolean function g as input and produces the Boolean function $T_\rho g : \{-1,1\}^n \to \mathbb{R}$ as output.

Operationally, the noise operator smooths out the high frequency components in the Fourier series of g. Note that here, we refer to Fourier analysis on the hypercube rather than the traditional setting of periodic functions. The notion of smoothing by a noise operator is concretely characterized by the hypercontractivity theorem [15].

Theorem 3.1.1 (Hypercontractivity Theorem)**.** *Suppose we are given a BSC with parameter $\rho \in [-1,1]$, with input Boolean random vector X_1^n, and output Boolean random vector Y_1^n. For every $1 \le q \le p$, if $\rho^2(p-1) \le q-1$, then for any Boolean function $g : \{-1,1\}^n \to \mathbb{R}$, we have:*

$$\| \left(T_\rho g\right)\left(X_1^n\right) \|_p \le \| g\left(Y_1^n\right) \|_q.$$

A proof of this theorem can be found in [15]. There are generalizations of hypercontractivity beyond Boolean functions, and [15] presents many such generalizations for the interested reader. In the inequality in Theorem 3.1.1, as $q \le p$, the norm on the left hand side gives more importance to larger values of $\left(T_\rho g\right)\left(X_1^n\right)$. Intuitively, the inequality resembles a Chebyshev or minimax optimization constraint when p is large. The hypercontractivity theorem provides sufficient conditions for the BSC parameter to ensure that larger values of $\left(T_\rho g\right)\left(X_1^n\right)$ are forced below an average of $g\left(Y_1^n\right)$. Hence, the theorem says that noise in a channel distributes out locally clustered peaks of energy in g. This intuition also gives meaning to the name "hypercontractivity."

We now generalize the notion of hypercontractivity for any two random variables (X, Y). In the remainder of this section, we will assume that we have some probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, with the random variables X and Y defined on this space. X and Y will take values in the discrete and finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Moreover, we will denote the joint pmf of (X, Y) as $P_{X,Y}$, and further assume that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$ where necessary. We first define the hypercontractivity ribbon below [16]. This concept will turn out to be deeply intertwined with the global case of problem 3.7 described earlier.

Definition 3.1.3 (Hypercontractivity Ribbon)**.** For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, we define the hypercontractivity ribbon, denoted $\mathcal{R}(X;Y)$, as:

$$\mathcal{R}(X;Y) \triangleq \left\{(p,q) \in \mathbb{R}^2 : 1 \le q \le p \text{ and for all functions } g : \mathcal{Y} \to \mathbb{R},\ \| \mathbb{E}\left[g(Y)|X\right] \|_p \le \|g(Y)\|_q \right\}.$$

We consider the hypercontractivity ribbon because it can be used to define a hypercontractive constant $s^*(X;Y)$ [16]. This is the quantity through which hypercontractivity interacts with information theory. To formally specify this quantity, we follow the exposition of [16] and use their notation. To this end, for any $p \ge 1$, we define:

$$s^{(p)}(X;Y) \triangleq \inf\{r \in \mathbb{R} : (p, pr) \in \mathcal{R}(X;Y)\}. \tag{3.10}$$

Since $s^{(p)}(X;Y)$ is monotonically decreasing in p [16], and bounded below by 0, its limit as $p \to \infty$ is well-defined. We call this limit the hypercontractive constant $s^*(X;Y)$. For fixed p, $s^{(p)}(X;Y)$

is the infimum ratio of $\frac{q}{p}$ such that $(p, q) \in \mathcal{R}(X;Y)$, where we note that $\mathcal{R}(X;Y)$ is a closed and connected set in $\mathbb{R}^2$ [17]. So, $(p, ps^{(p)}(X;Y))$ traces the lower boundary of $\mathcal{R}(X;Y)$ on the (p,q)-plane. $s^*(X;Y)$ is thus the infimum ratio of $\frac{q}{p}$ as $p \to \infty$, and characterizes the lower boundary of the hypercontractivity ribbon in the limit as $p \to \infty$. For this reason, [14] states that $s^*(X;Y)$ is the chordal slope of the boundary of the hypercontractivity ribbon $\mathcal{R}(X;Y)$ at infinity. We formally define the hypercontractive constant in the next definition.

Definition 3.1.4 (Hypercontractive Constant)**.** For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, we define the hypercontractive constant, denoted $s^*(X;Y)$, as:

$$s^*(X;Y) \triangleq \lim_{p \to \infty} s^{(p)}(X;Y).$$

The constant $s^*(X;Y)$ is a bounded quantity. Indeed, as $\mathcal{R}(X;Y) \subseteq \big\{(p,q) \in \mathbb{R}^2 : 1 \leq q \leq p\big\}$, for any $p \geq 1$, $0 \leq s^{(p)}(X;Y) \leq 1$. Hence, we have:

$$0 \leq s^*(X;Y) \leq 1. \tag{3.11}$$

Although we promised that the hypercontractive constant would be information theoretically meaningful, there has been little evidence of this so far. Thus, we provide an alternative and equivalent definition of $s^*(X;Y)$ which is more tangibly meaningful from an information theoretic perspective [14].

Definition 3.1.5 (Hypercontractive Constant)**.** For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we define the hypercontractive constant, denoted $s^*(X;Y)$, as:

$$s^*(X;Y) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where we optimize over all pmfs R_X on $\mathcal{X}$ such that $R_X \neq P_X$. Note that P_X and P_Y are the marginal pmfs of $P_{X,Y}$, and R_Y is the marginal pmf of the joint pmf $R_{X,Y} = P_{Y|X}R_X$. Furthermore, if X or Y is a constant almost surely, we define $s^*(X;Y) = 0$.

The ratio in Definition 3.1.5 resembles problem statement 3.7 introduced earlier. In fact, the ratio can be changed to one with mutual information terms rather than KL divergence terms. This offers another characterization of the hypercontractive constant. The next theorem from [14] presents this characterization.

Theorem 3.1.2 (Mutual Information Characterization of Hypercontractive Constant)**.** *For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$s^*(X;Y) = \sup_{\substack{P_U, P_{X|U} : U \to X \to Y \\ I(U;X) > 0}} \frac{I(U;Y)}{I(U;X)}$$

where $U \to X \to Y$ is a Markov chain, the random variable U takes values over the discrete and finite set $\mathcal{U}$, and we optimize over all possible pmfs P_U on $\mathcal{U}$ and all possible conditional pmfs $P_{X|U}$ so that the marginal pmf of X remains fixed at P_X.

In Theorem 3.1.2, we fix the joint pmf $P_{X,Y}$ and find the optimum P_U and $P_{X|U}$ with a constraint on the marginal P_X. Hence, the result of the optimization is truly a function of $P_{X,Y}$, and the notation $s^*(X;Y)$ captures this by not including the U. [14] provides a proof of this theorem using a geometric characterization of $s^*(X;Y)$, but we omit it here for the sake of brevity. It is worth noting that a supremum rather than a maximum is needed in both Theorem 3.1.2 and Definition 3.1.5. We must constrain the denominator to be strictly positive in both cases for the ratios to be well-defined, but the optimal value of the ratio may occur in the limit as the denominator tends to 0.

The characterization of $s^*(X;Y)$ in Theorem 3.1.2 matches the global case of problem 3.7. Hence, the hypercontractive constant equals the global optimal value of problem 3.7. This was our motivation to study hypercontractivity all along. Conveniently, the preceding discussion has also illustrated some of the wider theoretical significance of the hypercontractive constant. In section 3.3, we will compare $s^*(X;Y)$ to the local optimal value given in equation 3.8.

We now list a few properties of the hypercontractivity ribbon and the hypercontractive constant. Many of these properties can be found in [14], [16], and [17]. In each of the ensuing lemmata, all random variables are defined on the same probability space and take values on discrete and finite sets. Moreover, all probability masses in the marginal pmfs of these random variables are assumed to be strictly positive when required (for the hypercontractive constants to be well-defined). To avoid being pedantic, we do not state these conditions explicitly every time.

Lemma 3.1.3 (Tensorization)**.** *If (X_1, Y_1) and (X_2, Y_2) are independent, then:*

$$\mathcal{R}(X_1, X_2; Y_1, Y_2) = \mathcal{R}(X_1; Y_1) \cap \mathcal{R}(X_2; Y_2) \text{ and } s^*(X_1, X_2; Y_1, Y_2) = \max\{s^*(X_1; Y_1), s^*(X_2; Y_2)\}.$$

Lemma 3.1.4 (Data Processing Inequality)**.** *If the random variables $W \to X \to Y \to Z$ form a Markov Chain, then:*

$$\mathcal{R}(X;Y) \subseteq \mathcal{R}(W;Z) \text{ and } s^*(X;Y) \geq s^*(W;Z).$$

Lemma 3.1.5 (Vanishing Property)**.** *The random variables X and Y are independent if and only if $s^*(X;Y) = 0$.*

Proof.
($\Rightarrow$) If X and Y are independent, then $P_{Y|X} = P_Y$. By Definition 3.1.5:

$$s^*(X;Y) = \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}.$$

For any marginal pmf $R_X \neq P_X$, R_Y is the marginal pmf of Y corresponding to $R_{X,Y} = P_{Y|X}R_X = P_Y R_X$. This means $R_Y = P_Y$, which implies $D(R_Y||P_Y) = 0$. Hence, $s^*(X;Y) = 0$. This can also be shown using the characterization of $s^*(X;Y)$ in Theorem 3.1.2.
($\Leftarrow$) If $s^*(X;Y) = 0$, then by Definition 3.1.5:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} = 0$$

for every pmf $R_X \neq P_X$. This implies $\forall R_X \neq P_X,\ D(R_Y||P_Y) = 0$. So, $\forall R_X \neq P_X,\ R_Y = P_Y$. Let $\mathcal{X}$ and $\mathcal{Y}$ be the finite alphabet sets of X and Y, respectively. For every $i \in \mathcal{X}$, let R_X^i be the pmf such that $R_X^i(i) = 1$ and $\forall x \in \mathcal{X}\backslash\{i\},\ R_X^i(x) = 0$. Then, $\forall y \in \mathcal{Y},\ P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)R_X^i(x) = P_{Y|X}(y|i)$. Since this holds for every $i \in \mathcal{X}$, $P_{Y|X} = P_Y$. Hence, X and Y are independent. □

To avoid digressions, we do not prove Lemmata 3.1.3 and 3.1.4 here. [14] provides two proofs for the tensorization property of $s^*(X;Y)$. The first uses the KL divergence characterization of $s^*(X;Y)$, and the second uses an elegant geometric characterization.

We end this section having identified the hypercontractive constant, $s^*(X;Y)$, as the global optimal solution to problem 3.7, and explored some of its properties. We note that more generally, hypercontractivity has recently re-emerged as a powerful tool in information theory. It has been used in [16] to understand the mutual information between Boolean functions; an endeavor which originates from a conjecture in [18]. It has also been used in [17] to derive impossibility results for non-interactive simulation of joint distributions. While such literature is interesting in its own right, we will not need to delve any further into the depths and subtleties of hypercontractivity for our purposes.

3.2 Hirschfeld-Gebelein-Rényi Maximal Correlation

We now consider characterizing the local optimal value of problem 3.7 given in equation 3.8. In our ensuing discussion along this front, we will require a unique notion of correlation between two random variables X and Y. Recall that in the zero mean and unit variance case, the Pearson correlation coefficient is given by $\mathbb{E}[XY]$. Keeping this in mind, we define a stronger notion of correlation known as the Hirschfeld-Gebelein-Rényi maximal correlation [14], which we will refer to as the Rényi correlation from hereon.

Definition 3.2.1 (Hirschfeld-Gebelein-Rényi Maximal Correlation)**.** Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and jointly distributed random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ on this space such that $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. Then, we define the Hirschfeld-Gebelein-Rényi maximal correlation, or simply Rényi correlation, between X and Y as:

$$\rho(X;Y) \triangleq \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\ g:\mathcal{Y}\to\mathbb{R}\ :\\ \mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0\\ \mathbb{E}[f^2(X)]=\mathbb{E}[g^2(Y)]=1}} \mathbb{E}[f(X)g(Y)]$$

where the supremum is taken over all Borel measurable functions, f and g, subject to the zero mean and unit variance constraints. Furthermore, when one of X or Y is constant almost surely, there exist no functions f and g which satisfy the constraints, and we define $\rho(X;Y) = 0$.

This definition of Rényi correlation naturally extends to cover random vectors or random variables whose ranges are arbitrary measurable spaces. A compelling aspect of Rényi correlation is its ability to summarize the dependence between such random vectors or random variables with non-standard ranges with a single real scalar. We also note that Definition 3.2.1 is characteristic

of Rényi's style of generalizing fundamental metrics in probability and information theory using arbitrary functions. For example, Rényi generalized Shannon entropy to Rényi entropy by looking at the functional equation which forces the information in independent events to add. He changed the linear averaging of individual information terms to averaging functions of these terms and then applying the inverse function. This eventually led to the Rényi entropy family. Correspondingly, Definition 3.2.1 generalizes the Pearson correlation coefficient by finding alternative representations of X and Y by applying functions to them, where the functions are chosen to maximize the Pearson correlation coefficient.

Although we introduce Rényi correlaiton using Definition 3.2.1, in his paper [19], Rényi introduced it by demonstrating that it satisfied seven axioms which "natural" dependence measures should exhibit. We will not belabor these axioms, but develop them from Definition 3.2.1 when the need arises. We will however, take a moment to appreciate some properties of Rényi correlation with the hope that the reader will recognize the tacit parallels with the hypercontractive constant. We commence by noting that:

$$0 \leq \rho(X;Y) \leq 1. \tag{3.12}$$

This is Rényi's third axiom [19]. The upper bound follows from the Cauchy-Schwarz inequality. The lower bound can be argued by realizing that if $\rho(X;Y) < 0$, we may negate f or g (but not both) to get $\rho(X;Y) > 0$, and this contradicts the supremum in Definition 3.2.1. The singular value characterization of $\rho(X;Y)$ (soon to come in Theorem 3.2.4) also trivially implies both bounds, albeit only in the discrete and finite case for which the characterization is proved. We now list a few other properties of Rényi correlation (which hold for general random variables, not just discrete ones). In each of the ensuing lemmata, all random variables are defined on the same probability space although we do not state these conditions explicitly.

Lemma 3.2.1 (Tensorization)**.** *If (X_1, Y_1) and (X_2, Y_2) are independent, then:*

$$\rho(X_1, X_2; Y_1, Y_2) = \max\left\{\rho(X_1, Y_1), \rho(X_2, Y_2)\right\}.$$

Lemma 3.2.2 (Data Processing Inequality)**.** *If the random variables $W \to X \to Y \to Z$ form a Markov Chain, where X and Y take values on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $W = r(X)$ and $Z = s(Y)$ for Borel measurable functions $r : \mathcal{X} \to \mathbb{R}$ and $s : \mathcal{Y} \to \mathbb{R}$, then:*

$$\rho(X;Y) \geq \rho(W;Z).$$

Lemma 3.2.3 (Vanishing Property)**.** *The random variables X and Y are independent if and only if $\rho(X;Y) = 0$.*

These lemmata have been collated in [14] and [17]. In particular, lemmata 3.2.1 and 3.2.2 are proven in [20] and [17], respectively. Lemma 3.2.3 is Rényi's fourth axiom [19]; its forward statement is straightforward to derive from Definition 3.2.1. Indeed, if X and Y are independent, then $f(X)$ and $g(Y)$ are independent for any Borel measurable functions f and g. Thus,

$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)] = 0$, which means $\rho(X;Y) = 0$.

We now motivate the singular value characterization of Rényi correlation which will be pivotal to our arguments in section 3.3. To this end, if we briefly muse on information measures of discrete random variables, it becomes intuitively apparent that they are independent of the values taken by the random variables. This property manifests itself in the definition of discrete Shannon entropy in Definition 1.0.2, for example. Moreover, Definition 3.1.5 makes it clear that the hypercontractive constant, $s^*(X;Y)$, has this property. The Rényi correlation for discrete random variables shares this property as well. In fact, the optimization over all functions in Definition 3.2.1 renders the Rényi correlation independent of the values taken by the random variables in the discrete case. A more direct approach to illustrating that $\rho(X;Y)$ is solely a function of the joint distribution of (X,Y) is to use the singular value characterization of Rényi correlation. This characterization shows that Rényi correlation is the second largest singular value of the DTM for discrete and finite random variables. Recall that given discrete and finite random variables (X,Y), we may interpret X as the source and Y as the output of a channel. Then, letting W denote the column stochastic transition matrix of conditional probabilities $P_{Y|X}$ as shown in equation 1.23, the DTM is given by:

$$B = \left[\sqrt{P_Y}\right]^1 W \left[\sqrt{P_X}\right] \tag{3.13}$$

according to Definition 1.3.4. Assuming without loss of generality that X and Y take values on $\mathcal{X} = \{1,\dots,n\}$ and $\mathcal{Y} = \{1,\dots,m\}$, respectively, it is easy to derive that:

$$\forall y \in \mathcal{Y}, \forall x \in \mathcal{X},\ \ B_{yx} = \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}} \tag{3.14}$$

where B_{yx} is the entry in the yth row and xth column of B. Equation 3.14 illustrates how every entry of B displays a symmetry in X and Y. This offers some credence to the claim that Rényi correlation is the second largest singular value of B, because Rényi correlation is symmetric in its inputs X and Y; this is Rényi's second axiom [19]. The characterization is stated in the next theorem [14]. We also prove it here, because the statement is rather counter-intuitive.

Theorem 3.2.4 (DTM Singular Value Characterization of Rényi Correlation)**.** *Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ on this space such that $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$. Given further that the joint pmf, $P_{X,Y}$, is such that the marginals satisfy $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, the Rényi correlation, $\rho(X;Y)$, is the second largest singular value of the divergence transition matrix (DTM) B.*

Proof.
We use the notation from equations 1.11 and 1.12 in this proof. Moreover, we represent marginal pmfs, P_X and P_Y, as column vectors. Let f and g be the column vectors representing the range of the functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, respectively. From equation 3.13, we have:

$$B = \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{P_X}\right]$$

where the columns of W are conditional pmfs of Y given X.

We first show that the largest singular value of B is 1. Consider $M = \left[\sqrt{P_Y}\right]^{-1} BB^T \left[\sqrt{P_Y}\right]$.

On the one hand, M has the same set of eigenvalues as BB^T, because we are simply using a similarity transformation. As $BB^T \succeq 0$, the eigenvalues of M and BB^T are non-negative real numbers by the Spectral Theorem. On the other hand, we have:

$$M = [P_Y]^{-1} W [P_X] W^T = VW^T$$

where $V = [P_Y]^{-1} W [P_X]$ is the row stochastic reverse transition probability matrix of conditional pmfs $P_{X|Y}$ (i.e. each row of V is a conditional pmf). Since V and W^T are both row stochastic, their product $M = VW^T$ is also row stochastic. Hence, by the Perron-Frobenius Theorem, the largest eigenvalue of M (and thus BB^T) is 1. It follows that the largest singular value of B is 1. Notice further that $\sqrt{P_X}$ and $\sqrt{P_Y}$ are the right and left singular vectors of B, respectively, corresponding to singular value 1. Indeed, we have:

$$B\sqrt{P_X} = \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{P_X}\right] \sqrt{P_X} = \sqrt{P_Y},$$

$$\sqrt{P_Y}^T B = \sqrt{P_Y}^T \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{P_X}\right] = \sqrt{P_X}.$$

Next, note that we can express the expectations in Definition 3.2.1 of Rényi correlation in terms of B, P_X, P_Y, f, and g.

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \left(\left[\sqrt{P_Y}\right] g\right)^T B \left(\left[\sqrt{P_X}\right] f\right) \\
\mathbb{E}[f(X)] &= \left(\left[\sqrt{P_X}\right] f\right)^T \sqrt{P_X} \\
\mathbb{E}[g(Y)] &= \left(\left[\sqrt{P_Y}\right] g\right)^T \sqrt{P_Y} \\
\mathbb{E}\left[f^2(X)\right] &= \left\|\left[\sqrt{P_X}\right] f\right\|^2 \\
\mathbb{E}\left[g^2(Y)\right] &= \left\|\left[\sqrt{P_Y}\right] g\right\|^2
\end{aligned}$$

Letting $a = \left[\sqrt{P_X}\right] f$ and $b = \left[\sqrt{P_Y}\right] g$, we have from Definition 3.2.1:

$$\rho(X;Y) = \sup_{\substack{a,b: \\ a^T\sqrt{P_X}=b^T\sqrt{P_Y}=0 \\ \|\mathbf{a}\|^2=\|\mathbf{b}\|^2=1}} b^T B a.$$

Since $a^T\sqrt{P_X} = b^T\sqrt{P_Y} = 0$, a is orthogonal to the right singular vector of B corresponding to singular value 1, and b is orthogonal to the left singular vector of B corresponding to singular value 1. Hence, the above maximization clearly gives the second largest singular value of B as a and b are normalized. This completes the proof. □

It is worth mentioning that Rényi correlation is not the only measure of dependence based on singular values of the DTM. The authors of [21] define the k-correlation family by taking kth Ky Fan norms (sum of k largest singular values) of BB^T minus the first singular value of 1. In particular, this means that the squared Rényi correlation and the standard squared Schatten 2-norm minus 1

are both k-correlations.

Recall from equation 3.8 that the squared second largest singular value of B is the local optimal value of problem 3.7. Theorem 3.2.4 characterizes this local optimal value as the squared Rényi correlation. So, we have identified both local and global optimal values of problem 3.7 as fundamental quantities known in the information theory literature. Moreover, like the hypercontractive constant, the Rényi correlation has characterizations in terms of both mutual information and KL divergence. Equation 3.8 is its mutual information characterization. We now present a corollary of Theorem 3.2.4, which illustrates that $\rho(X;Y)$ can also be interpreted as a kind of second largest singular value of the channel matrix W when appropriate norms are used. This is essentially the KL divergence characterization. We note that notation from Definition 1.1.1 is used in the corollary.

Corollary 3.2.5 (Channel Characterization of Rényi Correlation)**.** *Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ on this space such that $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$, with marginal pmfs satisfying $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$. Letting W by the column stochastic matrix of conditional pmfs $P_{Y|X}$, the squared Rényi correlation is given by:*

$$\rho^2(X;Y) = \sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2}$$

where 1 denotes the vector of all ones, and we optimize over all valid additive perturbations of J_X.

Proof.
From Theorem 3.2.4, we have:

$$\rho^2(X;Y) = \sup_{\substack{K_X : K_X \neq 0 \\ \sqrt{P_X}^T K_X = 0}} \frac{\|BK_X\|^2}{\|K_X\|^2} = \sup_{\substack{K_X : K_X \neq 0 \\ \sqrt{P_X}^T K_X = 0}} \frac{\left\| \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{P_X}\right] K_X \right\|^2}{\left\| \left[\sqrt{P_X}\right]^{-1} \left[\sqrt{P_X}\right] K_X \right\|^2}$$

where we let $J_X = \left[\sqrt{P_X}\right] K_X$ to get:

$$\rho^2(X;Y) = \sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2}$$

as required. □

As mentioned earlier, Corollary 3.2.5 is actually a local KL divergence characterization of Rényi correlation. Under the assumptions of Corollary 3.2.5, let $R_X = P_X + \epsilon J_X$, where J_X is a valid additive perturbation and $\epsilon > 0$ is small enough so that R_X is a valid pmf. Moreover, let $R_Y = WR_X$. Then, from the discussion in chapter 1, we know that:

$$D(R_X||P_X) = \frac{1}{2}\epsilon^2 \|J_X\|_{P_X}^2 + o\left(\epsilon^2\right)$$

$$D(R_Y||P_Y) = \frac{1}{2}\epsilon^2 \|WJ_X\|_{P_Y}^2 + o\left(\epsilon^2\right)$$

which means that under local approximations:

$$\rho^2(X;Y) \overset{\text{local}}{=} \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \tag{3.15}$$

using Corollary 3.2.5. Notice that the right hand side of equation 3.15 is actually the hypercontractive constant (without approximations). Hence, under local approximations, the hypercontractive constant becomes the squared Rényi correlation regardless of whether we use the KL divergence or mutual information characterization (equations 3.15 and 3.8, respectively). The precise sense in which equations 3.8 and 3.15 are valid is accentuated in section 3.3. All in all, the singular value characterizations of Rényi correlation in Theorem 3.2.4 and Corollary 3.2.5 have many advantages. For example, computing the Rényi correlation is much easier when we find a singular value using linear algebra tools rather than solving the unnerving maximization in Definition 3.2.1. Moreover, such characterizations use the representation of singular values as extremal problems. This makes Rényi correlation a supremum of a ratio, which parallels the hypercontractive constant being a supremum of a ratio. This observation will be crucial in fulfilling the agenda of the next section: bounding the Rényi correlation and the hypercontractive constant with each other.

Before we end this section, we present a final theorem which finds necessary conditions on the optimizing functions f^* and g^* in the definition of Rényi correlation. This provides a deeper intuition on how $\rho(X;Y)$ measures the dependence between random variables.

Theorem 3.2.6 (MMSE Characterization of Rényi Correlation)**.** *Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with joint distribution $P_{X,Y}$ on this space, and Rényi correlation $\rho(X;Y)$. If the optimizing functions of the Rényi correlation are $f^* : \mathcal{X} \to \mathbb{R}$ and $g^* : \mathcal{Y} \to \mathbb{R}$, then f^* and g^* satisfy:*

$$\begin{aligned} \rho(X;Y) f^*(X) &= \mathbb{E}\left[g^*(Y)|X\right] \quad a.s. \\ \rho(X;Y) g^*(Y) &= \mathbb{E}\left[f^*(X)|Y\right] \quad a.s. \end{aligned}$$

where the equalities hold almost surely (with probability 1), and $\rho^2(X;Y) = \mathbb{E}\left[\mathbb{E}\left[f^*(X)|Y\right]^2\right] = \mathbb{E}\left[\mathbb{E}\left[g^*(Y)|X\right]^2\right]$.

Note that f^* and g^* are assumed to satisfy the zero mean, $\mathbb{E}\left[f^*(X)\right] = \mathbb{E}\left[g^*(Y)\right] = 0$, and unit variance, $\mathbb{E}\left[f^*(X)^2\right] = \mathbb{E}\left[g^*(Y)^2\right] = 1$, conditions as part of the premise of being valid optimizing functions for the Rényi correlation. A proof of Theorem 3.2.6 can be found in [19]. We provide an alternative proof using variational calculus techniques rather than statistical methods in Appendix A. The relation $\rho^2(X;Y) = \mathbb{E}\left[\mathbb{E}\left[f^*(X)|Y\right]^2\right] = \mathbb{E}\left[\mathbb{E}\left[g^*(Y)|X\right]^2\right]$ can be found in [19] and [14], but is derived in the appendix as well for completeness. We construe the above theorem as the minimum mean-square error (MMSE) characterization of Rényi correlation because $\mathbb{E}\left[g^*(Y)|X = x\right]$ is the MMSE estimator of $g^*(Y)$ given $X = x$, and $\mathbb{E}\left[f^*(X)|Y = y\right]$ is the MMSE estimator of $f^*(X)$ given $Y = y$. Hence, the theorem states that the optimizing function $f^*(x)$ (or $g^*(y)$) is the MMSE estimator of $g^*(Y)$ (or $f^*(X)$) given $X = x$ (or $Y = y$), normalized to have zero mean and unit variance. So, it unveils how Rényi correlation elegantly maximizes the correlation using MMSE estimation. It also illustrates the inherent coupling between the optimizing functions. Finally, we note that $f^*(x)$ and $g^*(y)$ may not be unique. Indeed, we may negate both functions to get the same $\rho(X;Y) \geq 0$. Such intuition from Theorem 3.2.6 will be valuable in section 3.4.

3.3 Discrete and Finite Case

We next consider assessing the performance of algorithms which employ the local approximation of KL divergence as in the linear information coupling problem. In this section, we assume that the random variables $U : \Omega \to \mathcal{U}$, $X : \Omega \to \mathcal{X}$, and $Y : \Omega \to \mathcal{Y}$ are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that $|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{U}| < \infty$, which means the random variables are discrete and take values on finite sets. Furthermore, we assume that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$. This ensures we can use P_X and P_Y as reference pmfs in local analysis, because they are in the relative interior of their respective probability simplexes. For the sake of clarity, we do not restate these assumptions in every theorem statement in this section.

Recall from Definition 3.1.5 and Theorem 3.1.2 that the hypercontractive constant can be characterized as:

$$s^*(X;Y) = \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \sup_{\substack{P_U, P_{X|U} : U \to X \to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)}. \tag{3.16}$$

On the other hand, using equation 3.8, Theorem 3.2.4, and equation 3.15, we have:

$$\rho^2(X;Y) \overset{\text{local}}{=} \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \sup_{\substack{P_U, P_{X|U} : U \to X \to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \tag{3.17}$$

where the equalities hold under local approximations. Equation 3.17 represents the local optimal value of problem 3.7. From our discussion at the outset of this chapter, we know that the Rényi correlation captures the optimal performance of algorithms using the local linear information coupling framework. On the other hand, equation 3.16 corresponds to the global optimal value of problem 3.7. Hence, comparing the Rényi correlation with the hypercontractive constant will indicate how effectively local algorithms perform. More concretely, we will upper and lower bound the hypercontractive constant with Rényi correlation to reveal how close the local and global optimal values are to each other.

3.3.1 Relationship between Hypercontractive Constant and Rényi Correlation

As promised, we will first rigorously establish the sense in which equation 3.17 holds. This analysis will illustrate that the hypercontractive constant is lower bounded by the squared Rényi correlation. To this end, we first consider the work in [22] which directly addresses problem 3.7. In part E of section IV [22], the authors analyze the function:

$$\forall R \geq 0,\ \ \Delta(R) \triangleq \sup_{\substack{P_U, P_{X|U} : U \to X \to Y \\ I(U;X)=R}} I(U;Y) \tag{3.18}$$

where $P_{X,Y}$ is fixed. $\Delta(R)$ has significance in understanding investment in the horse race market. In Theorem 8 [22], the authors compute the right derivative of this function at $R = 0$:

$$\left.\frac{d\Delta}{dR}\right|_{R=0} = \lim_{\epsilon \to 0^+} \frac{\Delta(\epsilon) - \Delta(0)}{\epsilon} = \lim_{\epsilon \to 0^+} \sup_{\substack{P_U, P_{X|U} : U \to X \to Y \\ I(U;X)=\epsilon}} \frac{I(U;Y)}{I(U;X)} \tag{3.19}$$

where the first equality holds by definition of right derivative (assuming the limit exists), and the second equality holds using equation 3.18 and the fact that $\Delta(0) = 0$. Moreover, we may rewrite equation 3.19 as:

$$\left.\frac{d\Delta}{dR}\right|_{R=0} = \lim_{I(U;X)\to 0^+} \sup_{P_U, P_{X|U}: U\to X\to Y} \frac{I(U;Y)}{I(U;X)} = \sup_{\substack{P_U, P_{X|U}: U\to X\to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} = s^*(X;Y) \tag{3.20}$$

where the second equality holds because $\Delta(R)$ is concave in R, $\Delta(R) \geq 0$, and $\Delta(0) = 0$ [22]. This is illustrated in Figure 3.1. The red line is the tangent to $\Delta(R)$ at $R = 0$. So, its gradient is given by equation 3.20. The gradients of the blue lines are various values of:

$$\sup_{\substack{P_U, P_{X|U}: U\to X\to Y \\ I(U;X)=R}} \frac{I(U;Y)}{I(U;X)}$$

for different values of R. Figure 3.1 portrays that:

$$\forall R > 0, \quad \left.\frac{d\Delta}{dR}\right|_{R=0} \geq \sup_{\substack{P_U, P_{X|U}: U\to X\to Y \\ I(U;X)=R}} \frac{I(U;Y)}{I(U;X)}$$

which produces equation 3.20 after taking suprema over $R > 0$ on both sides. Equation 3.20 demonstrates that the supremum in the mutual information characterization of $s^*(X;Y)$ is achieved when $I(U;X) \to 0$. This is why having a supremum (instead of a maximum) in Theorem 3.1.2 is essential.

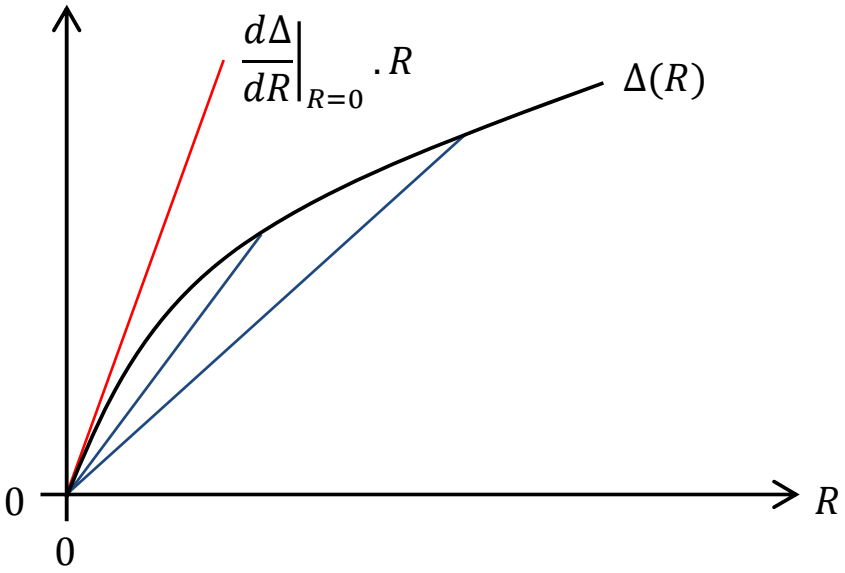

Figure 3.1: Plot of $\Delta(R)$ illustrating its salient features. The red line indicates the tangent of $\Delta(R)$ at $R = 0$, and the gradients of the blue lines are less than the gradient of the red line.

As correctly deduced in [14], the authors of [22] erroneously conclude that $\frac{d\Delta}{dR}\big|_{R=0} = \rho^2(X;Y)$. This is because they compute:

$$\left.\frac{d\Delta}{dR}\right|_{R=0} = \lim_{\epsilon\to 0^+} \sup_{\substack{P_{U|X}:U\to X\to Y \\ I(U;X)=\epsilon}} \frac{I(U;Y)}{I(U;X)} \tag{3.21}$$

by optimizing over $P_{U|X}$ (instead of $P_U, P_{X|U}$) since $P_{X,Y}$ is given. Recall that we have:

$$\epsilon = I(U;X) = \sum_{u\in\mathcal{U}} P_U(u) D(P_{X|U=u}||P_X) \tag{3.22}$$

$$= \sum_{x\in\mathcal{X}} P_X(x) D(P_{U|X=x}||P_U) \tag{3.23}$$

where the additional constraint $I(U;X) = \epsilon$ comes from equation 3.21. The authors of [22] show that the constraint $I(U;X) = \epsilon$ is equivalent to assuming the conditional pmfs, $P_{U|X}$, are local perturbations of P_U. This can be understood from equation 3.23. Since P_X is fixed, there is a bound on how much the KL divergences of equation 3.23 can vary. This bounds how much the conditional pmfs, $P_{U|X}$, can vary from P_U by Pinsker's inequality (which we will encounter soon). Propelled by this local perturbation view, the authors of [22] use Taylor approximations of conditional entropy terms with respect to P_U. [14] explains that this is incorrect because P_U may not be in the relative interior of the probability simplex. This may cause the first derivative term to be infinity, thereby rendering the Taylor approximation invalid.

Our local approximations are fundamentally different. We optimize the ratio in equation 3.21 over $P_U, P_{X|U}$ keeping P_X fixed. This means that we assume the condtional pmfs, $P_{X|U}$, are local perturbations of P_X, as shown in equation 3.2. Since P_X is fixed inside the relative interior of the probability simplex, our Taylor approximations are valid. However, a constraint like $I(U;X) = \epsilon$ is no longer equivalent to our local perturbation assumption. Indeed, from equation 3.22 it is evident that letting some $P_U(u)$ become very small and the corresponding $D(P_{X|U=u}||P_X)$ become very large will not violate the $I(U;X) = \epsilon$ constraint. In fact, proving such an equivalence requires $\min\{P_U(u) : u \in \mathcal{U},\ P_U(u) \neq 0\}$ to be well-defined (as can be inferred from [22]). This is no longer well-defined in our case as P_U is not fixed.

The previous discussion should convince the readers that our local approximation technique is mathematically sound and does not exhibit the pitfalls emphasized in [14]. Moreover, our local approximations do produce $\rho^2(X;Y)$ as the optimal value of problem 3.7. From Theorem 3.1.2, we have:

$$s^*(X;Y) = \sup_{\substack{P_U,P_{X|U}:U\to X\to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \geq \sup_{\substack{P_U,P_{X|U}:U\to X\to Y \\ \forall u\in\mathcal{U},\ P_{X|U=u}=P_X+\epsilon\left[\sqrt{P_X}\right]K_u \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)} \tag{3.24}$$

where $\{K_u, u \in \mathcal{U}\}$ are valid normalized perturbations, and the inequality follows from the additional perturbation constraints. Using the expression for local KL divergence in equation 1.15 and

our derivations in section 1.3, we have:

$$\begin{aligned}\sup_{\substack{P_U,P_{X|U}:U\to X\to Y\\ \forall u\in\mathcal{U},\ P_{X|U=u}=P_X+\epsilon\left[\sqrt{P_X}\right]K_u\\ I(U;X)>0}}\frac{I(U;Y)}{I(U;X)} &= \sup_{\substack{P_U,P_{X|U}:U\to X\to Y\\ \forall u\in\mathcal{U},\ P_{X|U=u}=P_X+\epsilon\left[\sqrt{P_X}\right]K_u\\ \exists u\in\mathcal{U},\ P_U(u)>0\wedge K_u\neq 0}}\frac{\sum_{u\in\mathcal{U}}P_U(u)\left\|BK_u\right\|^2+\frac{o(\epsilon^2)}{\epsilon^2}}{\sum_{u\in\mathcal{U}}P_U(u)\left\|K_u\right\|^2+\frac{o(\epsilon^2)}{\epsilon^2}}\\ &= \sup_{\substack{P_U,\{K_u,u\in\mathcal{U}\}:\\ \exists u\in\mathcal{U},\ P_U(u)>0\wedge K_u\neq 0}}\frac{\sum_{u\in\mathcal{U}}P_U(u)\left\|BK_u\right\|^2+o(1)}{\sum_{u\in\mathcal{U}}P_U(u)\left\|K_u\right\|^2+o(1)} \qquad (3.25)\end{aligned}$$

where B is the DTM, and the second equality holds because we are only optimizing over valid normalized perturbations $\{K_u, u\in\mathcal{U}\}$ and P_U such that the marginal pmf P_X is fixed. Note that $o(1)$ denotes functions which satisfy $\lim_{\epsilon\to 0^+} o(1)=0$. Letting $\epsilon\to 0^+$ on both sides produces:

$$\begin{aligned}\lim_{\epsilon\to 0^+}\sup_{\substack{P_U,P_{X|U}:U\to X\to Y\\ \forall u\in\mathcal{U},P_{X|U=u}=P_X+\epsilon\left[\sqrt{P_X}\right]K_u\\ I(U;X)>0}}\frac{I(U;Y)}{I(U;X)} &= \lim_{\epsilon\to 0^+}\sup_{\substack{P_U,\{K_u,u\in\mathcal{U}\}:\\ \exists u\in\mathcal{U},\ P_U(u)>0\wedge K_u\neq 0}}\frac{\sum_{u\in\mathcal{U}}P_U(u)\left\|BK_u\right\|^2+o(1)}{\sum_{u\in\mathcal{U}}P_U(u)\left\|K_u\right\|^2+o(1)}\\ &\geq \sup_{\substack{P_U,\{K_u,u\in\mathcal{U}\}:\\ \exists u\in\mathcal{U},\ P_U(u)>0\wedge K_u\neq 0}}\lim_{\epsilon\to 0^+}\frac{\sum_{u\in\mathcal{U}}P_U(u)\left\|BK_u\right\|^2+o(1)}{\sum_{u\in\mathcal{U}}P_U(u)\left\|K_u\right\|^2+o(1)}\\ &= \sup_{\substack{P_U,\{K_u,u\in\mathcal{U}\}:\\ \exists u\in\mathcal{U},\ P_U(u)>0\wedge K_u\neq 0}}\frac{\sum_{u\in\mathcal{U}}P_U(u)\left\|BK_u\right\|^2}{\sum_{u\in\mathcal{U}}P_U(u)\left\|K_u\right\|^2}\\ &= \rho^2(X;Y) \qquad (3.26)\end{aligned}$$

where the second line follows from the minimax inequality, and the last equality follows from inequality 3.5 (which we showed was tight). To see that the second line is indeed the minimax inequality, we can first replace the $\lim_{\epsilon\to 0^+}$ with $\liminf_{\epsilon\to 0^+}$, and then recognize that $\liminf_{\epsilon\to 0^+}$ is an asymptotic infimum. Note that we also have:

$$\begin{aligned}s^*(X;Y) &= \lim_{\epsilon\to 0^+}\sup_{\substack{P_U,P_{X|U}:U\to X\to Y\\ I(U;X)=\frac{1}{2}\epsilon^2}}\frac{I(U;Y)}{I(U;X)}\\ &\geq \lim_{\epsilon\to 0^+}\sup_{\substack{P_U,P_{X|U}:U\to X\to Y\\ \forall u\in\mathcal{U},\ P_{X|U=u}=P_X+\epsilon\left[\sqrt{P_X}\right]K_u\\ I(U;X)>0}}\frac{I(U;Y)}{I(U;X)}\\ &\geq \rho^2(X;Y) \qquad (3.27)\end{aligned}$$

where the first equality holds due to equations 3.19 and 3.20, the second inequality holds from equation 3.24, and the third inequality is the minimax inequality. This precisely characterizes the sense in which the mutual information based part of equation 3.17 holds. Indeed, to go from the global optimal solution of problem 3.7 (which is $s^*(X;Y)$) to the local optimal solution (which is $\rho^2(X;Y)$), we impose local perturbation constraints under the supremum instead of fixing $I(U;X)$ to be small, and then take the limit of the ratio of mutual informations before computing the supremum. We restate equation 3.27 below as a theorem.

Theorem 3.3.1 (Lower Bound on Hypercontractive Constant)**.** *For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$s^*(X;Y) \geq \rho^2(X;Y).$$

Proof.
Although we have already proven this theorem in the preceding discussion, we provide a separate proof using the KL divergence characterization of the hypercontractive constant. By Definition 3.1.5:

$$s^*(X;Y) = \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \geq \sup_{\substack{R_X : R_X = P_X + \epsilon J_X \\ J_X \neq 0}} \frac{\|WJ_X\|_{P_Y}^2 + o(1)}{\|J_X\|_{P_X}^2 + +o(1)}$$

where $o(1)$ denotes functions which satisfy $\lim_{\epsilon \to 0^+} o(1) = 0$, W is the column stochastic channel transition matrix as defined in equation 1.23, and we optimize the final expression over all valid additive perturbations J_X. The inequality holds because we have added the additional local perturbation constraint. As before, taking limits of both sides gives:

$$s^*(X;Y) \geq \lim_{\epsilon \to 0^+} \sup_{J_X : J_X \neq 0} \frac{\|WJ_X\|_{P_Y}^2 + o(1)}{\|J_X\|_{P_X}^2 + o(1)} \geq \sup_{J_X : J_X \neq 0} \lim_{\epsilon \to 0^+} \frac{\|WJ_X\|_{P_Y}^2 + o(1)}{\|J_X\|_{P_X}^2 + o(1)}$$

where the second inequality is the minimax inequality. Hence, we have:

$$s^*(X;Y) \geq \sup_{J_X : J_X \neq 0} \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2} = \rho^2(X;Y)$$

using Corollary 3.2.5. This completes the proof. □

This proof clarifies the KL divergence based part of equation 3.17. Theorem 3.3.1 is also proven in [17] using perturbation arguments. However, our derivations have a different flavor; they use the local perturbation concepts pertinent to our context, and clarify why our approximations are valid while those of [22] are not. We next observe that the inequality in Theorem 3.3.1 is tight and equality can be achieved. To see this, consider a doubly symmetric binary source (DSBS) with parameter $0 \leq \alpha \leq 1$. A DSBS describes a joint distribution of two binary random variables, X and Y, both defined on $\{0, 1\}$. In particular, a DSBS(α) represents a uniform Bernoulli input random variable X passing through a binary symmetric channel with crossover probability α to produce a uniform Bernoulli output random variable Y. As mentioned in [16], for $(X, Y) \sim$ DSBS(α):

$$s^*(X;Y) = \rho^2(X;Y) = (1 - 2\alpha)^2 \tag{3.28}$$

where $\rho^2(X;Y) = (1 - 2\alpha)^2$ can be readily computed using the singular value characterization of Rényi correlation in Theorem 3.2.4. As a final remark on Theorem 3.3.1, we explicate why $s^*(X;Y)$ is more appropriately compared with $\rho^2(X;Y)$ rather than $\rho(X;Y)$. This is because $s^*(X;Y)$ is a ratio between KL divergences and KL divergences behave like squared distances between distributions, as is evident from Pythagoras' theorem in information geometry. Thus, a squared correlation offers the right kind of comparison.

So far, we have characterized the precise sense in which Rényi correlation is the local optimal solution to problem 3.7, and derived (as expected) that it is less than or equal to the hypercontractive constant, which is the global optimal solution. We now portray the significance of these two quantities in terms of the Data Processing Inequality (DPI). To this end, we state the DPIs for KL divergence and mutual information below [2].

Theorem 3.3.2 (Data Processing Inequality for KL Divergence)**.** *For a fixed channel transition probability kernel $P_{Y|X}$, given marginal input distributions P_X and R_X, and marginal output distributions $P_Y = P_{Y|X}P_X$ and $R_Y = P_{Y|X}R_X$, we have:*

$$D(R_Y||P_Y) \leq D(R_X||P_X).$$

Theorem 3.3.3 (Data Processing Inequality for Mutual Information)**.** *Given the random variables U, X, and Y in a common probability space, such that $U \to X \to Y$ forms a Markov chain, we have:*

$$I(U;Y) \leq I(U;X).$$

When considering the DPIs, a natural question arises concerning the tightness of the inequalities when $P_{X,Y}$ is kept fixed. We now interpret the operational meanings of the hypercontractive constant and the Rényi correlation in terms of the tightness of the DPIs. From equation 3.16, we see that the hypercontractive constant, $s^*(X;Y)$, is the tightest factor we can insert into both DPIs. In other words, tighter DPIs for fixed P_X and $P_{Y|X}$ are:

$$D(R_Y||P_Y) \leq s^*(X;Y)D(R_X||P_X) \tag{3.29}$$

for the KL divergence case, and:

$$I(U;Y) \leq s^*(X;Y)I(U;X) \tag{3.30}$$

for the mutual information case. Conceptually, this gracefully unifies the two DPIs by characterizing the hypercontractive constant as the common tightest factor that can be inserted into either of them. Likewise, using equation 3.17, we see that the squared Rényi correlation is the tightest factor that can be inserted into both DPIs under local approximations. Hence, for fixed P_X and $P_{Y|X}$, we have:

$$D(R_Y||P_Y) \overset{\text{local}}{\leq} \rho^2(X;Y)D(R_X||P_X) \tag{3.31}$$

for the KL divergence case, and:

$$I(U;Y) \overset{\text{local}}{\leq} \rho^2(X;Y)I(U;X) \tag{3.32}$$

for the mutual information case, where the precise meaning of $\overset{\text{local}}{\leq}$ is exactly the sense in which Rényi correlation is the local optimal solution to problem 3.7.

Therefore, the hypercontractive constant depicts the largest fraction of information that can be

sent down a Markov chain in the global case, and the squared Rényi correlation also depicts this under local approximations. Furthermore, we know the intuitive fact that $s^*(X;Y) \geq \rho^2(X;Y)$ from Theorem 3.3.1. By associating the squared Rényi correlation to the optimal performance of algorithms which use local approximations along the linear information coupling framework, we effectively measure performance with respect to how well such algorithms preserve information down a Markov chain. Hence, our performance analysis is based on analyzing the tightness of the DPIs. $s^*(X;Y)$ represents the maximum amount of information that can be preserved, and our local algorithms evidently fall short of this.

3.3.2 KL Divergence Bounds using χ^2-Divergence

Having proved Theorem 3.3.1, we now seek to find an upper bound on the hypercontractive constant using the squared Rényi correlation. This will limit how much worse local algorithms can perform with respect to the globally optimal preservation of information down a Markov chain. The tightness of this bound will essentially assess the quality of the local approximation of KL divergence. The central idea to upper bound $s^*(X;Y)$ is to begin with Definition 3.1.5, upper and lower bound the KL divergences with appropriate χ^2-divergences (which are local KL divergences), and finally use the singular value characterization of Rényi correlation in Corollary 3.2.5.

We first consider lower bounding KL divergence with the χ^2-divergence. This can be accomplished using two different approaches. The first approach is more statistical in flavor. It hinges around using Pinsker's inequality, which is perhaps the most well-known lower bound on KL divergence using a norm. The next lemma presents our first lower bound on KL divergence.

Lemma 3.3.4 (KL Divergence Lower Bound)**.** *Given pmfs P_X and R_X on the discrete and finite alphabet $\mathcal{X}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $R_X = P_X + J_X$, where J_X is a valid additive perturbation, we have:*

$$D(R_X||P_X) \geq \frac{\|J_X\|_{P_X}^2}{2\left(\sum_{x\in\mathcal{X}} \frac{1}{P_X(x)}\right)} = \frac{\chi^2\left(R_X, P_X\right)}{2\left(\sum_{x\in\mathcal{X}} \frac{1}{P_X(x)}\right)}.$$

Proof.
By Pinsker's inequality, we have:

$$D(R_X||P_X) \geq 2D_{TV}^2(R_X, P_X)$$

where $D_{TV}(R_X, P_X) \triangleq \sup_{A\subseteq 2^{\mathcal{X}}} |R_X(A) - P_X(A)|$ is the total variation distance. For a discrete and finite alphabet $\mathcal{X}$, the total variation distance can be written in terms of the 1-norm:

$$D_{TV}(R_X, P_X) = \frac{1}{2}\|R_X - P_X\|_1 = \frac{1}{2}\sum_{x\in\mathcal{X}} |R_X(x) - P_X(x)|.$$

Hence, we have:

$$D(R_X||P_X) \geq \frac{1}{2}\|J_X\|_1^2 \geq \frac{1}{2}\|J_X\|_2^2$$

where the second inequality holds because the 1-norm of a finite dimensional vector is greater than or equal to its 2-norm. Next, from the Cauchy-Schwarz inequality, we get:

$$\|J_X\|_{P_X}^2 = \left\| \left[\sqrt{P_X}\right]^{-1} J_X \right\|_2^2 \leq \left\| \left[\sqrt{P_X}\right]^{-1} \right\|_{\text{Fro}}^2 \|J_X\|_2^2 = \left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right) \|J_X\|_2^2$$

where $\|\cdot\|_{\text{Fro}}$ denotes the Frobenius norm of a matrix. This implies:

$$\|J_X\|_2^2 \geq \frac{\|J_X\|_{P_X}^2}{\left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right)}$$

which gives:

$$D(R_X||P_X) \geq \frac{\|J_X\|_{P_X}^2}{2\left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right)} = \frac{\chi^2\left(R_X, P_X\right)}{2\left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right)}$$

as claimed. □

The second approach to proving Lemma 3.3.4 has a convex analysis flavor. It involves recognizing that KL divergence is a Bregman divergence associated with the negative Shannon entropy function. Then, the convexity properties of the negative Shannon entropy function can be used to bound KL divergence. This alternative proof of Lemma 3.3.4 is provided next.

Proof.
Without loss of generality, let $\mathcal{X} = \{1, \ldots, n\}$. Moreover, let $\mathcal{P} \subseteq \mathbb{R}^n$ be the probability simplex in $\mathbb{R}^n$. Then, recall from equation 2.22 that the negative Shannon entropy function is:

$$\forall P = [p_1 \ \ \ldots \ \ p_n]^T \in \mathcal{P}, \ \ H_-(P) = \sum_{i=1}^{n} p_i \log\left(p_i\right)$$

and its corresponding Bregman divergence (Definition 2.2.1) is the KL divergence:

$$\forall P \in \mathcal{P}, \forall Q \in \text{relint(P)}, \ \ D(P||Q) = H_-(P) - H_-(Q) - \nabla H_-(Q)^T(P - Q).$$

Next, we recall from equation 2.23 that H_- is twice differentiable on $\text{relint}(\mathcal{P})$:

$$\forall P \in \text{relint}(\mathcal{P}), \ \ \nabla^2 H_-(P) = [P]^{-1} \succeq I$$

where $\succeq$ denotes the Löwner partial order on symmetric matrices, I is the identity matrix, and $[P]^{-1} - I$ is positive semidefinite because it is a diagonal matrix with non-negative diagonal entries (as the elements of P are between 0 and 1). From chapter 9 of [13], we know that a twice continuously differentiable convex function $f : S \to \mathbb{R}$, where the domain $S \subseteq \mathbb{R}^n$ is open, is called strongly convex if and only if $\exists m > 0$ such that $\forall x \in S, \ \nabla^2 f(x) \succeq mI$. This means that H_- is a strongly convex function on $\text{relint}(\mathcal{P})$. A consequence of strong convexity is the following quadratic lower bound [13]:

$$\forall P \in \mathcal{P}, \forall Q \in \text{relint}(\mathcal{P}), \ \ H_-(P) \geq H_-(Q) + \nabla H_-(Q)^T(P - Q) + \frac{1}{2}\|P - Q\|_2^2$$

where we allow $P \in \mathcal{P} \backslash \text{relint}(\mathcal{P})$ due to the continuity of H_-. This gives us:

$$\forall P \in \mathcal{P}, \forall Q \in \text{relint}(\mathcal{P}), \;\; D(P||Q) \geq \frac{1}{2} \|P - Q\|_2^2 .$$

Hence, for $P_X \in \text{relint}(\mathcal{P})$ and $R_X \in \mathcal{P}$, we have:

$$D(R_X||P_X) \geq \frac{1}{2} \|J_X\|_2^2$$

which is precisely what we had in the previous proof after loosening Pinsker's inequality using the fact that 1-norm of a finite dimensional vector is greater than or equal to its 2-norm. Hence, the rest of this proof is identical to the previous proof. □

We note that unfortunately, such a Bregman divergence and convexity based approach cannot be used to easily derive an upper bound on KL divergence. It is known that if $\exists r > 0$ such that $\forall P \in \text{relint}(\mathcal{P}), \;\; \nabla^2 H_-(P) \preceq rI$, or if ∇H_- is Lipschitz continuous on $\text{relint}(\mathcal{P})$, then a quadratic upper bound on H_- can be derived [13]. However, the natural logarithm is not Lipschitz continuous on the domain $(0, \infty)$. So, such approaches do not work. Returning to our first proof of Lemma 3.3.4, we see that it should be possible to tighten our bound after the use of Pinsker's inequality. A variant of Lemma 3.3.4 is presented next, which essentially uses Hölder's inequality instead of the norm inequality and Cauchy-Schwarz inequality to get a tighter bound.

Lemma 3.3.5 (KL Divergence Tighter Lower Bound)**.** *Given distinct pmfs P_X and R_X on the discrete and finite alphabet $\mathcal{X}$, such that $\forall x \in \mathcal{X}, \;\; P_X(x) > 0$ and $R_X = P_X + J_X$, where J_X is a valid additive perturbation, we have:*

$$D(R_X||P_X) \geq \frac{\|J_X\|_{P_X}^4}{2\left(\max\limits_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right)^2} = \frac{\chi^2 \left(R_X, P_X\right)^2}{2\left(\max\limits_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right)^2}.$$

Proof.
As in the first proof of Lemma 3.3.4, from Pinsker's inequality, we have:

$$D(R_X||P_X) \geq \frac{1}{2} \|J_X\|_1^2 .$$

Next, note that:

$$\|J_X\|_{P_X}^2 = \sum_{x \in \mathcal{X}} |J_X(x)| \left|\frac{J_X(x)}{P_X(x)}\right| \leq \left(\max_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right) \sum_{x \in \mathcal{X}} |J_X(x)| = \left(\max_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right) \|J_X\|_1 .$$

This can also be obtained from Hölder's inequality:

$$\|J_X\|_{P_X}^2 = \sum_{x \in \mathcal{X}} |J_X(x)| \left|\frac{J_X(x)}{P_X(x)}\right| \leq \|J_X\|_p \left\|[P_X]^{-1} J_X\right\|_q$$

where $p, q \in (1, \infty)$ are the Hölder conjugates which satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Taking the limit of the right hand side as $p \to 1^+$, we have:

$$\|J_X\|_{P_X}^2 \leq \lim_{p \to 1^+} \|J_X\|_p \left\|[P_X]^{-1} J_X\right\|_q = \lim_{p \to 1^+} \|J_X\|_p \lim_{q \to \infty} \left\|[P_X]^{-1} J_X\right\|_q = \|J_X\|_1 \left\|[P_X]^{-1} J_X\right\|_\infty$$

Hence, we have:

$$\|J_X\|_{P_X}^2 \leq \|J_X\|_1 \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)$$

as before. This gives us:

$$\|J_X\|_1 \geq \frac{\|J_X\|_{P_X}^2}{\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right|}$$

where the denominator on the right hand side is strictly positive as $R_X \neq P_X$. Using the above inequality and the result from Pinsker's inequality: $D(R_X||P_X) \geq \frac{1}{2} \|J_X\|_1^2$, we get:

$$D(R_X||P_X) \geq \frac{\|J_X\|_{P_X}^4}{2 \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2} = \frac{\chi^2 (R_X, P_X)^2}{2 \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2}$$

as claimed. □

We now consider upper bounding KL divergence with χ^2-divergence. To do this, we use a well-known bound provided in [9] which is actually quite easy to prove using Jensen's inequality. The bound is presented in Lemma 3.3.6.

Lemma 3.3.6 (KL Divergence Upper Bound)**.** *Given pmfs P_X and R_X on the discrete and finite alphabet $\mathcal{X}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $R_X = P_X + J_X$, where J_X is a valid additive perturbation, we have:*

$$D(R_X||P_X) \leq \|J_X\|_{P_X}^2 = \chi^2 (R_X, P_X) .$$

Proof.
Noting that the natural logarithm is a concave function, using Jensen's inequality, we have:

$$D(R_X||P_X) = \mathbb{E}_{R_X} \left[\log \left(\frac{R_X(X)}{P_X(X)} \right) \right] \leq \log \left(\mathbb{E}_{R_X} \left[\frac{R_X(X)}{P_X(X)} \right] \right) .$$

Observe that:

$$\mathbb{E}_{R_X} \left[\frac{R_X(X)}{P_X(X)} \right] = \sum_{x \in \mathcal{X}} \frac{R_X^2(x)}{P_X(x)} = \sum_{x \in \mathcal{X}} \frac{J_X^2(x)}{P_X(x)} + P_X(x) + 2J_X(x) = 1 + \|J_X\|_{P_X}^2 .$$

Hence, we have:

$$D(R_X||P_X) \leq \log \left(1 + \|J_X\|_{P_X}^2 \right) \leq \|J_X\|_{P_X}^2 = \chi^2 (R_X, P_X) \tag{3.33}$$

where the second inequality follows from: $\forall x > -1,\ \log(1 + x) \leq x$, which is fondly known as Gallager's favorite inequality in MIT. □

We remark that the first bound in equation 3.33 is clearly tighter than the second bound. However, we will not use this tighter bound, because we will require the χ^2-divergence term to be isolated in the proofs that follow.

3.3.3 Performance Bound

Using the lemmata from the previous section, we can upper bound the hypercontractive constant in terms of the squared Rényi correlation. Combining Lemmata 3.3.4 and 3.3.6 gives Theorem 3.3.7, and combining Lemmata 3.3.5 and 3.3.6 gives Theorem 3.3.8. These are presented next.

Theorem 3.3.7 (Upper Bound on Hypercontractive Constant)**.** *For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$s^*(X;Y) \leq 2 \left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right) \rho^2(X;Y).$$

Proof.
Let W be the column stochastic conditional probability matrix with conditional pmfs $P_{Y|X}$ along its columns, as shown in equation 1.23. For any pmf $R_X = P_X + J_X$ on $\mathcal{X}$ such that $J_X \neq 0$ is a valid additive perturbation, we have $R_Y = WR_X$. Using Lemmata 3.3.4 and 3.3.6, we get:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq 2 \left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right) \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2}.$$

Taking the supremum over R_X on the left hand side first, and then the supremum over J_X on the right hand side, we have:

$$\sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq 2 \left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right) \sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2}$$

Using Definition 3.1.5 and Corollary 3.2.5, we get:

$$s^*(X;Y) \leq 2 \left(\sum_{x \in \mathcal{X}} \frac{1}{P_X(x)} \right) \rho^2(X;Y).$$

This completes the proof. □

Theorem 3.3.8 (Tighter Upper Bound on Hypercontractive Constant)**.** *For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$s^*(X;Y) \leq \left(\frac{2}{\min\limits_{x \in \mathcal{X}} P_X(x)} \right) \rho^2(X;Y).$$

Proof.
Once again, let W be the column stochastic conditional probability matrix with conditional pmfs $P_{Y|X}$ along its columns, as shown in equation 1.23. For any pmf $R_X = P_X + J_X$ on $\mathcal{X}$ such that $J_X \neq 0$ is a valid additive perturbation, we have $R_Y = WR_X$. Using Lemmata 3.3.5 and 3.3.6, we get:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{2}{\|J_X\|_{P_X}^2} \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2 \frac{\|WJ_X\|_{P_Y}^2}{\|J_X\|_{P_X}^2}.$$

As in the proof of Theorem 3.3.7, we take the supremum over R_X on the left hand side first, and then the supremum over J_X on the right hand side. Moreover, since the supremum of a non-negative product is less than or equal to the product of the suprema, we have:

$$s^*(X;Y) \leq 2\rho^2(X;Y) \sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{1}{\|J_X\|_{P_X}^2} \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2 \tag{3.34}$$

using Definition 3.1.5 and Corollary 3.2.5. Furthermore, note that:

$$\frac{1}{\|J_X\|_{P_X}^2} \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2 \leq \frac{1}{\min\limits_{x \in \mathcal{X}} P_X(x)} \frac{\max\limits_{x \in \mathcal{X}} \frac{J_X^2(x)}{P_X(x)}}{\|J_X\|_{P_X}^2} \leq \frac{1}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

which gives us:

$$\sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{1}{\|J_X\|_{P_X}^2} \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2 \leq \frac{1}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Hence, we get:

$$s^*(X;Y) \leq \left(\frac{2}{\min\limits_{x \in \mathcal{X}} P_X(x)} \right) \rho^2(X;Y).$$

This completes the proof. □

Theorem 3.3.8 is a tighter upper bound on the hypercontractive constant than Theorem 3.3.7, because we use the tighter Lemma 3.3.5 to prove it. We note that inequality 3.34 is a tighter bound than Theorem 3.3.8, but we loosen it to provide a more agreeable form. This loosening can be understood through the equivalence of norms. For any vector $x \in \mathbb{R}^n$, the 2-norm is equivalent to the ∞-norm, because:

$$\|x\|_\infty^2 \leq \|x\|_2^2 \leq n \|x\|_\infty^2 \tag{3.35}$$

where n, which is the dimension of the vector space, is the tightest factor that does not invalidate the upper bound. In inequality 3.34, we have the term:

$$\sup_{\substack{J_X : J_X \neq 0 \\ 1^T J_X = 0}} \frac{1}{\|J_X\|_{P_X}^2} \left(\max_{x \in \mathcal{X}} \left| \frac{J_X(x)}{P_X(x)} \right| \right)^2$$

which can be intuitively perceived as the tightest factor that can be inserted when a squared weighted 2-norm is upper bounded by a squared weighted ∞-norm. From inequality 3.35, we know that this tightest factor must correspond to the dimension of the vector space. Indeed, our factor satisfies:

$$\frac{1}{\min\limits_{x \in \mathcal{X}} P_X(x)} \geq |\mathcal{X}|$$

and it models the dimension of the vector space. We now present the main result of this section. The tighter bound in Theorem 3.3.8 is used to conclude this result (rather than Theorem 3.3.7).

Theorem 3.3.9 (Performance Bound)**.** *For random variables X and Y with joint pmf $P_{X,Y}$ defined over $\mathcal{X} \times \mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$\rho^2(X;Y) \leq s^*(X;Y) \leq \left(\frac{2}{\min\limits_{x \in \mathcal{X}} P_X(x)} \right) \rho^2(X;Y).$$

This theorem does not require a proof; it is a direct consequence of writing Theorems 3.3.1 and 3.3.8 together. The lower bound of Theorem 3.3.9 asserts the intuitive fact that local algorithms perform poorer than global ones in a data processing sense. The upper bound limits how much worse local optimal algorithms can perform with respect to the global optimal performance. For this reason, we designate Theorem 3.3.9 as the "performance bound."

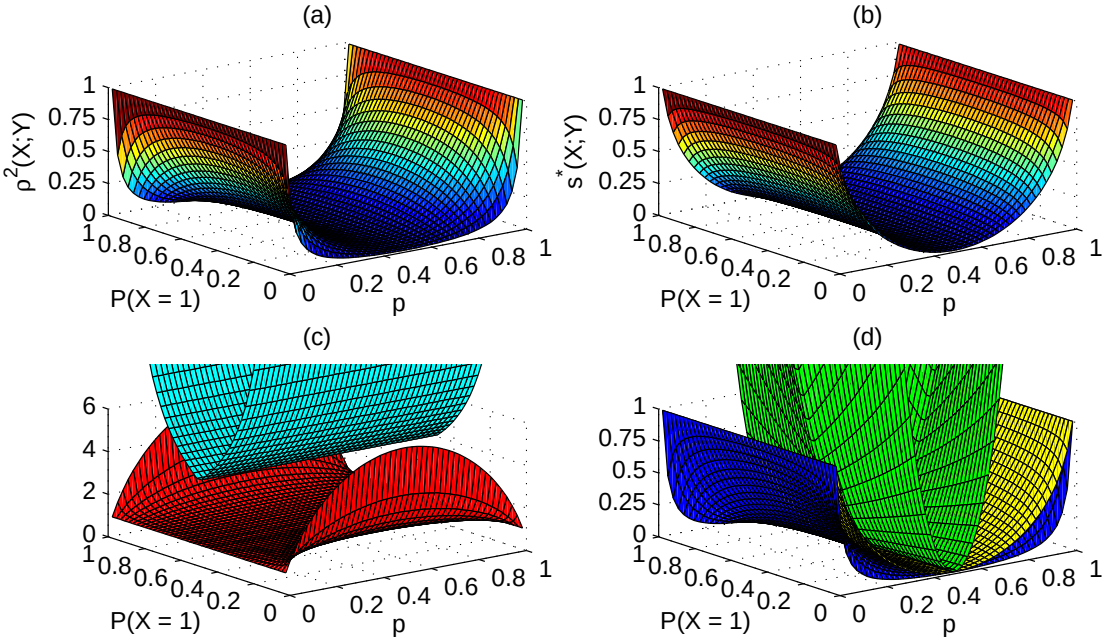

Figure 3.2: Plots of quantities corresponding to (X, Y), where $X \sim \text{Bernoulli}\,(\mathbb{P}(X = 1))$ is passed through a BSC with flip-over probability p to produce Y. (a) Squared Rényi correlation of (X, Y), $\rho^2(X;Y)$. (b) Hypercontractive constant of (X, Y), $s^*(X;Y)$. (c) Comparison of $\frac{s^*(X;Y)}{\rho^2(X;Y)}$ (red plot) and its upper bound $\frac{2}{\min(\mathbb{P}(X=0),P(X=1))}$ (cyan plot). (d) Performance bound illustrating $\rho^2(X;Y)$ (blue plot), $s^*(X;Y)$ (yellow plot), and $\left(\frac{2}{\min(\mathbb{P}(X=0),\mathbb{P}(X=1))}\right)\rho^2(X;Y)$ (green plot).

Figure 3.2 illustrates many facets of the performance bound in the special case where X is an input Bernoulli random variable to a BSC and Y is the corresponding output Bernoulli random variable. In particular, Figures 3.2(a) and 3.2(b) depict how the squared Rényi correlation and hypercontrative constant values change with the parameters of the input Bernoulli random variable and the BSC. Figure 3.2(d) depicts the performance bound itself, and verifies that the upper bound (green plot) is non-trivial since it remains below 1 for a large subset of the parameter space. In

fact, if we let $(X,Y) \sim \text{DSBS}(p)$ for $0 \leq p \leq 1$ (which is a slice along $\mathbb{P}(X=1)=0.5$ in Figure 3.2(d)), we know from equation 3.28 that $\rho^2(X;Y) = s^*(X;Y) = (1-2p)^2$. The upper bound in Theorem 3.3.9 is:

$$\frac{2}{\min\left(\mathbb{P}(X=0), \mathbb{P}(X=1)\right)}\rho^2(X;Y) = 4\left(1-2p\right)^2$$

and it satisfies:

$$4\left(1-2p\right)^2 < 1 \Leftrightarrow \frac{1}{4} < p < \frac{3}{4}.$$

So, although the upper bound is loose in this scenario, it is tighter than the trivial bound of 1 for $\frac{1}{4} < p < \frac{3}{4}$.

Simulations displayed in Figure 3.2(c) illustrate that the ratio between $s^*(X;Y)$ and $\rho^2(X;Y)$ increases significantly near the edges of the input probability simplex when one or more of the probability masses of the input pmf are close to 0. This effect is unsurprising given the skewed nature of stochastic manifolds at the edges of the probability simplex. It is captured in the upper bound of Theorem 3.3.9 because the constant $\frac{2}{\min_{x\in\mathcal{X}} P_X(x)}$ increases when one of the probability masses tends to 0. However, the constant $\frac{2}{\min_{x\in\mathcal{X}} P_X(x)}$ does not tensorize, while both $s^*(X;Y)$ and $\rho^2(X;Y)$ tensorize (Lemmata 3.1.3 and 3.2.1). This can make the upper bound quite loose. For example, if $X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$, then $\frac{2}{\min_{x\in\mathcal{X}} P_X(x)} = 4$. If we now consider $X_1^n = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are i.i.d. $\text{Bernoulli}\left(\frac{1}{2}\right)$, then X_1^n has a 2^n-point uniform pmf. Here, the constant of the upper bound is: $\frac{2}{\min_{x\in\{0,1\}^n} P_{X_1^n}(x_1^n)} = 2^{n+1}$. However, $s^*(X_1^n;Y_1^n) = s^*(X_1;Y_1)$ and $\rho^2(X_1^n;Y_1^n) = \rho^2(X_1;Y_1)$, due to their tensorization properties when the channel is memoryless. There is a quick fix for this i.i.d. loosening attack, which is presented in the next corollary.

Corollary 3.3.10 (I.I.D. Performance Bound)**.** *For jointly distributed random vectors X_1^n and Y_1^n, where $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$ defined over $\mathcal{X}\times\mathcal{Y}$, such that $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$, we have:*

$$\rho^2(X_1^n;Y_1^n) \leq s^*(X_1^n;Y_1^n) \leq \left(\frac{2}{\min\limits_{x\in\mathcal{X}} P_X(x)}\right)\rho^2(X_1^n;Y_1^n)$$

where $\min\limits_{x\in\mathcal{X}} P_X(x) = \left(\min\limits_{x_1^n\in\mathcal{X}^n} P_{X_1^n}\left(x_1^n\right)\right)^{\frac{1}{n}}$.

The corollary is trivially seen to be true since Lemmata 3.1.3 and 3.2.1 ensure that $s^*(X_1^n;Y_1^n) = s^*(X_1;Y_1)$ and $\rho^2(X_1^n;Y_1^n) = \rho^2(X_1;Y_1)$. This means we may use the factor:

$$\frac{2}{\min\limits_{x\in\mathcal{X}} P_X(x)}$$

instead of the looser factor:

$$\frac{2}{\min\limits_{x_1^n\in\mathcal{X}^n} P_{X_1^n}\left(x_1^n\right)}$$

in our upper bound. Therefore, Theorem 3.3.9 and Corollary 3.3.10 provide appropriate performance bounds for our local approximations in the discrete and finite setting. While Corollary

3.3.10 partially remedies the tensorization issue that ails Theorem 3.3.9, we ideally seek a constant in the upper bound which naturally tensorizes and does not change blindly with the dimension of the problem. This is a possible direction of future research.

3.4 Gaussian Case

The local approximation technique of [4] has also been applied to additive white Gaussian noise (AWGN) channels in [23]. So, we now consider the relationship between the local and global optimal values of problem 3.7 in the Gaussian case. As stated however, this is an ill-defined pursuit. Indeed, problem 3.7 is only defined for the discrete and finite setting. Moreover, the local approximation set-up in chapter 1 is also for the discrete and finite case. Recall that in section 2.1, we defined local perturbations and derived the local f-divergence (and hence, local KL divergence) in the continuous case. Using these definitions, we may define an analogous problem to problem 3.7 for the AWGN channel. For the sake of brevity, we do not explicitly define this problem, but the ensuing discourse reveals why the hypercontractive constant (to be defined in this scenario) and Rényi correlation represent the global and local optimal performance values, respectively. Furthermore, for simplicity in this section, we will abstain from unnecessary rigor like appropriately defining the underlying probability space every time.

We first introduce the AWGN channel. The AWGN channel is a well-known classical channel in information theory. [2] contains a very accessible introduction to it. The channel is discrete-time, but at each time instance we assume that the input and output are continuous random variables, X and Y, respectively. The additive Gaussian noise at each time instance, $W \sim \mathcal{N}(0, \sigma_W^2)$, is independent of other time instances, and also independent of the input X. The channel model is given in the next definition.

Definition 3.4.1 (Single Letter AWGN Channel)**.** The single letter AWGN channel has jointly distributed input random variable X and output random variable Y, where X and Y are related by the equation:

$$Y = X + W, \;\; X \perp\!\!\!\perp W \sim \mathcal{N}\left(0, \sigma_W^2\right)$$

where X is independent of the Gaussian noise $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, $\sigma_W^2 > 0$, and X must satisfy the average power constraint:

$$\mathbb{E}\left[X^2\right] \leq \sigma_X^2$$

for some given power $\sigma_X^2 > 0$.

Traditionally, the average power constraint in Definition 3.4.1 is defined over codewords [2], but we provide an alternative definition to avoid considering channel coding in our arguments. Given the model in Definition 3.4.1, it is well-known that the information capacity of the AWGN channel is:

$$C \triangleq \frac{1}{2} \log\left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right) \tag{3.36}$$

where the capacity achieving input distribution is $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$.

We next observe that Definition 3.2.1 of Rényi correlation is valid for continuous random variables. So, $\rho(X;Y)$ is a well-defined quantity for the AWGN channel. The DPI for KL divergence in Theorem 3.3.2 also holds for continuous random variables with pdfs. However, the hypercontractive constant is defined only for discrete and finite random variables in Definition 3.1.5. Hence, to study the DPI for AWGN channels, we define a new and analogous notion of $s^*(X;Y)$ for continuous random variables with a power constraint.

Definition 3.4.2 (Hypercontractive Constant with Power Constraint)**.** For a pair of jointly continuous random variables X and Y with joint pdf $P_{X,Y}$ and average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, the hypercontractive constant is defined as:

$$s^*_{\sigma_X^2}(X;Y) \triangleq \sup_{\substack{R_X : R_X \neq P_X \\ \mathbb{E}_{R_X}\left[X^2\right] \leq \sigma_X^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where we take the supremum over all pdfs $R_X \neq P_X$, which means that the pdfs R_X and P_X must differ on a set with non-zero Lebesgue measure. Moreover, R_Y denotes the marginal pdf of $R_{X,Y} = P_{Y|X} R_X$, and we assume that $P_{X,Y}$ is fixed.

Although we call $s^*_{\sigma_X^2}(X;Y)$ the hypercontractive constant with a power constraint and use similar notation to describe it, the nomenclature and notation do not reflect any deep relation to hypercontractivity. We simply wish to conform to the naming convention in the discrete and finite case. For the ensuing discussion, we fix the capacity achieving distribution, $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, $\sigma_X^2 > 0$, as the input distribution to the AWGN channel with average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$. This defines a joint distribution $P_{X,Y}$ for the AWGN channel from which we can compute the hypercontractive constant with power constraint and the Rényi correlation. Is is easy to observe from Definition 3.4.2 that the hypercontractive constant with power constraint represents the tightest factor that can be inserted into the DPI for KL divergence (as in equation 3.29) if the power constraint is imposed on all input distributions R_X. Moreover, it turns out that the squared Rényi correlation represents the tightest DPI in the local case [23]. Therefore, we explicitly compute the Rényi correlation, $\rho(X;Y)$, and the hypercontractive constant with power constraint, $s^*_{\sigma_X^2}(X;Y)$, of the AWGN channel for comparison.

3.4.1 Rényi Correlation of AWGN Channel

In order to find the Rényi correlation of the AWGN channel with capacity achieving input distribution, we first compute the Rényi correlation of two jointly Gaussian random variables. The strategy for this is to use Theorem 3.2.6, the MMSE characterization of Rényi correlation, to infer the optimizing functions for the Rényi correlation. It is well-known that if (X,Y) are jointly Gaussian, then the MMSE estimator of X given Y is also the linear least squares error (LLSE) estimator. Since Theorem 3.2.6 states that the optimizing functions of Definition 3.2.1 of Rényi correlation satisfy $f^*(X) \propto \mathbb{E}\left[g^*(Y)|X\right]$ and $g^*(Y) \propto \mathbb{E}\left[f^*(X)|Y\right]$, we can conjecture that f^* and g^* are linear functions of X and Y, respectively, normalized to have zero mean and unit variance. This is proven in the next theorem.

Theorem 3.4.1 (Gaussian Rényi Correlation)**.** *For jointly Gaussian random variables (X,Y) with distribution $(X,Y) \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$, the following are true:*

1. *The Rényi correlation is the absolute value of the Pearson correlation coefficient:*

$$\rho(X;Y) = \left| \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \right|.$$

2. *A pair of optimizing functions of the Rényi correlation, $\rho(X;Y)$, are:*

$$f^*(x) = \pm \frac{x - \mu_X}{\sigma_X} \text{ and } g^*(y) = \pm \frac{y - \mu_Y}{\sigma_Y},$$

where the signs of f^* and g^* are chosen so that their covariance is non-negative. This means that $(f^*(X), g^*(Y)) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho(X;Y) \\ \rho(X;Y) & 1 \end{bmatrix} \right)$.

Proof.
Part 1 of the theorem is Rényi's seventh axiom [19]. So, the Rényi correlation must satisfy it.

We now prove part 2. Consider a guess of the optimizing function, $f(X) = \frac{X - \mu_X}{\sigma_X}$. Clearly, $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}\left[f^2(X)\right] = 1$. We will derive $g(Y)$ from this guess, and show that the resulting $f(X)$ and $g(Y)$ pair satisfy part 1. (The pair will also satisfy the conditions in Theorem 3.2.6.) To this end, note that:

$$\mathbb{COV}\left(f(X), Y\right) = \mathbb{E}\left[f(X)Y\right] = \mathbb{E}\left[\frac{XY - \mu_X Y}{\sigma_X}\right] = \frac{\sigma_{XY}}{\sigma_X}$$

and hence, $(f(X), Y) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu_Y \end{bmatrix}, \begin{bmatrix} 1 & \frac{\sigma_{XY}}{\sigma_X} \\ \frac{\sigma_{XY}}{\sigma_X} & \sigma_Y^2 \end{bmatrix} \right)$. We now find the MMSE estimator of $f(X)$ given $Y = y$:

$$\mathbb{E}\left[f(X)|Y = y\right] = \frac{\mathbb{E}\left[X|Y = y\right] - \mu_X}{\sigma_X}.$$

For jointly Gaussian (X, Y), the MMSE estimator is the LLSE estimator, given by $\mathbb{E}\left[X|Y = y\right] = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \mu_Y)$. Using this, we get:

$$\mathbb{E}\left[f(X)|Y = y\right] = \frac{\sigma_{XY}}{\sigma_X \sigma_Y^2}\left(y - \mu_Y\right)$$

from which, we have:

$$\mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]\right] = \mathbb{E}\left[f(X)\right] = 0,$$

$$\mathbb{VAR}\left(\mathbb{E}\left[f(X)|Y\right]\right) = \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]^2\right] = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^4}\mathbb{E}\left[\left(Y - \mu_Y\right)^2\right] = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}.$$

Using intuition from Theorem 3.2.6, we normalize $\mathbb{E}\left[f(X)|Y = y\right]$ to have unit variance and let this be $g(y)$:

$$g(y) = \pm \frac{\sigma_X \sigma_Y}{\sigma_{XY}} \frac{\sigma_{XY}}{\sigma_X \sigma_Y^2}\left(y - \mu_Y\right) = \pm \frac{y - \mu_Y}{\sigma_Y}$$

where the sign of g is chosen such that $\mathbb{E}\left[f(X)g(Y)\right] \geq 0$. Observe that g has the same form as f. So, starting with g, we may use a similar argument to get f. (This shows that f and g satisfy the conditions of Theorem 3.2.6, which optimizing functions of Rényi correlation must satisfy.)

Hence, f and g are valid candidates for the optimizing functions. In fact, the Pearson correlation coefficient between them is the Rényi correlation given in part 1 of this theorem as shown below:

$$\mathbb{E}\left[f(X)g(Y)\right] = \left|\mathbb{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]\right| = \left|\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right| = \rho(X;Y)$$

where the absolute values ensure the covariance is non-negative. So, f and g are indeed the optimizing functions of Rényi correlation: $f^* = f$ and $g^* = g$. This proves part 2. □

Using Theorem 3.4.1, we may compute the Rényi correlation of the AWGN channel with capacity achieving input distribution. This is shown in the next corollary.

Corollary 3.4.2 (AWGN Rényi Correlation)**.** *Given an AWGN channel, $Y = X+W$, with average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, $X \perp\!\!\!\perp W$, and $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, if X takes on the capacity achieving distribution $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, then the Rényi correlation between X and Y is:*

$$\rho(X;Y) = \frac{\sigma_X}{\sqrt{\sigma_X^2 + \sigma_W^2}}.$$

Proof.
Using Theorem 3.4.1, we have:

$$\rho(X;Y) = \frac{|\mathbb{COV}(X,Y)|}{\sqrt{\mathbb{VAR}(X)\mathbb{VAR}(Y)}} = \frac{|\mathbb{E}[XY]|}{\sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}}$$

where the last equality follows from $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Moreover, $\mathbb{E}[XY] = \mathbb{E}[X(X+W)] = \mathbb{E}\left[X^2\right] + \mathbb{E}[X]\mathbb{E}[W] = \sigma_X^2$ and $\mathbb{E}\left[Y^2\right] = \sigma_X^2 + \sigma_W^2$. Hence, we get:

$$\rho(X;Y) = \frac{\sigma_X^2}{\sqrt{\sigma_X^2\left(\sigma_X^2 + \sigma_W^2\right)}} = \frac{\sigma_X}{\sqrt{\sigma_X^2 + \sigma_W^2}}.$$

□

We remark that just as in the discrete and finite setting, for the AWGN channel, $\rho(X;Y)$ is the second largest singular value of the divergence transition map which takes perturbations of the Gaussian input along right singular vector directions of Hermite polynomials to perturbations of the Gaussian output [23]. Hence, it represents the local optimal performance in a linear information coupling sense. The spectral decomposition of continuous channels will be the agenda of chapter 5, so we do not discuss the Gaussian case in detail here.

3.4.2 Hypercontractive Constant of AWGN Channel

To compute the hypercontractive constant with power constraint for an AWGN channel with capacity achieving input distribution, we first use the notion of exponential families to estimate it. The exponential family is a framework for studying large classes of distributions. It unifies many areas of probability and statistics including efficient estimation and large deviation bounds. We

find an explicit form of the hypercontractive constant with the additional constraint that all distributions lie along an exponential family. Then, we prove that this constrained hypercontractive constant is actually the correct hypercontractive constant using the entropy power inequality [6] and a variant of Bernoulli's inequality [24]. We begin by recalling the definition and pertinent properties of canonical exponential families [25].

Definition 3.4.3 (Canonical Exponential Family)**.** The parametrized family of pdfs with parameter x, $\{P_Y(\cdot;x)\}$, is called a regular canonical exponential family when the support of the pdfs do not depend on x, and each pdf in the family has the form:

$$P_Y(y;x) = \exp\left[xt(y) - \alpha(x) + \beta(y)\right], \;\; y \in \mathbb{R}$$

where $t(y)$ is the sufficient statistic of the distribution, $P_Y(y;0) = \exp\left[\beta(y)\right]$ is a valid pdf known as the base distribution, and:

$$\exp\left[\alpha(x)\right] = \int_{-\infty}^{+\infty} \exp\left[xt(y) + \beta(y)\right] d\lambda(y)$$

is the partition function with $\alpha(0) = 0$ without loss of generality, where λ denotes the Lebesgue measure and the integral is the Lebesgue integral. The parameter x is called the natural parameter, and it takes values from the natural parameter space $\mathcal{X} \subseteq \mathbb{R}$, defined as:

$$\mathcal{X} \triangleq \{x \in \mathbb{R} : \alpha(x) < \infty\}$$

which ensures that $P_Y(\cdot;x)$ is a valid pdf when $x \in \mathcal{X}$.

While Definition 3.4.3 defines exponential families as pdfs (because this is all we require in this chapter), the definition can be generalized to pmfs and other distributions [25]. Moreover, Definition 3.4.3 only introduces the one parameter exponential family. In general, exponential families are defined with any finite number of parameters. We now list some properties of canonical exponential families in the next lemma. The proofs of these properties are not provided as they can be easily derived or found in reference texts [25].

Lemma 3.4.3 (Properties of Canonical Exponential Family)**.** *For a canonical exponential family $\{P_Y(\cdot;x)\}$ with natural parameter $x \in \mathbb{R}$, such that $\mathbb{R}$ is the natural parameter space and $P_Y(y;x) = \exp\left[xt(y) - \alpha(x) + \beta(y)\right], \;\; y \in \mathbb{R}$, the following statements are true:*

1. *Complementarity:*

$$\forall x \in \mathbb{R}, \;\; D\left(P_Y(\cdot;x)||P_Y(\cdot;0)\right) + \alpha(x) = x\mathbb{E}_{P_Y(\cdot;x)}[t(Y)].$$

2. *Properties of log-partition function:*

$$\forall x \in \mathbb{R}, \;\; \alpha(x) = \log\left(\mathbb{E}_{P_Y(\cdot;0)}\left[e^{xt(Y)}\right]\right)$$

$$\forall x \in \mathbb{R}, \;\; \alpha'(x) = \mathbb{E}_{P_Y(\cdot;x)}[t(Y)]$$

$$\forall x \in \mathbb{R}, \;\; \alpha''(x) = \mathbb{VAR}_{P_Y(\cdot;x)}\left(t(Y)\right) = \mathbb{VAR}_{P_Y(\cdot;x)}\left(S(Y;x)\right) = -\frac{\partial}{\partial x}S(Y;x) = J_Y(x)$$

where under regularity conditions such that the order of differentiation and integration can be changed:

$$\forall x \in \mathbb{R}, \;\; J_Y(x) \triangleq -\mathbb{E}_{P_Y(\cdot;x)}\left[\frac{\partial^2}{\partial x^2}\log\left(P_Y(Y;x)\right)\right]$$

is the Fisher information, and:

$$\forall x \in \mathbb{R}, \;\; S(Y;x) \triangleq \frac{\partial}{\partial x}\log\left(P_Y(Y;x)\right)$$

is the score function, which can be written as $S(Y;x) = t(Y) - \mathbb{E}_{P_Y(\cdot;x)}[t(Y)]$ *for a canonical exponential family.*

3. *Derivative of KL divergence:*

$$\forall x \in \mathbb{R}, \;\; \frac{\partial}{\partial x} D\left(P_Y(\cdot;x)||P_Y(\cdot;0)\right) = x\alpha''(x) = xJ_Y(x).$$

Given these properties of canonical exponential families, we compute the hypercontractive constant with power constraint with an additional constraint on the marginal distributions being canonical exponential families. Theorem 3.4.4 illustrates that the constant becomes dependent on the Fisher information of the input and output marginal distributions. Note that once again, we describe the theorem as "hypercontractive constant" for consistency in nomenclature rather than any deep ties with hypercontractivity itself.

Theorem 3.4.4 (Hypercontractive Constant with Exponential Family Constraint)**.** *For a pair of jointly continuous random variables* (X, Y) *with joint pdf* $P_{X,Y}$*, such that the marginal pdfs are canonical exponential families:*

$$P_X(x;\mu) = \exp\left[\mu t(x) - \alpha(\mu) + \beta(x)\right], \;\; x \in \mathbb{R}$$

$$P_Y(y;\mu) = \exp\left[\mu\tau(y) - A(\mu) + B(y)\right], \;\; y \in \mathbb{R}$$

with common natural parameter $\mu \in \mathbb{R}$*, we have:*

$$\sup_{\mu:\mu\neq 0} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{J_Y(\mu^*)}{J_X(\mu^*)}$$

where $\mu^* = \arg\sup_{\mu:\mu\neq 0} \dfrac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))}$.

Proof.
We prove this by recognizing that μ^* is also the arg sup of:

$$\log\left(D(P_Y(\cdot;\mu)||P_Y(\cdot;0))\right) - \log\left(D(P_X(\cdot;\mu)||P_X(\cdot;0))\right).$$

Using Lemma 3.4.3, we have:

$$\mu^* = \arg\sup_{\mu:\mu\neq 0} \log\left(\mu A'(\mu) - A(\mu)\right) - \log\left(\mu\alpha'(\mu) - \alpha(\mu)\right).$$

Differentiating the right hand side expression with respect to μ and setting it equal to 0, we have:

$$\frac{\mu A'(\mu) - A(\mu)}{\mu\alpha'(\mu) - \alpha(\mu)} = \frac{A''(\mu)}{\alpha''(\mu)}.$$

The above equation is satisfied by $\mu = \mu^*$. Hence, using Lemma 3.4.3, we get:

$$\frac{D(P_Y(\cdot;\mu^*)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu^*)||P_X(\cdot;0))} = \frac{J_Y(\mu^*)}{J_X(\mu^*)}.$$

This completes the proof. □

The elegance in the way Fisher information appears in Theorem 3.4.4 is primarily due to the canonical exponential family assumptions. Essentially, the canonical exponential family assumption can be thought of as locally approximating some arbitrary family of distributions. Under such a local view, KL divergences begin to look like Fisher information metrics, as mentioned in section 2.1 in chapter 2. This local assumption will in general not give us the global supremum $s^*_{\sigma_X^2}(X;Y)$. However, it turns out that in the jointly Gaussian case, the local supremum is also the global supremum. Intuitively, this is because Gaussian distributions have the fundamental property that they are completely characterized locally (by only first and second order moments).

Theorem 3.4.4 is used to derive the hypercontractive constant with power constraint for an AWGN channel with the additional constraint that the input and output marginal distributions are in canonical exponential families with common natural parameter. This derivation is presented in the next lemma. Observe that if $J_Y(\mu) = J_Y$ and $J_X(\mu) = J_X$ are constant, then Theorem 3.4.4 implies that:

$$\sup_{\mu:\mu\neq 0} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{J_Y}{J_X}.$$

Gaussian random variables with fixed variance and exponentially tilted means form a canonical exponential family with constant Fisher information. This allows a simple derivation of the additionally constrained hypercontractive constant for the AWGN channel, but we must take care to ensure that the usual average power constraint is met.

Lemma 3.4.5 (AWGN Hypercontractive Constant with Exponential Family Constraint)**.** *Given an AWGN channel, $Y = X + W$, with $X \perp\!\!\!\perp W$, $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, and average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2 + \epsilon$ for any $\epsilon > 0$, the hypercontractive constant with average power and marginal canonical exponential family constraints is:*

$$\sup_{\substack{\mu:\mu\neq 0 \\ P_X(\cdot;\mu)=\mathcal{N}\left(\mu,\sigma_X^2\right) \\ \mathbb{E}_{P_X(\cdot;\mu)}\left[X^2\right]\leq\sigma_X^2+\epsilon}} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

where we fix the marginal distribution of $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, and restrict the distributions we take the supremum over to be $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$.

Proof.
Let $P_X(x;\mu) = \exp\left[\mu t(x) - \alpha(\mu) + \beta(x)\right]$ with natural parameter space $\mathbb{R}$ (i.e. $\mu \in \mathbb{R}$) be a canonical exponential family with:

$$\exp\left[\beta(x)\right] = \mathcal{N}\left(0, \sigma_X^2\right), \;\; \alpha(\mu) = \frac{\mu^2}{2\sigma_X^2}, \;\text{ and }\; t(x) = \frac{x}{\sigma_X^2}.$$

It is easily verified that $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$. We now show that $P_Y(\cdot;\mu)$ is a canonical exponential family with natural parameter $\mu \in \mathbb{R}$ as well. For the AWGN channel, $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$ has corresponding output distribution $P_Y(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2 + \sigma_W^2\right)$. So, $P_Y(y;\mu) = \exp\left[\mu\tau(y) - A(\mu) + B(y)\right]$ is also a canonical exponential family with:

$$\exp\left[B(y)\right] = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right), \;\; A(\mu) = \frac{\mu^2}{2(\sigma_X^2 + \sigma_W^2)}, \;\text{ and }\; \tau(y) = \frac{y}{\sigma_X^2 + \sigma_W^2}.$$

Using Lemma 3.4.3, we compute the Fisher information of the input and the output models:

$$J_X(\mu) = \alpha''(\mu) = \frac{1}{\sigma_X^2},$$

$$J_Y(\mu) = A''(\mu) = \frac{1}{\sigma_X^2 + \sigma_W^2}.$$

Hence, using Theorem 3.4.4, we have:

$$\sup_{\substack{\mu:\mu\neq 0 \\ P_X(\cdot;\mu)=\mathcal{N}\left(\mu,\sigma_X^2\right)}} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}.$$

However, we also need to satisfy the average power constraint: $\mathbb{E}_{P_X(\cdot;\mu)}\left[X^2\right] \leq \sigma_X^2 + \epsilon$, where $\epsilon > 0$ is some small additional power. To this end, notice from Lemma 3.4.3 that for every $\mu \neq 0$:

$$\frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{\mu A'(\mu) - A(\mu)}{\mu\alpha'(\mu) - \alpha(\mu)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} \tag{3.37}$$

which does not depend on μ. Since $X \sim P_X(\cdot;\mu)$ has expectation $\mathbb{E}_{P_X(\cdot;\mu)}[X] = \mu$ and variance $\mathbb{VAR}_{P_X(\cdot;\mu)}(X) = \sigma_X^2$, the average power constraint corresponds to:

$$\mathbb{E}_{P_X(\cdot;\mu)}\left[X^2\right] = \mu^2 + \sigma_X^2 \leq \sigma_X^2 + \epsilon \;\; \Leftrightarrow \;\; |\mu| \leq \sqrt{\epsilon}.$$

As $\epsilon > 0$, $\exists \mu \neq 0$ such that $|\mu| \leq \sqrt{\epsilon}$. So there exists μ which satisfies the average power constraint. Such a μ also satisfies equation 3.37. Thus, we have:

$$\sup_{\substack{\mu:\mu\neq 0 \\ P_X(\cdot;\mu)=\mathcal{N}\left(\mu,\sigma_X^2\right) \\ \mathbb{E}_{P_X(\cdot;\mu)}\left[X^2\right]\leq\sigma_X^2+\epsilon}} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

as claimed. This completes the proof. □

We remark that Theorem 3.4.4 is not essential to the proof of Lemma 3.4.5. Indeed, the observation in equation 3.37 suffices in playing the same role in the proof. Moreover, the proof of Lemma 3.4.5 elucidates why we use the additional slack of $\epsilon > 0$ in the average power constraint: $\mathbb{E}\left[X^2\right] \leq \sigma_X^2 + \epsilon$. If $\epsilon = 0$ and we take the supremum over $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$ for $\mu \neq 0$, then the average power constraint is never satisfied, and the supremum over an empty set is $-\infty$. Hence, in order to make the supremum well-defined, we add a slack of ϵ to the average power constraint and use a nearly capacity achieving input distribution like $X \sim \mathcal{N}\left(\sqrt{\epsilon}, \sigma_X^2\right)$ as input to the AWGN channel. We next prove that the constrained hypercontractive constant in Lemma 3.4.5 equals $s^*_{\sigma_X^2}(X;Y)$ for the AWGN channel.

Theorem 3.4.6 (AWGN Hypercontractive Constant)**.** *Given an AWGN channel, $Y = X + W$, with $X \perp\!\!\!\perp W$, $X \sim P_X = \mathcal{N}\left(0, \sigma_X^2\right)$, $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, and average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, the hypercontractive constant with average power constraint is:*

$$s^*_{\sigma_X^2}(X;Y) = \sup_{\substack{R_X : R_X \neq P_X \\ \mathbb{E}_{R_X}\left[X^2\right] \leq \sigma_X^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

where $R_Y = P_{Y|X} R_X$ and $P_{Y|X}(\cdot|x) = \mathcal{N}\left(x, \sigma_W^2\right), \ \ x \in \mathbb{R}$.

Proof.
Lemma 3.4.5 shows that the ratio of KL divergences can achieve $\frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$ if some additional slack of $\epsilon > 0$ is allowed in the average power constraint. Since ϵ is arbitrary, this means that the supremum of the ratio of KL divergences should be able to achieve $\frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$ with effectively no additional slack. So, it is sufficient to prove that:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2} \tag{3.38}$$

for every pdf $R_X \neq P_X$ (which means that the pdfs R_X and P_X must differ on a set with non-zero Lebesgue measure) satisfying $\mathbb{E}_{R_X}\left[X^2\right] \leq \sigma_X^2$. Let R_X and R_Y have second moments $\mathbb{E}_{R_X}\left[X^2\right] = \sigma_X^2 + p$ and $\mathbb{E}_{R_Y}\left[Y^2\right] = \sigma_X^2 + \sigma_W^2 + p$, for some $-\sigma_X^2 < p \leq 0$. Since $P_X = \mathcal{N}\left(0, \sigma_X^2\right)$ and $P_Y = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right)$ are Gaussian, we have from Definition 1.0.1 that:

$$\begin{aligned}
D(R_X||P_X) &= \mathbb{E}_{R_X}\left[\log\left(R_X\right)\right] - \mathbb{E}_{R_X}\left[\log\left(P_X\right)\right] \\
&= \frac{1}{2}\log\left(2\pi\sigma_X^2\right) + \frac{1}{2\sigma_X^2}\mathbb{E}_{R_X}\left[X^2\right] - h\left(R_X\right) \\
&= \frac{1}{2}\log\left(2\pi e\sigma_X^2\right) + \frac{p}{2\sigma_X^2} - h\left(R_X\right) \\
&= h\left(P_X\right) - h\left(R_X\right) + \frac{p}{2\sigma_X^2}
\end{aligned}$$

and in a similar manner:

$$D(R_Y||P_Y) = h\left(P_Y\right) - h\left(R_Y\right) + \frac{p}{2\left(\sigma_X^2 + \sigma_W^2\right)}$$

where for any pdf $P : \mathbb{R} \to \mathbb{R}^+$, we define:

$$h\left(P\right) \triangleq -\mathbb{E}_P\left[\log\left(P\right)\right]$$

as the differential entropy of P. Hence, we know from inequality 3.38 that it is sufficient to prove:

$$\left(\sigma_X^2+\sigma_W^2\right)\left(h(P_Y)-h(R_Y)+\frac{p}{2\left(\sigma_X^2+\sigma_W^2\right)}\right)\leq\sigma_X^2\left(h(P_X)-h(R_X)+\frac{p}{2\sigma_X^2}\right)$$

which simplifies to:

$$\left(\sigma_X^2+\sigma_W^2\right)\left(h(P_Y)-h(R_Y)\right)\leq\sigma_X^2\left(h(P_X)-h(R_X)\right). \tag{3.39}$$

We transform this conjectured inequality into one with entropy power terms rather than differential entropy terms.

$$\begin{aligned}\left(e^{2h(P_Y)-2h(R_Y)}\right)^{\sigma_X^2+\sigma_W^2}&\leq\left(e^{2h(P_X)-2h(R_X)}\right)^{\sigma_X^2}\\ \left(\frac{\frac{1}{2\pi e}e^{2h(P_Y)}}{\frac{1}{2\pi e}e^{2h(R_Y)}}\right)^{\sigma_X^2+\sigma_W^2}&\leq\left(\frac{\frac{1}{2\pi e}e^{2h(P_X)}}{\frac{1}{2\pi e}e^{2h(R_X)}}\right)^{\sigma_X^2}\end{aligned}$$

Introducing entropy powers, we have:

$$\left(\frac{N(P_Y)}{N(R_Y)}\right)^{\sigma_X^2+\sigma_W^2}\leq\left(\frac{N(P_X)}{N(R_X)}\right)^{\sigma_X^2}$$

where for any pdf $P:\mathbb{R}\to\mathbb{R}^+$, the entropy power of P is defined as:

$$N(P)\triangleq\frac{1}{2\pi e}e^{2h(P)}.$$

For $P_X=\mathcal{N}\left(0,\sigma_X^2\right)$ and $P_Y=\mathcal{N}\left(0,\sigma_X^2+\sigma_W^2\right)$, the entropy powers are $N(P_X)=\sigma_X^2$ and $N(P_Y)=\sigma_X^2+\sigma_W^2$. Applying the entropy power inequality [6] to the AWGN channel, we have:

$$N(R_Y)\geq N(R_X)+N\left(\mathcal{N}(0,\sigma_W^2)\right)=N(R_X)+\sigma_W^2.$$

Hence, it is sufficient to prove that:

$$\left(\frac{\sigma_X^2+\sigma_W^2}{N(R_X)+\sigma_W^2}\right)^{\sigma_X^2+\sigma_W^2}\leq\left(\frac{\sigma_X^2}{N(R_X)}\right)^{\sigma_X^2}.$$

Since a Gaussian uniquely maximizes entropy under a second moment constraint and $R_X\neq P_X$, we know that $h(R_X)<h(P_X)\Rightarrow N(R_X)<N(P_X)=\sigma_X^2$. We also assume that $h(R_X)>-\infty$. Let $a=\sigma_X^2+\sigma_W^2$, $b=\sigma_X^2-N(R_X)$, and $c=\sigma_X^2$. Then, we have $a>c>b>0$, and we are required to prove:

$$\left(\frac{a}{a-b}\right)^a\leq\left(\frac{c}{c-b}\right)^c$$

which is equivalent to proving:

$$a>c>b>0\Rightarrow\left(1-\frac{b}{c}\right)^c\leq\left(1-\frac{b}{a}\right)^a.$$

This statement is a variant of Bernoulli's inequality proved in equation (r_7'') in [24]. Thus, we have shown:

$$\left(\frac{\sigma_X^2+\sigma_W^2}{N(R_X)+\sigma_W^2}\right)^{\sigma_X^2+\sigma_W^2}\leq\left(\frac{\sigma_X^2}{N(R_X)}\right)^{\sigma_X^2}$$

which completes the proof. □

We now derive an interesting corollary of Theorem 3.4.6. The corollary can be regarded as a fundamental bound on the deviation of the mutual information from the capacity of an AWGN channel in terms of the deviation of the differential entropy of the input distribution from the maximum differential entropy of the capacity achieving input distribution, where the input distributions satisfy the average power constraint. Note that mutual information for continuous random variables is defined in an analogous manner to Definition 1.0.3 by using integrals and pdfs instead of summations and pmfs [2]. In fact, the KL divergence characterization of mutual information in Definition 1.0.3 holds for both discrete and continuous random variables.

Corollary 3.4.7 (Mutual Information and Entropy Deviation Bound)**.** *Given an AWGN channel, $Y = X + W$, with $X \perp\!\!\!\perp W$, $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, and average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, for any input pdf R_X satisfying the average power constraint, we have:*

$$C - I\left(R_X; P_{Y|X}\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right)$$

where $C \triangleq \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right)$ is the AWGN channel capacity, and $P_{Y|X}(.|x) = \mathcal{N}\left(x, \sigma_W^2\right)$, $x \in \mathbb{R}$ is fixed.

Proof.
Let $P_X = \mathcal{N}\left(0, \sigma_X^2\right)$ be the capacity achieving input pdf; it is also the maximum differential entropy pdf given the second moment constraint. Let $P_Y = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right)$ be the output pdf corresponding to P_X. From Theorem 3.4.6, we have:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}$$

for any pdf $R_X \neq P_X$ (which means that the pdfs R_X and P_X must differ on a set with non-zero Lebesgue measure) satisfying $\mathbb{E}_{R_X}\left[X^2\right] \leq \sigma_X^2$. Note that if $R_X = P_X$ *a.e.*, then the inequality in the statement of Corollary 3.4.7 trivially holds with equality. So, we only prove the inequality for the case when $R_X \neq P_X$. Following the proof of Theorem 3.4.6, we again get inequality 3.39:

$$\left(h(P_Y) - h(R_Y)\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(h(P_X) - h(R_X)\right).$$

Now observe that $h(X + W, X) = h(X + W|X) + h(X) = h(W) + h(X)$ by the chain rule [2] and $X \perp\!\!\!\perp W$. Moreover, $h(X + W, X) = h(X|X + W) + h(X + W)$ by the chain rule. So, $h(X + W) = h(W) + h(X) - h(X|X + W) = h(W) + I(X + W; X)$. This means:

$$h(R_Y) = h\left(R_X \star \mathcal{N}\left(0, \sigma_W^2\right)\right) = h\left(\mathcal{N}\left(0, \sigma_W^2\right)\right) + I\left(R_X; P_{Y|X}\right)$$

where $\star$ denotes the convolution operator. This gives:

$$h(P_Y) - h\left(\mathcal{N}\left(0, \sigma_W^2\right)\right) - I\left(R_X; P_{Y|X}\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(h(P_X) - h(R_X)\right).$$

By definition of channel capacity, $C = h(P_Y) - h\left(\mathcal{N}\left(0, \sigma_W^2\right)\right)$. Hence, we have:

$$C - I\left(R_X; P_{Y|X}\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right)$$

as claimed. □

Corollary 3.4.7 has a cute analog in the discrete and finite case. Consider a discrete memoryless channel (Definition 1.3.1) with input random variable X which takes values on a finite alphabet $\mathcal{X}$, output random variable Y which takes values on a finite alphabet $\mathcal{Y}$, and conditional pmfs $P_{Y|X}$. Let P_X be the capacity achieving input pmf, and P_Y be the corresponding output pmf. If $\forall x \in \mathcal{X},\ P_X(x) > 0$, then for every pmf R_X on $\mathcal{X}$:

$$I\left(P_X; P_{Y|X}\right) - I\left(R_X; P_{Y|X}\right) = D\left(R_Y || P_Y\right) \tag{3.40}$$

where $R_Y = P_{Y|X} R_X$. This can be proved using the equidistance property of channel capacity [3] which states that P_X achieves capacity C if and only if $D\left(P_{Y|X=x} || P_Y\right) = C$ for every $x \in \mathcal{X}$ such that $P_X(x) > 0$ and $D\left(P_{Y|X=x} || P_Y\right) \leq C$ for every $x \in \mathcal{X}$ such that $P_X(x) = 0$. Equations 3.29 and 3.40 imply that:

$$I\left(P_X; P_{Y|X}\right) - I\left(R_X; P_{Y|X}\right) = D\left(R_Y || P_Y\right) \leq s^*(X;Y) D\left(R_X || P_X\right) \tag{3.41}$$

where $s^*(X;Y)$ is computed using P_X and $P_{Y|X}$. This parallels Corollary 3.4.7 in the discrete and finite setting.

We return to briefly discuss Corollary 3.4.7 itself. The inequality in Corollary 3.4.7 is tight and equality can be achieved. The inequality can also be written in terms of the signal-to-noise ratio, $SNR \triangleq \frac{\sigma_X^2}{\sigma_W^2}$, as follows:

$$C - I\left(R_X; P_{Y|X}\right) \leq \frac{SNR}{1 + SNR}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right). \tag{3.42}$$

Intuitively, this says that if $SNR \to 0$, then $I\left(R_X; P_{Y|X}\right) \to C$, which means that any input distribution satisfying the power constraint achieves capacity. This is because in the low SNR regime, capacity is also very small and it is easier to achieve it. On the other hand, under moderate or high SNR regimes, the capacity gap (between capacity and mutual information) is controlled solely by the input distribution. In particular, the capacity gap is sensitive to perturbations of the input distribution from the capacity achieving input distribution, and the input distribution achieves capacity if and only if it is Gaussian.

3.4.3 AWGN Channel Equivalence

In this final subsection, we state the main result we have been developing. The proof is trivial from Corollary 3.4.2 and Theorem 3.4.6, and is thus omitted.

Theorem 3.4.8 (AWGN Channel Equivalence)**.** *For an AWGN channel, $Y = X + W$, with $X \perp\!\!\!\perp W$, $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, $W \sim \mathcal{N}\left(0, \sigma_W^2\right)$, and average power constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, the hypercontractive constant with average power constraint equals the squared Rényi correlation:*

$$s^*_{\sigma_X^2}(X;Y) = \rho^2(X;Y) = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}.$$

The implications of this theorem are quite profound. Recall that $s^*_{\sigma_X^2}(X;Y)$ represents the tightest factor that can be inserted into the DPI for KL divergence with power constraint. Hence, it

represents the globally optimal performance in the sense of preserving information down an AWGN channel. On the other hand, $\rho^2(X;Y)$ represents the locally optimal performance in the same sense. Theorem 3.4.8 conveys that the local and global optima coincide in the Gaussian case. This conforms to our understanding of Gaussian distributions, where many local properties become global ones. We have essentially proved that for AWGN channels, the local approximation approach performs as well as any global approach in an information preservation sense.

Chapter 4

Large Deviations and Source-Channel Decomposition

We have heretofore established the local approximation of KL divergence by constructing perturbation spaces around reference distributions, and assessed the performance of algorithms which might employ such approximations by examining the tightness of the DPI. Moving forward, we develop an application of this approximation technique to better understand the large deviation behavior of sources and channels. Some of this work has been published in [8], and [26] provides a rigorous development of the pertinent large deviation theory for the interested reader.

In our discussion, we assume that all random variables are defined on a common underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will seldom refer to this probability space explicitly. Suppose we are given a pair of jointly distributed discrete random variables (X, Y) with joint pmf $P_{X,Y}$. Let the alphabet sets of X and Y be $\mathcal{X}$ and $\mathcal{Y}$ respectively, where $|\mathcal{X}| < \infty$ and $|\mathcal{Y}| < \infty$. Moreover, we assume that the joint pmf satisfies $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ P_{X,Y}(x, y) > 0$. This means the marginal pmfs satisfy $\forall x \in \mathcal{X},\ P_X(x) > 0$ and $\forall y \in \mathcal{Y},\ P_Y(y) > 0$. Then, we can think of X as the source or input random variable to a discrete memoryless channel (Definition 1.3.1) with transition probability matrix W. W is column stochastic as in equation 1.23; its columns are the conditional pmfs $P_{Y|X}$, and all its entries are strictly positive. The output to the channel is the random variable Y and its marginal pmf satisfies $P_Y = WP_X$. We will conform to the notation introduced in sections 1.1, 1.2, and 1.3 of chapter 1. For example, pmfs will be represented as column vectors. The channel view of (X, Y) is given in Figure 4.1.

Suppose further that we draw i.i.d. $(X_i, Y_i) \sim P_{X,Y}$ for $i = 1, \ldots, n$ to get (X_1^n, Y_1^n). Let us denote the empirical distribution of a sequence of i.i.d. random variables X_1^n by $\hat{P}_{X_1^n}$. We formally define this below.

Definition 4.0.4 (Empirical Distribution)**.** For a sequence of random variables X_1^n, for some $n \in \mathbb{Z}^+$, such that $\forall i \in \{1, \ldots, n\},\ X_i$ takes values on $\mathcal{X}$, we define the empirical distribution of the sequence as:

$$\forall x \in \mathcal{X},\ \hat{P}_{X_1^n}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i = x\right)$$

where $\mathbb{I}\left(X_i = x\right) = \mathbb{I}\left(\{\omega \in \Omega : X_i(\omega) = x\}\right)$, and the indicator function for any event $A \in \mathcal{F}$,

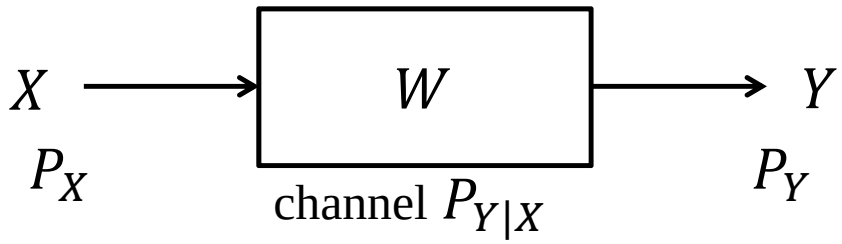

Figure 4.1: Channel view of (X, Y).

denoted by $\mathbb{I}(A)$, is defined as:

$$\mathbb{I}(A) = \begin{cases} 1 & , \text{ if } \omega \in A \\ 0 & , \text{ if } \omega \notin A \end{cases} .$$

If n is large, then $\hat{P}_{X_1^n}$ and $\hat{P}_{Y_1^n}$ are "close" to P_X and P_Y, respectively, with high probability by Sanov's Theorem [26]. In particular, the empirical distributions, $\hat{P}_{X_1^n}$ and $\hat{P}_{Y_1^n}$, lie on uniform KL divergence balls around the theoretical distributions, P_X and P_Y. This is illustrated in Figure 4.2.

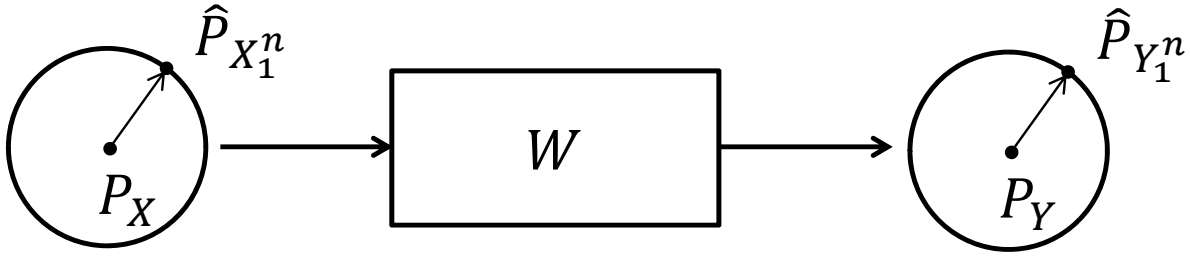

Figure 4.2: Uniform KL divergence balls.

Since we may think of i.i.d. $(X_i, Y_i) \sim P_{X,Y}$ for $i = 1, \ldots, n$ as inputting i.i.d. X_1^n into a discrete memoryless channel to get i.i.d. Y_1^n, the fact that the empirical output distribution resides uniformly on a divergence ball with high probability seemingly contradicts the elegant work in [4]. For large n, $\hat{P}_{X_1^n}$ can be perceived as a perturbation of P_X as it is close to P_X: $\hat{P}_{X_1^n} = P_X + \epsilon J_X$, where $\epsilon > 0$ is small and J_X is a valid additive perturbation. [4] portrays that although the input perturbation can be in any uniform direction around P_X, the output perturbation directions are not uniform because different directions are scaled by different singular values. This is depicted in Figure 4.3. In particular, the scenario in [4] assumes that $\hat{P}_{X_1^n}$ is another theoretical distribution perturbed from P_X, and then finds the theoretical perturbation of $\hat{P}_Y = W\hat{P}_{X_1^n}$ from $P_Y = WP_X$ using the SVD on the DTM. Since all the singular values of the DTM are less than or equal to 1 (from the proof of Theorem 3.2.4), the input sphere shrinks to the output ellipsoid. The ellipsoidal shape naturally occurs because certain singular vector directions have larger associated singular

values than others. The apparent paradox is that the output empirical perturbations should lie on a uniform sphere as Y_1^n are i.i.d. random variables. This is resolved by realizing that $\hat{P}_Y \neq \hat{P}_{Y_1^n}$. $\hat{P}_Y$ is the theoretical output when $\hat{P}_{X_1^n}$ is used as the theoretical input distribution, but $\hat{P}_{Y_1^n}$ is the actual empirical output distribution. It is fruitful to consider the relationship between $\hat{P}_Y$ and $\hat{P}_{Y_1^n}$ more closely. This is pursued in the next section.

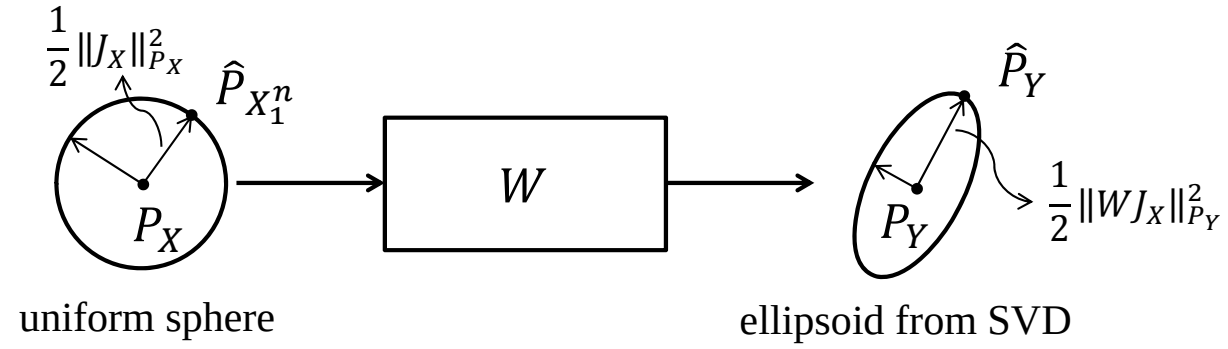

Figure 4.3: SVD characterization of output perturbations due to source.

For brevity and notational clarity, we will not write the ϵ factor when using perturbations in the remainder of this chapter. We will also neglect all $o\left(\epsilon^2\right)$ terms. For example, we will use $\hat{P}_{X_1^n} = P_X + J_X$ to denote the perturbed empirical distribution instead of the more rigorous $\hat{P}_{X_1^n} = P_X + \epsilon J_X$, where J_X is a valid additive perturbation. As another example, we will write:

$$D(\hat{P}_{X_1^n}||P_X) \overset{\text{local}}{=} \frac{1}{2}\|J_X\|^2_{P_X}$$

instead of the more rigorous statement from Definition 1.1.2:

$$D(\hat{P}_{X_1^n}||P_X) = \frac{1}{2}\epsilon^2\|J_X\|^2_{P_X} + o\left(\epsilon^2\right).$$

The notation $\overset{\text{local}}{=}$ will be used throughout this chapter to mean "equality under appropriate local approximations."

4.1 Source-Channel Perturbation Decomposition

From the previous discussion, we see that the perturbation between $\hat{P}_{Y_1^n}$ and P_Y is located on a uniform divergence ball. It is caused in part by the perturbation between $\hat{P}_{X_1^n}$ and P_X, which is also located on a uniform divergence ball. The effect of the source perturbation is presented in the SVD analysis of [4] and leads to the ellipsoidal perturbation $\hat{P}_Y - P_Y$. When we compute $\hat{P}_Y$, we take the perturbation of the source $\hat{P}_{X_1^n}$ into account, but assume that the channel is fixed at W. In reality, the conditional distributions of the channel will also perturb to give empirical conditional distributions. This channel perturbation also causes a perturbation of the output, and transforms the ellipsoid (due to the source perturbations) back into a sphere. Hence, the perturbation between

$\hat{P}_{Y_1^n}$ and P_Y is caused by a combination of source and channel perturbations.

Since empirical distributions are close to theoretical distributions when n is large, we let:

$$\hat{P}_{X_1^n} = P_X + J_X \tag{4.1}$$

and:

$$\hat{P}_{Y_1^n} = P_Y + J_Y \tag{4.2}$$

where J_X and J_Y are valid additive perturbations. We further let the column stochastic transition probability matrix be:

$$W = \left[W_1 \cdots W_{|\mathcal{X}|}\right] \tag{4.3}$$

where for each $x \in \mathcal{X}$, W_x is the conditional pmf of Y given $X = x$ written as a column vector. We define the column stochastic empirical transition probability matrix as:

$$V = \left[V_1 \cdots V_{|\mathcal{X}|}\right] = W + J_W \tag{4.4}$$

where for each $x \in \mathcal{X}$, $V_x = W_x + J_x$ is the empirical conditional pmf of Y given $X = x$, and $J_W = \left[J_1 \cdots J_{|\mathcal{X}|}\right]$ is a perturbation matrix whose columns are the valid additive perturbation vectors of the conditional pmfs $W_x,\ x \in \mathcal{X}$. Recall that a valid additive perturbation vector has the properties that the perturbed probability masses are between 0 and 1 (inclusive), and the sum of all entries in the vector is 0. The former constraint is not imposed in ensuing optimizations because it can be imposed by inserting the arbitrarily small $\epsilon > 0$ parameter in front of the perturbation vectors. The latter constraint is always imposed when optimizing.

Using these definitions, we derive the relationship between $\hat{P}_Y$ and $\hat{P}_{Y_1^n}$. First, we note that:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x(y) = \frac{\sum_{j=1}^{n} \mathbb{I}\left(X_j = x, Y_j = y\right)}{\sum_{i=1}^{n} \mathbb{I}\left(X_i = x\right)} \tag{4.5}$$

which correctly extends Definition 4.0.4 for empirical conditional distributions. We have to assume the denominator of equation 4.5 is non-zero for the ratio to be well-defined; this holds with probability 1 asymptotically as $\forall x \in \mathcal{X},\ \ P_X(x) > 0$. From equation 4.5, it is easily shown that:

$$\forall y \in \mathcal{Y},\ \ \hat{P}_{Y_1^n}(y) = \sum_{x \in \mathcal{X}} \hat{P}_{X_1^n}(x) V_x(y)$$

which simplifies to:

$$\hat{P}_{Y_1^n} = V \hat{P}_{X_1^n} \tag{4.6}$$

using matrix-vector notation. Since $V = W + J_W$, we have:

$$\hat{P}_{Y_1^n} = \hat{P}_Y + J_W \hat{P}_{X_1^n} \tag{4.7}$$

where $\hat{P}_Y = W \hat{P}_{X_1^n}$. This is the relationship between $\hat{P}_Y$ and $\hat{P}_{Y_1^n}$. To see this relationship in terms of perturbation vectors, we can subtract $P_Y = W P_X$ on both sides of equation 4.7 and use equations 4.1 and 4.2 to get:

$$J_Y = W J_X + J_W P_X + J_W J_X \tag{4.8}$$

where J_X and J_Y are uniform perturbations around P_X and P_Y respectively, and WJ_X is the ellipsoidal perturbation analyzed in [4]. If we neglect the second order term $J_W J_X$ (as this term would technically have an ϵ^2 factor in front of it), we have:

$$J_Y \overset{\text{local}}{=} WJ_X + J_W P_X \tag{4.9}$$

where equality holds under the obvious local approximations. This first order perturbation model is very elegant because it conveys that J_Y is the result of adding the source perturbation going through the unperturbed channel and the channel perturbation induced by the unperturbed source. J_Y is thus caused in part by the source perturbation, WJ_X, and in part by the channel perturbation, $J_W P_X$. This is a source-channel decomposition of the output perturbation J_Y. The decomposition setup engenders several natural questions:

1. If we observe J_Y, what are the most probable source and channel perturbations?
2. If we observe J_Y, what is the most probable source perturbation with fixed channel?
3. If we observe J_Y, what is the most probable channel perturbation with fixed source?

Analysis of the second question was the subject of [4]. The first and third questions will be explored in the sections 4.2 and 4.3, respectively. Section 4.4 will use the results of section 4.3 to model channel perturbations as additive Gaussian noise.

4.2 Most Probable Source and Channel Perturbations

Suppose we only observe the output Y_1^n of the discrete memoryless channel W as i.i.d. X_1^n are inputted into it. We assume we have full knowledge of the theoretical distribution $P_{X,Y}$. All probabilities in this section will be computed with respect to this distribution. In particular, we will use the notation $P_{X,Y}^n$ to denote the probability distribution of (X_1^n, Y_1^n). Observing Y_1^n tells us $\hat{P}_{Y_1^n}$, but there are many possible $\hat{P}_{X_1^n}$ and V which can cause $\hat{P}_{Y_1^n} = V\hat{P}_{X_1^n}$. A pertinent question is to try to find the most probable $\hat{P}_{X_1^n}$ and V which produce $\hat{P}_{Y_1^n}$. Our analysis to answer this question will require a well-established concept from large deviation and information theory. This concept is defined next.

Definition 4.2.1 (Exponential Approximation)**.** Given two functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we let $\doteq$ denote the exponential approximation which is defined as:

$$f(n) \doteq g(n) \Leftrightarrow \lim_{n\to\infty} \frac{1}{n} \log\left(f(n)\right) = \lim_{n\to\infty} \frac{1}{n} \log\left(g(n)\right)$$

where the limits are assumed to be well-defined.

Let $\mathcal{P}_\mathcal{X}$ be the probability simplex of pmfs on $\mathcal{X}$, and $\mathcal{P}_\mathcal{Y}$ be the probability simplex of pmfs on $\mathcal{Y}$. Using the notation from Definition 4.1.1, simple combinatorial arguments can be used to derive the probabilities of the following empirical marginal distributions [26]:

$$\forall p_X \in \mathcal{P}_\mathcal{X}, \;\; P_{X,Y}^n\left(\hat{P}_{X_1^n} = p_X\right) \doteq e^{-nD(p_X||P_X)} \tag{4.10}$$

$$\forall p_Y \in \mathcal{P}_\mathcal{Y}, \;\; P_{X,Y}^n\left(\hat{P}_{Y_1^n} = p_Y\right) \doteq e^{-nD(p_Y||P_Y)} \tag{4.11}$$

and empirical conditional distributions:

$$\forall x \in \mathcal{X}, \forall p_X \in \mathcal{P}_{\mathcal{X}}, \forall v_x \in \mathcal{P}_{\mathcal{Y}}, \;\; P_{X,Y}^n \left(V_x = v_x \mid \hat{P}_{X_1^n} = p_X \right) \doteq e^{-np_X(x)D(v_x||W_x)} \tag{4.12}$$

where we require the conditioning on $\hat{P}_{X_1^n} = p_X$ in equation 4.12 because we need to know $\sum_{i=1}^n \mathbb{I}(X_i = x) = np_X(x)$ in order to define V_x. Note that equations 4.10, 4.11, and 4.12 are (rigorously) true when p_X, p_Y, and v_x are type classes. However, we do not use a rigorous treatment in this chapter for the sake of brevity. Using equation 4.12, we derive the exponential approximation of $P_{X,Y}^n \left(V = v \mid \hat{P}_{X_1^n} = p_X \right)$, where $v = \left[v_1 \cdots v_{|\mathcal{X}|} \right]$, and $\{V = v\}$ denotes the event $\{V = v\} = \{V_1 = v_1\} \cap \cdots \cap \left\{ V_{|\mathcal{X}|} = v_{|\mathcal{X}|} \right\}$. For any $p_X \in \mathcal{P}_{\mathcal{X}}$, and for any $v = \left[v_1 \cdots v_{|\mathcal{X}|} \right]$ such that $\forall x \in \mathcal{X}, \;\; v_x \in \mathcal{P}_{\mathcal{Y}}$, we have:

$$P_{X,Y}^n \left(V = v \mid \hat{P}_{X_1^n} = p_X \right) = \prod_{x \in \mathcal{X}} P_{X,Y}^n \left(V_x = v_x \mid \hat{P}_{X_1^n} = p_X \right) \tag{4.13}$$

because the probabilities of empirical conditional distributions for different values of x are conditionally independent given $\hat{P}_{X_1^n} = p_X$. This can be understood by considering how we infer information about V_2 from V_1, for example. Let $N(X = x)$ denote the number of samples in X_1^n which are equal to x. For fixed n, we can infer information about $N(X = 1)$ given $V_1 = v_1$, and this gives information about $N(X = 2)$. This in turn provides information about V_2. So, the way in which different empirical conditional distributions reveal information about a particular V_x is by narrowing down the possible choices for $N(X = x)$. We cannot get any other information about the distribution of the Y_i within these $N(X = x)$ samples because the samples are drawn i.i.d. Due to the conditioning in equation 4.13, we know n and $\hat{P}_{X_1^n} = p_X$, which means we also know $N(X = x)$ for all $x \in \mathcal{X}$. Given this information, as the (X_1^n, Y_1^n) are drawn i.i.d., the different V_x become conditionally independent.

We slightly digress to point out that asymptotically:

$$P_{X,Y}^n \left(V = v \right) = \prod_{x \in \mathcal{X}} P_{X,Y}^n \left(V_x = v_x \right)$$

is also intuitively true, because the events $\{V_x = v_x\}$ are asymptotically independent as $n \to \infty$. This is certainly not a trivial observation. For example, if n is fixed, we can infer information about $N(X = 1)$ from V_1 and this gives information about $N(X = 2)$. So, we effectively get some information about V_2 from V_1, as discussed earlier. However, if we let $n \to \infty$, then the laws of large numbers imply that there are roughly $nP_X(1)$ samples where $X_i = 1$ and $nP_X(2)$ samples where $X_i = 2$. Since the samples are drawn i.i.d., the resulting empirical conditional pmfs, V_1 and V_2, are independent. This result can be rigorously established, but we do not pursue it here.

Returning to our discussion, using equations 4.12 and 4.13, we have:

$$P_{X,Y}^n \left(V = v \mid \hat{P}_{X_1^n} = p_X \right) \doteq \exp\left(-n \sum_{x \in \mathcal{X}} p_X(x) D(v_x||W_x) \right). \tag{4.14}$$

We can write this expression more elegantly using the notation for conditional KL divergence. This is defined next.

Definition 4.2.2 (Discrete Conditional KL Divergence)**.** Given a marginal pmf P_X on the countable set $\mathcal{X}$, and two sets of conditional pmfs V and W, representing $\{V_x : x \in \mathcal{X}\}$ and $\{W_x : x \in \mathcal{X}\}$ respectively, such that V_x and W_x are conditional pmfs on the countable set $\mathcal{Y}$ given $x \in \mathcal{X}$, the conditional KL divergence between V and W, denoted $D(V||W|P_X)$, is given by:

$$D(V||W|P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(V_x||W_x).$$

Using the notation from Definition 4.1.2, we rewrite equation 4.14 as:

$$P^n_{X,Y}\left(V = v \mid \hat{P}_{X_1^n} = p_X\right) \doteq e^{-nD(v||W|p_X)}. \tag{4.15}$$

Next, we consider $P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y\right)$ more closely. From the total probability law, we get:

$$\forall p_Y \in \mathcal{P}_\mathcal{Y}, \;\; P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y\right) = \sum_{\substack{p_X, v: \\ vp_X = p_Y}} P^n_{X,Y}\left(\hat{P}_{X_1^n} = p_X\right) P^n_{X,Y}\left(V = v \mid \hat{P}_{X_1^n} = p_X\right)$$

which using equation 4.10 and 4.15, we can write (non-rigorously) as:

$$\forall p_Y \in \mathcal{P}_\mathcal{Y}, \;\; P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y\right) = \sum_{\substack{p_X, v: \\ vp_X = p_Y}} e^{-n(D(p_X||P_X) + D(v_x||W|p_X))}$$

where the sum indexes over a polynomial in n number of terms, because the number of possible empirical pmfs p_X and $v = \begin{bmatrix} v_1 \cdots v_{|\mathcal{X}|} \end{bmatrix}$ which satisfy $vp_X = p_Y$ are both polynomial in n for fixed n. As all the terms in the sum decay exponentially, the term with the slowest decay rate dominates the sum by the Laplace principle [26]. This gives us:

$$\forall p_Y \in \mathcal{P}_\mathcal{Y}, \;\; P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y\right) \doteq \exp\left(-n \min_{\substack{p_X, v: \\ vp_X = p_Y}} D(p_X||P_X) + D(v||W|p_X)\right). \tag{4.16}$$

However, comparing equations 4.11 and 4.16, we have:

$$\forall p_Y \in \mathcal{P}_\mathcal{Y}, \;\; D(p_Y||P_Y) = \min_{\substack{p_X, v: \\ vp_X = p_Y}} D(p_X||P_X) + D(v||W|p_X) \tag{4.17}$$

where the minimization is over all $p_X \in \mathcal{P}_\mathcal{X}$ and all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$. This is an appealing decomposition of the output perturbation in terms of the source perturbation and the channel perturbation. The perturbations are measured in KL divergence and the minimization intuitively identifies the most probable source-channel decomposition.

We now set up the problem of finding the most probable source and channel empirical pmfs, $\hat{P}^*_{X_1^n}$ and V^*, such that $V^* \hat{P}^*_{X_1^n} = p_Y$, where $\hat{P}_{Y_1^n} = p_Y$ is the observed output empirical pmf. As one would intuitively expect, the problem turns out to equivalent to the right hand side of equation 4.17. We want to maximize $P^n_{X,Y}(\hat{P}_{X_1^n} = p_X, V = v \mid \hat{P}_{Y_1^n} = p_Y)$ over all $p_X \in \mathcal{P}_\mathcal{X}$ and all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$ such that $vp_X = p_Y$. Using Bayes' rule, we have:

$$P^n_{X,Y}\left(\hat{P}_{X_1^n} = p_X, V = v \mid \hat{P}_{Y_1^n} = p_Y\right) = \begin{cases} \frac{P^n_{X,Y}\left(\hat{P}_{X_1^n} = p_X\right) P^n_{X,Y}\left(V = v | \hat{P}_{X_1^n} = p_X\right)}{P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y\right)} & , \text{ if } vp_X = p_Y \\ 0 & , \text{ otherwise} \end{cases}$$

because:

$$P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y \mid \hat{P}_{X_1^n} = p_X, V = v\right) = \begin{cases} 1 & , \text{ if } vp_X = p_Y \\ 0 & , \text{ otherwise} \end{cases}.$$

Hence, assuming that the constraint $vp_X = p_Y$ is satisfied, we get:

$$P^n_{X,Y}\left(\hat{P}_{X_1^n} = p_X, V = v \mid \hat{P}_{Y_1^n} = p_Y\right) \doteq e^{-n(D(p_X||P_X)+D(v||W|p_X)-D(p_Y||P_Y))} \tag{4.18}$$

using equations 4.10, 4.11, and 4.15. Note that $D(p_Y||P_Y)$ is a constant as $\hat{P}_{Y_1^n} = p_Y$ is given. Hence, to maximize $P^n_{X,Y}(\hat{P}_{X_1^n} = p_X, V = v \mid \hat{P}_{Y_1^n} = p_Y)$, we must minimize the remaining portion of the exponent in equation 4.18. Our optimization problem to find $\hat{P}^*_{X_1^n}$ and V^* is thus:

$$\min_{\substack{p_X, v: \\ vp_X = p_Y}} D(p_X||P_X) + D(v||W|p_X) \tag{4.19}$$

where $\hat{P}_{Y_1^n} = p_Y$ is known, and the minimization is over all $p_X \in \mathcal{P}_\mathcal{X}$ and all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$. The problem in statement 4.19 is precisely the right hand side of equation 4.17. In fact, equation 4.17 illustrates that the minimum value of the objective function is $D(p_Y||P_Y)$. So, we simply need to recover the optimal arguments, $\hat{P}^*_{X_1^n}$ and V^*, which generate this minimum value. This is equivalent to finding the most probable source and channel perturbations, $J^*_X = \hat{P}^*_{X_1^n} - P_X$ and $J^*_W = V^* - W$, when the output empirical distribution is perturbed by $J_Y = p_Y - P_Y$. We note that we mean "most probable" in the exponential approximation sense, as is evident from our derivation of the problem. The next subsection solves problem 4.19.

4.2.1 Global Solution using Information Projection

For the sake of clarity, in our ensuing derivations, we will use the notation $\hat{P}_{X_1^n}$, $\hat{P}_{Y_1^n}$, and V, instead of p_X, p_Y, and v, to mean particular values of the empirical distributions rather than random variables which represent these empirical distributions. This should not generate any confusion since we will not compute any large deviation probabilities in this subsection. Using this notation, we can rewrite the optimization problem in statement 4.19 as:

$$\min_{\substack{\hat{P}_{X_1^n}, V: \\ V\hat{P}_{X_1^n} = \hat{P}_{Y_1^n}}} D(\hat{P}_{X_1^n}||P_X) + D(V||W|\hat{P}_{X_1^n}) \tag{4.20}$$

where $\hat{P}_{Y_1^n}$ is known, and the minimization is over all empirical pmfs $\hat{P}_{X_1^n}$ and all empirical channels (conditional pmfs) V. The optimizing $\hat{P}^*_{X_1^n}$ and V^* are the most probable (in the exponential approximation sense) source and channel empirical pmfs given the output empirical pmf is $\hat{P}_{Y_1^n}$.

We will globally solve problem 4.20 by setting it up as an information projection (i-projection). The i-projection is a highly useful notion which imparts elegant geometric structure into many large deviation theoretic arguments in information theory; [1] and [10] are both wonderful resources which introduce it. We will assume some familiarity with the i-projection in this chapter. The next theorem solves problem 4.20 using i-projections.

Theorem 4.2.1 (Most Probable Empirical Source and Channel PMFs)**.** *The most probable empirical source pmf and empirical channel conditional pmfs, $\hat{P}^*_{X_1^n}$ and V^*, which are global optimizing arguments of the extremal problem in statement 4.20, are given by:*

$$\begin{aligned} \forall x \in \mathcal{X},\ \hat{P}^*_{X_1^n}(x) &= P_X(x)\left(1 + \sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)}\right), \text{ and} \\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V^*_x(y) &= W_x(y)\left(\frac{1 + \frac{J_Y(y)}{P_Y(y)}}{1 + \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}}\right), \end{aligned}$$

where $J_Y = \hat{P}_{Y_1^n} - P_Y$, and the optimal value of the objective function of problem 4.20 is:

$$D(\hat{P}^*_{X_1^n}||P_X) + D(V^*||W|\hat{P}^*_{X_1^n}) = D(\hat{P}_{Y_1^n}||P_Y).$$

Proof.
The optimal value of the objective function holds from equation 4.17. To compute the optimizing arguments, we use the chain rule for KL divergence on the objective function of problem 4.20 to get:

$$D(\hat{P}_{X_1^n}||P_X) + D(V||W|\hat{P}_{X_1^n}) = D(\hat{P}_{X_1^n,Y_1^n}||P_{X,Y}).$$

The constraint $V\hat{P}_{X_1^n} = \hat{P}_{Y_1^n}$ is a constraint on the marginal pmf of Y for $\hat{P}_{X_1^n,Y_1^n}$. It can be written as a linear family of distributions [10]:

$$\mathcal{L} \triangleq \left\{Q_{X,Y} : Q_Y = \hat{P}_{Y_1^n}\right\} = \left\{Q_{X,Y} : \forall y \in \mathcal{Y},\ \mathbb{E}_{Q_{X,Y}}\left[\mathbb{I}(Y = y)\right] = \hat{P}_{Y_1^n}(y)\right\}$$

where $Q_{X,Y}$ denotes any joint pmf on $\mathcal{X}\times\mathcal{Y}$. With this linear family $\mathcal{L}$, problem 4.20 can be recast as an i-projection:

$$\hat{P}^*_{X_1^n,Y_1^n} = \underset{\hat{P}_{X_1^n,Y_1^n}\in\mathcal{L}}{\arg\min}\ D(\hat{P}_{X_1^n,Y_1^n}||P_{X,Y}).$$

Solving this i-projection produces the unique solution:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \hat{P}^*_{X_1^n,Y_1^n}(x,y) = P_{X|Y}(x|y)\hat{P}_{Y_1^n}(y)$$

where we substitute $\hat{P}_{Y_1^n} = P_Y + J_Y$ to get:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \hat{P}^*_{X_1^n,Y_1^n}(x,y) = P_{X,Y}(x,y)\left(1 + \frac{J_Y(y)}{P_Y(y)}\right).$$

Marginalizing this solution gives:

$$\forall x \in \mathcal{X},\ \hat{P}^*_{X_1^n}(x) = P_X(x)\left(1 + \sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)}\right) = P_X(x)\left(1 + \mathbb{E}_{W_x}\left[\frac{J_Y(Y)}{P_Y(Y)}\right]\right)$$

and the definition of conditional probability gives:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; V_x^*(y) = \frac{\hat{P}^*_{X_1^n, Y_1^n}(x,y)}{\hat{P}^*_{X_1^n}(x)} = W_x(y) \left(\frac{1 + \frac{J_Y(y)}{P_Y(y)}}{1 + \sum_{z \in \mathcal{Y}} W_x(z) \frac{J_Y(z)}{P_Y(z)}} \right)$$

which are the global optimizing solutions of problem 4.20. □

We note that there is a more elementary (but perhaps less insightful) method of proving Theorem 4.2.1. We can directly derive the optimal value of the objective function of problem 4.20 (or equation 4.17) using the chain rule for KL divergence. Observe that for any two distributions $Q_{X,Y}$ and $P_{X,Y}$, the chain rule gives:

$$\begin{aligned} D(Q_{X,Y}||P_{X,Y}) &= D(Q_X||P_X) + D(Q_{Y|X}||P_{Y|X}|Q_X) \\ &= D(Q_Y||P_Y) + D(Q_{X|Y}||P_{X|Y}|Q_Y) \end{aligned}$$

which by Gibbs' inequality, implies that:

$$D(Q_Y||P_Y) \leq D(Q_X||P_X) + D(Q_{Y|X}||P_{Y|X}|Q_X)$$

with equality if and only if $D(Q_{X|Y}||P_{X|Y}|Q_Y) = 0$. Suppose we fix $P_{X,Y}$ and Q_Y, and minimize the right hand side of this inequality over all Q_X and $Q_{Y|X}$ such that the marginal distribution of Y of $Q_{X,Y}$ is Q_Y. Then, choosing $Q_{X|Y} = P_{X|Y}$ ensures that $D(Q_{X|Y}||P_{X|Y}|Q_Y) = 0$, which means that:

$$D(Q_Y||P_Y) = \min_{\substack{Q_X, Q_{Y|X}: \\ Q_{Y|X} Q_X = Q_Y}} D(Q_X||P_X) + D(Q_{Y|X}||P_{Y|X}|Q_X).$$

This is precisely equation 4.17 and the optimal value statement in Theorem 4.2.1. Moreover, the optimizing $Q^*_{X,Y}$ has marginal distribution $Q^*_Y = Q_Y$ and conditional distributions $Q^*_{X|Y} = P_{X|Y}$, as we saw in the proof of Theorem 4.2.1. Hence, this is an alternative derivation of Theorem 4.2.1.

Although Theorem 4.2.1 states that the minimum value of the objective function of problem 4.20 is $D(\hat{P}_{Y_1^n}||P_Y) = D(\hat{P}^*_{X_1^n}||P_X) + D(V^*||W|\hat{P}^*_{X_1^n})$, it is instructive to closely examine what $D(\hat{P}^*_{X_1^n}||P_X)$ and $D(V^*||W|\hat{P}^*_{X_1^n})$ are. To derive more insight on these quantities, we employ local approximations of KL divergence and solve problem 4.20 in the local case in subsection 4.2.2. For now, we form local approximations of the global solution directly. This can be compared to the results of subsection 4.2.2 later. We first reproduce the solutions in Theorem 4.2.1 by reinserting the $\epsilon > 0$ factor:

$$\forall x \in \mathcal{X}, \;\; \hat{P}^*_{X_1^n}(x) = P_X(x) + \epsilon \sum_{y \in \mathcal{Y}} P_{X|Y}(x|y) J_Y(y) \tag{4.21}$$

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; V_x^*(y) = W_x(y) \left(\frac{1 + \epsilon \frac{J_Y(y)}{P_Y(y)}}{1 + \epsilon \sum_{z \in \mathcal{Y}} W_x(z) \frac{J_Y(z)}{P_Y(z)}} \right) \tag{4.22}$$

from which we see that $\hat{P}^*_{X_1^n}$ is already in local form because the second term which constitutes it is a valid additive perturbation:

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X|Y}(x|y)J_Y(y) = \sum_{y\in\mathcal{Y}} J_Y(y)\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) = \sum_{y\in\mathcal{Y}} J_Y(y) = 0.$$

The $V^*_x,\ x\in\mathcal{X}$ are not in local form because we cannot easily isolate additive, log-likelihood, or normalized perturbations from them. To locally approximate them, we use the following substitution:

$$\frac{1}{1+\epsilon\sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}} = 1-\epsilon\sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)} + o(\epsilon)$$

where $o(\epsilon)$ denotes a function which satisfies $\lim_{\epsilon\to 0^+}\frac{o(\epsilon)}{\epsilon} = 0$. The substitution produces:

$$\forall x\in\mathcal{X}, \forall y\in\mathcal{Y},\ V^*_x(y) = W_x(y)\left(1+\epsilon\frac{J_Y(y)}{P_Y(y)} - \epsilon\sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)} + o(\epsilon)\right)$$

which we rearrange to get:

$$\forall x\in\mathcal{X}, \forall y\in\mathcal{Y},\ V^*_x(y) = W_x(y) + \epsilon\left(W_x(y)\frac{J_Y(y)}{P_Y(y)} - W_x(y)\sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}\right) + o(\epsilon) \tag{4.23}$$

where the second term on the right hand side is a valid additive perturbation:

$$\sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)} - \sum_{y\in\mathcal{Y}} W_x(y)\sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)} = \sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)} - \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)} = 0.$$

Hence, equations 4.21 and 4.23 are the local approximations to the global solution in Theorem 4.2.1. We rewrite them below by again neglecting the ϵ factors and terms:

$$\forall x\in\mathcal{X},\ \hat{P}^*_{X_1^n}(x) = P_X(x) + P_X(x)\sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)} \tag{4.24}$$

$$\forall x\in\mathcal{X}, \forall y\in\mathcal{Y},\ V^*_x(y) \overset{\text{local}}{=} W_x(y) + W_x(y)\left(\frac{J_Y(y)}{P_Y(y)} - \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}\right) \tag{4.25}$$

where equation 4.24 is identical to the global solution, but equation 4.25 requires local approximations. Recognizing that $\hat{P}^*_{X_1^n} = P_X + J_X$ and $\forall x\in\mathcal{X},\ V^*_x = W_x + J^*_x$, the most probable local perturbations are:

$$\forall x\in\mathcal{X},\ J^*_X(x) = P_X(x)\sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)} \tag{4.26}$$

$$\forall x\in\mathcal{X}, \forall y\in\mathcal{Y},\ J^*_x(y) = W_x(y)\left(\frac{J_Y(y)}{P_Y(y)} - \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}\right) \tag{4.27}$$

using equations 4.24 and 4.25. In the next subsection, we verify that equations 4.26 and 4.27 are indeed the solutions to the locally approximated problem 4.20.

4.2.2 Local Solution using Lagrangian Optimization

To locally approximate problem 4.20, we assume that empirical distributions perturb very slightly from theoretical distributions. This is a sound assumption when $n \to \infty$. Recall that we have:

$$\begin{aligned} \hat{P}_{X_1^n} &= P_X + J_X \\ \hat{P}_{Y_1^n} &= P_Y + J_Y \\ \forall x \in \mathcal{X},\ V_x &= W_x + J_x \end{aligned}$$

where the additive perturbation vectors have elements that are small (or the ϵ factor which is understood to be in front of them is small). This means that KL divergences can be approximated using χ^2-divergences.

$$D(\hat{P}_{X_1^n}||P_X) \overset{\text{local}}{=} \frac{1}{2}\|J_X\|^2_{P_X} \tag{4.28}$$

$$D(\hat{P}_{Y_1^n}||P_Y) \overset{\text{local}}{=} \frac{1}{2}\|J_Y\|^2_{P_Y} \tag{4.29}$$

$$\forall x \in \mathcal{X},\ D(V_x||W_x) \overset{\text{local}}{=} \frac{1}{2}\|J_x\|^2_{W_x} \tag{4.30}$$

We may also locally approximate the conditional KL divergence, $D(V||W|\hat{P}_{X_1^n})$. Using Definition 4.2.2 and equation 4.30, we have:

$$D(V||W|\hat{P}_{X_1^n}) = \sum_{x\in\mathcal{X}} \hat{P}_{X_1^n}(x) D(V_x||W_x) \overset{\text{local}}{=} \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\|J_x\|^2_{W_x} + \frac{1}{2}\sum_{x\in\mathcal{X}} J_X(x)\|J_x\|^2_{W_x}$$

which gives:

$$D(V||W|\hat{P}_{X_1^n}) \overset{\text{local}}{=} \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\|J_x\|^2_{W_x} \tag{4.31}$$

where we neglect the third order terms $J_X(x)\|J_x\|^2_{W_x}$ (as they are $o\left(\epsilon^2\right)$ terms when we write the ϵ factors explicitly). Equations 4.28 and 4.31 can be used to recast the extremal problem in statement 4.20 in local form:

$$\begin{aligned} \min_{J_X, J_W=\left[J_1 \cdots J_{|\mathcal{X}|}\right]} \quad & \frac{1}{2}\|J_X\|^2_{P_X} + \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\|J_x\|^2_{W_x} \\ \text{subject to:} \quad & J_X^T 1 = 0, \\ & J_W^T 1 = 0, \\ \text{and} \quad & WJ_X + J_W P_X = J_Y \end{aligned} \tag{4.32}$$

where 1 denotes the vector of with all entries equal to 1, 0 is a scalar in the first constraint and a column vector with all entries equal to 0 in the second constraint, and the third constraint imposes equation 4.9 which ensures that $\hat{P}_{Y_1^n}$ is the observed output empirical pmf. The next theorem solves the optimization problem in statement 4.32.

Theorem 4.2.2 (Most Probable Local Source and Channel Perturbations)**.** *The most probable local source perturbation and local channel perturbations, which are optimizing arguments of the*

extremal problem in statement 4.32, are given by:

$$\begin{aligned} \forall x \in \mathcal{X},\ J_X^*(x) &= P_X(x) \sum_{y \in \mathcal{Y}} W_x(y) \frac{J_Y(y)}{P_Y(y)},\ \textit{and} \\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ J_x^*(y) &= W_x(y) \left(\frac{J_Y(y)}{P_Y(y)} - \sum_{z \in \mathcal{Y}} W_x(z) \frac{J_Y(z)}{P_Y(z)} \right), \end{aligned}$$

and the optimal value of the objective function of problem 4.32 is:

$$\frac{1}{2} \left\| J_X^* \right\|_{P_X}^2 + \frac{1}{2} \sum_{x \in \mathcal{X}} P_X(x) \left\| J_x^* \right\|_{W_x}^2 = \frac{1}{2} \left\| J_Y \right\|_{P_Y}^2 .$$

Proof.
We first set up the Lagrangian, $\mathcal{L}\left(J_X, J_W, \lambda, \mu, \tau\right)$, for problem 4.32:

$$\mathcal{L} = J_X^T \left[P_X\right]^{-1} J_X + \sum_{x \in \mathcal{X}} P_X(x) J_x^T \left[W_x\right]^{-1} J_x + \lambda^T \left(W J_X + J_W P_X - J_Y\right) + \mu J_X^T 1 + \tau^T J_W^T 1$$

where $\mu \in \mathbb{R}$, $\lambda = \left[\lambda_1 \cdots \lambda_{|\mathcal{Y}|}\right]^T \in \mathbb{R}^{|\mathcal{Y}|}$, and $\tau = \left[\tau_1 \cdots \tau_{|\mathcal{X}|}\right]^T \in \mathbb{R}^{|\mathcal{X}|}$ are Lagrange multipliers, and we omit the factor of $\frac{1}{2}$ in the objective function. We will take partial derivatives of $\mathcal{L}$ using denominator layout notation, which means:

$$\frac{\partial \mathcal{L}}{\partial J_X} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial J_X(1)} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial J_X(|\mathcal{X}|)} \end{bmatrix} \text{ and } \forall x \in \mathcal{X},\ \frac{\partial \mathcal{L}}{\partial J_x} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial J_x(1)} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial J_x(|\mathcal{Y}|)} \end{bmatrix}.$$

Taking partial derivatives and setting them equal to 0, we get:

$$\frac{\partial \mathcal{L}}{\partial J_X} = 2 \left[P_X\right]^{-1} J_X + W^T \lambda + \mu 1 = 0$$

and:

$$\forall x \in \mathcal{X},\ \frac{\partial \mathcal{L}}{\partial J_x} = 2 P_X(x) \left[W_x\right]^{-1} J_x + P_X(x) \lambda + \tau_x 1 = 0.$$

Appropriately absorbing constants into the Lagrange multipliers, we have:

$$\begin{aligned} J_X &= \left[P_X\right] W^T \lambda + \mu P_X, \\ \forall x \in \mathcal{X},\ J_x &= \left[W_x\right] \lambda + \tau_x W_x. \end{aligned}$$

We then impose the constraint, $J_X^T 1 = 0$:

$$J_X^T 1 = \lambda^T W \left[P_X\right] 1 + \mu P_X^T 1 = \lambda^T P_Y + \mu = 0$$

which gives:

$$\mu = -\lambda^T P_Y.$$

We also impose the constraint, $J_W^T 1 = 0$ (or $\forall x \in \mathcal{X},\ J_x^T 1 = 0$):

$$\forall x \in \mathcal{X},\ J_x^T 1 = \lambda^T \left[W_x\right] 1 + \tau_x W_x^T 1 = \lambda^T W_x + \tau_x = 0$$

which gives:

$$\forall x \in \mathcal{X}, \tau_x = -\lambda^T W_x.$$

Hence, we have:

$$\begin{aligned} J_X &= \left([P_X]\, W^T - P_X P_Y^T\right) \lambda, \\ \forall x \in \mathcal{X},\ J_x &= \left([W_x] - W_x W_x^T\right) \lambda. \end{aligned}$$

Finally, we impose the constraint, $WJ_X + J_W P_X = WJ_X + \sum_{x\in\mathcal{X}} P_X(x) J_x = J_Y$:

$$WJ_X + J_W P_X = W\left([P_X]\, W^T - P_X P_Y^T\right)\lambda + \sum_{x\in\mathcal{X}} P_X(x)\left([W_x] - W_x W_x^T\right)\lambda = J_Y$$

$$\left(W\,[P_X]\, W^T - P_Y P_Y^T + [P_Y] - \sum_{x\in\mathcal{X}} P_X(x) W_x W_x^T\right)\lambda = J_Y$$

which is equivalent to:

$$\left([P_Y] - P_Y P_Y^T\right)\lambda = J_Y.$$

To find λ, we must invert the matrix $[P_Y] - P_Y P_Y^T$, which is a rank 1 perturbation of the matrix $[P_Y]$. It is tempting to apply the Sherman-Morrison-Woodbury formula to invert $[P_Y] - P_Y P_Y^T$, but we find that $1 - P_Y^T\,[P_Y]^{-1}\,P_Y = 0$. This means that the Sherman-Morrison-Woodbury formula cannot invert $[P_Y] - P_Y P_Y^T$ because it is singular. On the other hand, we observe that:

$$\lambda = [P_Y]^{-1}\, J_Y$$

satisfies the constraint $\left([P_Y] - P_Y P_Y^T\right)\lambda = J_Y$:

$$\left([P_Y] - P_Y P_Y^T\right)[P_Y]^{-1}\, J_Y = J_Y - P_Y 1^T J_Y = J_Y$$

where we use the fact that $1^T J_Y = 0$, because J_Y is a valid additive perturbation. Therefore, the optimal J_X and $J_x,\ x \in \mathcal{X}$ are:

$$\begin{aligned} J_X^* &= [P_X]\, W^T\, [P_Y]^{-1}\, J_Y, \\ \forall x \in \mathcal{X},\ J_x^* &= \left([W_x] - W_x W_x^T\right)[P_Y]^{-1} J_Y. \end{aligned}$$

which we can rewrite as:

$$\begin{aligned} \forall x \in \mathcal{X},\ J_X^*(x) &= P_X(x) \sum_{y\in\mathcal{Y}} W_x(y) \frac{J_Y(y)}{P_Y(y)}, \\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ J_x^*(y) &= W_x(y)\left(\frac{J_Y(y)}{P_Y(y)} - \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}\right). \end{aligned}$$

This completes the proof of the optimizing arguments.

From Theorem 4.2.1, we have that the global solution satisfies:

$$D(\hat{P}^*_{X_1^n}||P_X) + D(V^*||W|\hat{P}^*_{X_1^n}) = D(\hat{P}_{Y_1^n}||P_Y).$$

In analogy with this relation, we would like the local solution to satisfy:

$$\frac{1}{2}\left\|J_X^*\right\|_{P_X}^2 + \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\left\|J_x^*\right\|_{W_x}^2 = \frac{1}{2}\left\|J_Y\right\|_{P_Y}^2$$

which is clearly equivalent to satisfying:

$$\left\|J_X^*\right\|_{P_X}^2 + \sum_{x\in\mathcal{X}} P_X(x)\left\|J_x^*\right\|_{W_x}^2 = \left\|J_Y\right\|_{P_Y}^2. \tag{4.33}$$

We now directly prove equation 4.33. Observe that:

$$\begin{aligned}
\left\|J_X^*\right\|_{P_X}^2 &= J_Y^T[P_Y]^{-1}W[P_X]W^T[P_Y]^{-1}J_Y \\
&= \sum_{x\in\mathcal{X}} P_X(x)\left(\sum_{y\in\mathcal{Y}} W_x(y)\frac{J_Y(y)}{P_Y(y)}\right)^2 \\
&= \mathbb{VAR}\left(\mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right]\right)
\end{aligned} \tag{4.34}$$

where the third equality follows from $\mathbb{E}\left[\mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right]\right] = \mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\right] = 0$. Likewise, observe that:

$$\begin{aligned}
\sum_{x\in\mathcal{X}} P_X(x)\left\|J_x^*\right\|_{W_x}^2 &= \sum_{x\in\mathcal{X}} P_X(x)\sum_{y\in\mathcal{Y}} W_x(y)\left(\frac{J_Y(y)}{P_Y(y)} - \sum_{z\in\mathcal{Y}} W_x(z)\frac{J_Y(z)}{P_Y(z)}\right)^2 \\
&= \mathbb{E}\left[\left(\frac{J_Y(Y)}{P_Y(Y)} - \mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right]\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{VAR}\left(\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right)\right] \\
&= \mathbb{VAR}\left(\frac{J_Y(Y)}{P_Y(Y)}\right) - \mathbb{VAR}\left(\mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right]\right) \\
&= \left\|J_Y\right\|_{P_Y}^2 - \mathbb{VAR}\left(\mathbb{E}\left[\frac{J_Y(Y)}{P_Y(Y)}\middle| X\right]\right)
\end{aligned} \tag{4.35}$$

where the third equality holds by the tower property, the fourth equality holds by the law of total variance, and the last equality holds because $\mathbb{VAR}\left(\frac{J_Y(Y)}{P_Y(Y)}\right) = \left\|J_Y\right\|_{P_Y}^2$. Adding equations 4.34 and 4.35 produces equation 4.33. This completes the proof. □

The most probable source and channel perturbations given in Theorem 4.2.2 are the same as those in equations 4.26 and 4.27. Thus, the solutions to the locally approximated global problem are consistent with the local approximations to the global solutions. We next consider explicitly identifying the local approximations of $D(\hat{P}^*_{X_1^n}||P_X)$ and $D(V^*||W|\hat{P}^*_{X_1^n})$. Using equation 4.28 and Theorem 4.2.2, we have:

$$D(\hat{P}^*_{X_1^n}||P_X) \overset{\text{local}}{=} \frac{1}{2}\left\|J_X^*\right\|_{P_X}^2 = \frac{1}{2}\left\|[P_X]W^T[P_Y]^{-1}J_Y\right\|_{P_X}^2 = \frac{1}{2}J_Y^T[P_Y]^{-1}W[P_X]W^T[P_Y]^{-1}J_Y.$$

Recall from section 1.2 in chapter 1 that:

$$K_Y = \left[\sqrt{P_Y}\right]^{-1} J_Y$$

is the normalized perturbation corresponding to the additive perturbation J_Y. Using K_Y, we get:

$$D(\hat{P}^*_{X_1^n}||P_X) \overset{\text{local}}{=} \frac{1}{2}\left\|J^*_X\right\|^2_{P_X} = \frac{1}{2}K_Y^T BB^T K_Y \tag{4.36}$$

where B is the DTM defined in Definition 1.3.4. Likewise, using equation 4.31 and Theorem 4.2.2, we have:

$$D(V^*||W|\hat{P}^*_{X_1^n}) \overset{\text{local}}{=} \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\left\|J^*_x\right\|^2_{W_x} = \frac{1}{2}\left(\|J_Y\|^2_{P_Y} - \left\|J^*_X\right\|^2_{P_X}\right) = \frac{1}{2}\left(K_Y^T K_Y - K_Y^T BB^T K_Y\right)$$

which simplifies to:

$$D(V^*||W|\hat{P}^*_{X_1^n}) \overset{\text{local}}{=} \frac{1}{2}\sum_{x\in\mathcal{X}} P_X(x)\left\|J^*_x\right\|^2_{W_x} = \frac{1}{2}K_Y^T\left(I - BB^T\right)K_Y. \tag{4.37}$$

Hence, equations 4.36 and 4.37 illustrate local approximations of the KL divergences, $D(\hat{P}^*_{X_1^n}||P_X)$ and $D(V^*||W|\hat{P}^*_{X_1^n})$. They intuitively convey that in the most probable scenario (under the exponential approximation), BB^T controls the fraction of local output KL divergence arising from the source perturbation, and $I - BB^T$ controls the fraction of local output KL divergence arising from the channel perturbations. Hence, under local approximations, BB^T determines the exponent of the large deviation probability of observing $\hat{P}^*_{X_1^n}$, and $I - BB^T$ determines the exponent of the large deviation probability of observing V^*.

Finally, we portray that BB^T and $I - BB^T$ also govern the source-channel decomposition of the output perturbation. From equation 4.9 and the third constraint of problem 4.32, we have:

$$J_Y = WJ^*_X + J^*_W P_X \tag{4.38}$$

where $J^*_W = \left[J^*_1 \cdots J^*_{|\mathcal{X}|}\right]$, and we neglect all second order terms. We may rewrite this source-channel decomposition using the normalized perturbation K_Y:

$$K_Y = \left[\sqrt{P_Y}\right]^{-1} WJ^*_X + \left[\sqrt{P_Y}\right]^{-1} J^*_W P_X. \tag{4.39}$$

The source perturbation component of K_Y is:

$$\left[\sqrt{P_Y}\right]^{-1} WJ^*_X = \left[\sqrt{P_Y}\right]^{-1} W[P_X]W^T[P_Y]^{-1}J_Y = BB^T K_Y \tag{4.40}$$

using Theorem 4.2.2. This means the channel perturbation component of K_Y is:

$$\left[\sqrt{P_Y}\right]^{-1} J^*_W P_X = \left(I - BB^T\right)K_Y \tag{4.41}$$

because the proof of Theorem 4.2.2 ensures that equation 4.38 holds, which implies equation 4.39 also holds. Therefore, we may write equation 4.39 as:

$$K_Y = BB^T K_Y + \left(I - BB^T\right)K_Y. \tag{4.42}$$

This is a source-channel decomposition of the normalized output perturbation K_Y, where the source part is determined by BB^T, as shown in equation 4.40, and the channel part is determined by $I - BB^T$, as shown in equation 4.41.

4.3 Most Probable Channel Perturbations with Fixed Source

We now address the third question at the end of section 4.1. Suppose we observe both the i.i.d. input X_1^n and the i.i.d. output Y_1^n of the discrete memoryless channel W. A theoretical example of this would be a sender in a perfect feedback channel. Once again, we assume we have full knowledge of the theoretical distribution $P_{X,Y}$, and calculate all probabilities with respect to this distribution. In particular, we again use the notation $P_{X,Y}^n$ to denote the probability distribution of (X_1^n, Y_1^n). Observing X_1^n and Y_1^n reveals $\hat{P}_{X_1^n}$ and $\hat{P}_{Y_1^n}$, but there are many possible V which are consistent with $\hat{P}_{Y_1^n} = V\hat{P}_{X_1^n}$. In this scenario, the pertinent question is to try and find the most probable V which satisfies $\hat{P}_{Y_1^n} = V\hat{P}_{X_1^n}$.

We want to maximize $P_{X,Y}^n(V = v \mid \hat{P}_{X_1^n} = p_X, \hat{P}_{Y_1^n} = p_Y)$ over all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$ such that $vp_X = p_Y$, where p_X and p_Y are any observed input and output empirical pmfs. To this end, we have using Bayes' rule:

$$P_{X,Y}^n\left(V = v \mid \hat{P}_{X_1^n} = p_X, \hat{P}_{Y_1^n} = p_Y\right) = \begin{cases} \frac{P_{X,Y}^n\left(V=v|\hat{P}_{X_1^n}=p_Y\right)}{P_{X,Y}^n\left(\hat{P}_{Y_1^n}=p_Y|\hat{P}_{X_1^n}=p_X\right)} & , \text{ if } vp_X = p_Y \\ 0 & , \text{ otherwise} \end{cases}$$

because:

$$P_{X,Y}^n\left(\hat{P}_{Y_1^n} = p_Y \mid \hat{P}_{X_1^n} = p_X, V = v\right) = \begin{cases} 1 & , \text{ if } vp_X = p_Y \\ 0 & , \text{ otherwise} \end{cases}.$$

Moreover, $P_{X,Y}^n(\hat{P}_{Y_1^n} = p_Y \mid \hat{P}_{X_1^n} = p_X)$ is a constant because $\hat{P}_{X_1^n} = p_X$ and $\hat{P}_{Y_1^n} = p_Y$ are given. So, maximizing $P_{X,Y}^n(V = v \mid \hat{P}_{X_1^n} = p_X, \hat{P}_{Y_1^n} = p_Y)$ is equivalent to maximizing its numerator $P_{X,Y}^n(V = v \mid \hat{P}_{X_1^n} = p_X)$ assuming that the constraint $vp_X = p_Y$ is satisfied. Recall from equation 4.15 that:

$$P_{X,Y}^n\left(V = v \mid \hat{P}_{X_1^n} = p_X\right) \doteq e^{-nD(v||W|p_X)}.$$

Hence, maximizing $P_{X,Y}^n(V = v \mid \hat{P}_{X_1^n} = p_X)$ is equivalent to minimizing $D(v||W|p_X)$ over empirical channel matrices v, under the exponential approximation. This means our optimization problem to find the most probable empirical channel conditional pmfs V^* is:

$$\min_{v:\ vp_X = p_Y} D(v||W|p_X) \tag{4.43}$$

where $\hat{P}_{X_1^n} = p_X$ and $\hat{P}_{Y_1^n} = p_Y$ are known, and the minimization is over all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$. Finding V^* also delivers us the most probable channel perturbations $J_W^* = V^* - W$ when the input and output empirical pmfs are perturbed by $J_X = p_X - P_X$ and $J_Y = p_Y - P_Y$, respectively. J_W^* contains intriguing information regarding which conditional distributions perturb more than others.

Although problem 4.43 is soundly set up, it does not address the title of this section: "most probable channel perturbations with fixed source." Indeed, the source is not fixed in the setup of problem 4.43 as a perturbed empirical source pmf is observed. Another way to perceive this problem is to think of the fixed $\hat{P}_{X_1^n} = p_X$ as a fixed composition which does not come from i.i.d. X_1^n. The sender fixes an empirical distribution $\hat{P}_{X_1^n} = p_X$ and sends X_1^n according to this distribution. Each X_i is sent through the discrete memoryless channel. So, the theoretical input distribution is

$\hat{P}_{X_1^n}$ and the theoretical output distribution is $\hat{P}_Y = W\hat{P}_{X_1^n} = Wp_X$; the original P_X is meaningless in this scenario. The receiver observes the output empirical distribution $\hat{P}_{Y_1^n} = p_Y$ and tries to find the most probable empirical conditional distributions V given $\hat{P}_{X_1^n} = p_X$ and $\hat{P}_{Y_1^n} = p_Y$. What we have effectively done is assumed that there is no source perturbation. So, the output perturbation is only caused by the channel perturbations. This is the opposite scenario of [4] where the channel did not perturb but the source did.

We now derive the problem of finding the most probable V under this alternative interpretation. Let the theoretical joint pmf of (X, Y) be $P_{X,Y}$, where $P_{X,Y}$ now has marginal pmfs Q_X and $P_Y = WQ_X$, and conditional probabilities $P_{Y|X}$ given by W. Moreover, we assume that $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ P_{X,Y}(x, y) > 0$. We again calculate all probabilities with respect to $P^n_{X,Y}$, which denotes the distribution of (X_1^n, Y_1^n). Since the empirical input distribution equals the theoretical input distribution in the fixed composition scenario, we have:

$$Q_X = \hat{P}_{X_1^n}. \tag{4.44}$$

Moreover, we define:

$$Q_Y \triangleq \hat{P}_{Y_1^n} = P_Y + J_Y \tag{4.45}$$

as the empirical output pmf, where J_Y is now the additive perturbation of $Q_Y = \hat{P}_{Y_1^n}$ from P_Y. Note that:

$$Q_Y = VQ_X = (W + J_W)\, Q_X = P_Y + J_W Q_X$$

where J_W is the matrix of channel perturbations. Hence, the source-channel decomposition in equation 4.9 becomes:

$$J_Y = J_W Q_X \tag{4.46}$$

because the source perturbation is 0. We want to maximize $P^n_{X,Y}(V = v \mid Q_Y = q_Y)$ over all $v_x \in \mathcal{P}_\mathcal{Y}$ for each $x \in \mathcal{X}$ such that $vQ_X = q_Y$, where q_Y is any observed output empirical pmf. Using Bayes' rule, we have:

$$P^n_{X,Y}\left(V = v \mid Q_Y = q_Y\right) = \begin{cases} \frac{P^n_{X,Y}(V=v)}{P^n_{X,Y}(Q_Y=q_Y)} & ,\ \text{if } vQ_X = q_Y \\ 0 & ,\ \text{otherwise} \end{cases}$$

because:

$$P^n_{X,Y}\left(Q_Y = q_Y \mid V = v\right) = \begin{cases} 1 & ,\ \text{if } vQ_X = q_Y \\ 0 & ,\ \text{otherwise} \end{cases}$$

where we do not require any conditioning on $\hat{P}_{X_1^n}$ in the above equations because we use a fixed composition $Q_X = \hat{P}_{X_1^n}$. Thus, assuming that the constraint $vQ_X = q_Y$ is satisfied, maximizing $P^n_{X,Y}(V = v \mid Q_Y = q_Y)$ is equivalent to maximizing $P^n_{X,Y}(V = v)$, because $P^n_{X,Y}(Q_Y = q_Y)$ is constant since $Q_Y = q_Y$ is given.

To compute $P^n_{X,Y}(V = v)$, we observe that for every $x \in \mathcal{X}$ and every empirical pmf v_x:

$$P^n_{X,Y}\left(V_x = v_x\right) \doteq e^{-nQ_X(x)D(v_x||W_x)} \tag{4.47}$$

in analogy with equation 4.12. This is because the channel is memoryless:

$$\forall x_1^n \in \mathcal{X}^n, \forall y_1^n \in \mathcal{Y}^n, \;\; P_{Y_1^n|X_1^n}\left(y_1^n \mid x_1^n\right) = \prod_{i=1}^{n} W_{x_i}\left(y_i\right)$$

which means that if we are given the X_i values, the Y_i values are drawn independently. Once again, note that we do not need to condition on Q_X when computing probabilities of V (or V_x) because Q_X is a known fixed composition. Furthermore, for any $v = \left[v_1 \cdots v_{|\mathcal{X}|}\right]$, we have:

$$P_{X,Y}^n\left(V = v\right) = \prod_{x \in \mathcal{X}} P_{X,Y}^n\left(V_x = v_x\right) \tag{4.48}$$

which holds because empirical conditional distributions for different values of $x \in \mathcal{X}$ are independent for a fixed composition Q_X. Indeed, knowing Q_X and n implies we know $nQ_X(x)$ for all $x \in \mathcal{X}$. This is the only kind of information that can be inferred about one empirical conditional distribution from another. The reason for this is no longer because samples are i.i.d. as in equation 4.13 in section 4.2. In fact, the samples are clearly not i.i.d. However, memorylessness of the channel implies that Y_i is conditionally independent of all $Y_j, \;\; j \neq i$ and $X_j, \;\; j \neq i$, given X_i. To find V_x, we must collect all $X_i = x$ samples, and no other empirical conditional distribution requires these samples. This means that the corresponding Y_i values of the $X_i = x$ samples (which determine V_x) are independent of all samples with $X_i \neq x$. Hence, the empirical conditional distributions are independent of one another. Combining this conclusion in equation 4.48 with equation 4.47, we have:

$$P_{X,Y}^n\left(V = v\right) \doteq \exp\left(-n \sum_{x \in \mathcal{X}} Q_X(x) D(w_x||W_x)\right)$$

which we may write using Definition 4.2.2 as:

$$P_{X,Y}^n\left(V = v\right) \doteq e^{-nD(v||W|Q_X)} \tag{4.49}$$

in analogy with equation 4.15.

Using equation 4.49, we see that the most probable V^* is given by minimizing the exponent $D(v||W|Q_X)$ over all empirical channel conditional pmfs v, under the exponential approximation. Hence, the optimization problem which finds V^* is:

$$\min_{v: \;\; vQ_X = q_Y} D(v||W|Q_X) \tag{4.50}$$

where the empirical pmf of Y, $Q_Y = q_y$, is given, and we use a fixed composition Q_X. This is exactly the same problem as that in statement 4.43. Therefore, both problems find the most probable channel perturbations $J_W^* = V^* - W$ with fixed source composition Q_X given that Q_Y is observed. The ensuing subsections solve problem 4.50 in the global and local scenarios.

Before delving into the solutions, we draw attention to some miscellaneous observations. Firstly, it is worth mentioning that the optimal value of problem 4.50 will be greater than or equal to the optimal value of problem 4.19. Indeed, if we let $Q_X = P_X$, then the convexity of KL divergence in its arguments and Jensen's inequality imply that:

$$D(v||W|Q_X) \geq D(q_Y||P_Y)$$

for every set of empirical channel conditional pmfs v such that $vQ_X = q_Y$, where $D(q_Y||P_Y) = D(\hat{P}_{Y_1^n}||P_Y)$ is the optimal value of problem 4.19 according to equation 4.17. This trivially gives us:

$$\min_{v:\ vQ_X=q_Y} D(v||W|Q_X) \geq \min_{\substack{p_X,v:\\ vp_X=q_Y}} D(p_X||Q_X) + D(v||W|p_X) = D(q_Y||P_Y). \tag{4.51}$$

On another note, we observe that Problem 4.50 actually finds the maximum a posteriori probablity (MAP) estimate V^* after observing Q_Y, because it maximizes $P^n_{X,Y}(V = v \mid Q_Y = q_y)$ under the exponential approximation. It turns out that we can also perceive V^* as the set of dominating empirical channel conditional distributions which lead to q_Y. To elucidate this, we first find $P^n_{X,Y}(Q_Y = q_Y)$. Unfortunately, we cannot blindly use equation 4.11:

$$P^n_{X,Y}\left(Q_Y = q_Y\right) \doteq e^{-nD(q_Y||P_Y)}$$

because the Y_1^n are no longer i.i.d. Hence, we have from first principles:

$$P^n_{X,Y}\left(Q_Y = q_Y\right) = \sum_{v:\ vQ_X=q_Y} P^n_{X,Y}\left(V = v\right)$$

into which we can substitute equation 4.49 (non-rigorously) to get:

$$P^n_{X,Y}\left(Q_Y = q_Y\right) = \sum_{v:\ vQ_X=q_Y} e^{-nD(v||W|Q_X)}$$

where the sum indexes over a polynomial in n number of terms, because the number of possible empirical channel conditional pmfs $v = \left[v_1 \cdots v_{|\mathcal{X}|}\right]$ which satisfy $vQ_X = q_Y$ is polynomial in n for fixed n. Then, the Laplace principle [26] gives:

$$P^n_{X,Y}\left(Q_Y = q_Y\right) \doteq \exp\left(-n \min_{v:\ vQ_X=q_Y} D(v||W|Q_X)\right). \tag{4.52}$$

The exponent in equation 4.52 is precisely the extremal problem in statement 4.50. Thus, for a fixed composition Q_X, the exponential approximation of the probability of observing $Q_Y = q_Y$ equals the exponential approximation of the probability of the dominating (most probable) V^*. This also agrees with equation 4.16, where letting $\hat{P}_{X_1^n} = p_X = P_X$ (which corresponds to fixing the source composition or allowing no source perturbation) gives the same exponent as equation 4.52, because $D(p_X||P_X) = 0$ when $p_X = P_X$. In the next subsection, we find an implicit global solution for problem 4.50.

4.3.1 Implicit Global Solution

As in subsection 4.2.1, for the sake of clarity, we will alter notation slightly before solving problem 4.50. In our ensuing derivations, we will use the notation of problem 4.50 rather than its equivalent formulation in statement 4.43. Furthermore, we will use V and Q_Y, instead of v and q_Y, to mean particular values of the empirical distributions rather than random variables which represent these empirical distributions. This should not generate any confusion since we will not compute any large deviation probabilities in this subsection. Using this altered notation, we can rewrite the optimization problem in statement 4.50 as:

$$\min_{V:\ VQ_X=Q_Y} D(V||W|Q_X) \tag{4.53}$$

where the source composition Q_X is fixed, the empirical pmf Q_Y is observed, and the minimization is over all empirical channels (conditional pmfs) V. The optimizing V^* is the most probable empirical channel with fixed source composition Q_X. We will first attack problem 4.53 using the method of Lagrange multipliers. The next theorem presents this Lagrangian optimization approach.

Theorem 4.3.1 (Most Probable Empirical Channel PMFs)**.** *The most probable empirical channel conditional pmfs, V^*, which are global optimizing arguments of the extremal problem in statement 4.53, are given by:*

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x^*(y) = \frac{W_x(y)e^{\lambda_y}}{\sum\limits_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}}$$

where the $\lambda_y,\ \ y \in \mathcal{Y}$ satisfy the marginal constraint:

$$\forall y \in \mathcal{Y},\ \ \sum_{x\in\mathcal{X}} Q_X(x)\left(\frac{W_x(y)e^{\lambda_y}}{\sum\limits_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}}\right) = Q_Y(y).$$

Proof.
We seek to minimize $D(V||W|Q_X)$ over all empirical channel conditional pmfs V subject to the constraint:

$$\forall y \in \mathcal{Y},\ \ \sum_{x\in\mathcal{X}} Q_X(x)V_x(y) = Q_Y(y)$$

by employing the method of Lagrange multipliers. We additionally impose the constraint:

$$\forall x \in \mathcal{X},\ \ \sum_{y\in\mathcal{Y}} V_x(y) = 1$$

to ensure that each $V_x,\ \ x \in \mathcal{X}$ is normalized like a valid pmf. However, we do not impose the non-negativity constraints on $V_x,\ \ x \in \mathcal{X}$ as they will turn out to hold naturally. The Lagrangian, $\mathcal{L}(V, \lambda, \mu)$, for the problem is:

$$\mathcal{L} = \sum_{x\in\mathcal{X}} Q_X(x) \sum_{y\in\mathcal{Y}} V_x(y) \log\left(\frac{V_x(y)}{W_x(y)}\right) + \sum_{y\in\mathcal{Y}} \lambda_y \sum_{x\in\mathcal{X}} Q_X(x)V_x(y) + \sum_{x\in\mathcal{X}} \mu_x \sum_{y\in\mathcal{Y}} V_x(y)$$

where $\lambda = \begin{bmatrix}\lambda_1 \cdots \lambda_{|\mathcal{Y}|}\end{bmatrix}^T$ and $\mu = \begin{bmatrix}\mu_1 \cdots \mu_{|\mathcal{X}|}\end{bmatrix}^T$ are Lagrange multipliers. Taking the partial derivatives of $\mathcal{L}$ with respect to $V_x(y)$ and setting them equal to 0, we get:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ \frac{\partial \mathcal{L}}{\partial V_x(y)} = Q_X(x)\left(1 + \lambda_y + \log\left(\frac{V_x(y)}{W_x(y)}\right)\right) + \mu_x = 0.$$

By appropriately absorbing constants into the Lagrange multipliers and rearranging, we have:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x(y) = W_x(y)e^{\lambda_y - \mu_x}.$$

This clearly shows that each empirical conditional distribution, V_x, is exponentially tilted from the theoretical conditional distribution W_x. The $\lambda_y,\ \ y \in \mathcal{Y}$ are natural parameters corresponding to indicator sufficient statistics $\mathbb{I}(Y = y)$, and the $\mu_x,\ \ x \in \mathcal{X}$ are the log-partition functions. We will

provide an alternative proof of Theorem 4.3.1 using i-projections which will make this structure clear. To find $\mu_x,\ x \in \mathcal{X}$, we impose the valid pmf normalization constraints:

$$\forall x \in \mathcal{X},\ \sum_{y\in\mathcal{Y}} V_x(y) = e^{-\mu_x} \sum_{y\in\mathcal{Y}} W_x(y)e^{\lambda_y} = 1$$

which give:

$$\forall x \in \mathcal{X},\ e^{\mu_x} = \sum_{y\in\mathcal{Y}} W_x(y)e^{\lambda_y}.$$

This agrees with the observation that $e^{\mu_x},\ x \in \mathcal{X}$ are partition functions. Using the expressions for e^{μ_x}, the optimal V_x^* have the form:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = \frac{W_x(y)e^{\lambda_y}}{\sum_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}}$$

where the $\lambda_y,\ y \in \mathcal{Y}$ must satisfy the the marginal constraint:

$$\forall y \in \mathcal{Y},\ \sum_{x\in\mathcal{X}} Q_X(x)V_x^*(y) = \sum_{x\in\mathcal{X}} Q_X(x) \left(\frac{W_x(y)e^{\lambda_y}}{\sum_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}} \right) = Q_Y(y).$$

Hence, we have found the optimizing V^*. Note that the non-negativity constraints on $V_x^*,\ x \in \mathcal{X}$ which ensure they are valid pmfs automatically hold. We now justify that the V^* is indeed a minimizer of problem 4.53 (as we have only shown it is a stationary point of the Lagrangian so far). Observe that our objective function $D(V||W|Q_X)$ is a strictly convex function of V for fixed $W$and Q_X, because $\forall x \in \mathcal{X},\ Q_X(x) > 0$, and for each $x \in \mathcal{X}$, $D(V_x||W_x)$ is a strictly convex function of V_x for fixed W_x. Moreover, the constraints $VQ_X = Q_Y$ and $1^T V = 1^T$ (where 1 is the column vector with all entries equal to 1) are affine equality constraints. Theorem 2.2.8 in [27] asserts that such convex minimization problems (with affine equality constraints) attain their global minimum at the stationary point of their Lagrangian. This completes the proof. □

Theorem 4.3.1 does not provide an explicit characterization of V^*. It only presents an implicit global solution of problem 4.53 by presenting each $V_x^*,\ x \in \mathcal{X}$ as functions of $\lambda_y,\ y \in \mathcal{Y}$, where the λ_y are implicitly defined by a marginal constraint. In fact, our proof using Lagrange multipliers does not elaborate upon why such $\lambda_y,\ y \in \mathcal{Y}$ must exist. We now present an alternative proof of Theorem 4.3.1 by setting problem 4.53 up as an i-projection. This proof offers more insight into the existence, uniqueness, and exponential family form of V^*.

Proof.
We first set up the extremal problem in statement 4.53 as an i-projection. Observe that:

$$D(V||W|Q_X) = D(Q_{X,Y}||P_{X,Y})$$

where $Q_{X,Y}$ denotes the empirical joint pmf, with empirical conditional distributions of Y given X given by V, and empirical marginal pmf of X given by Q_X (which is both the empirical and the

theoretical distribution of X in the fixed composition scenario). Note that $P_{X,Y}$ is the theoretical joint pmf, with empirical conditional distributions of Y given X given by W, and empirical marginal pmf of X given by Q_X. The constraint $VQ_X = Q_Y$ is a marginal constraint on $Q_{X,Y}$. So, consider the linear family of distributions:

$$\mathcal{L} \triangleq \left\{Q_{X,Y} : \forall x \in \mathcal{X},\ \mathbb{E}_{Q_{X,Y}}\left[\mathbb{I}\left(X = x\right)\right] = Q_X(x)\ \ \wedge\ \ \forall y \in \mathcal{Y},\ \mathbb{E}_{Q_{X,Y}}\left[\mathbb{I}\left(Y = y\right)\right] = Q_Y(y)\right\}$$

which is the set of joint pmfs on $\mathcal{X} \times \mathcal{Y}$ with marginal pmfs Q_X and Q_Y. With this linear family $\mathcal{L}$, problem 4.53 can be recast as an i-projection:

$$Q^*_{X,Y} = \underset{Q_{X,Y} \in \mathcal{L}}{\arg\min}\, D(Q_{X,Y}||P_{X,Y})$$

where we additionally know that the marginal pmf of X of $P_{X,Y}$ is also Q_X.

We now solve this i-projection. Let $\mathcal{E}$ be the $(|\mathcal{X}| + |\mathcal{Y}|)$-dimensional canonical exponential family which is "orthogonal" to the linear family $\mathcal{L}$:

$$\mathcal{E} \triangleq \left\{ Q_{X,Y} : \forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ Q_{X,Y}(x,y) = \frac{P_{X,Y}(x,y)e^{\tau_x + \lambda_y}}{\sum\limits_{a \in \mathcal{X}} \sum\limits_{b \in \mathcal{Y}} P_{X,Y}(a,b)e^{\tau_a + \lambda_b}} \text{ for some } \tau_x, \lambda_y \in \mathbb{R} \right\}$$

where $Q_{X,Y}$ denotes any joint pmf on $\mathcal{X} \times \mathcal{Y}$, $P_{X,Y}$ is the base distribution, $\tau_x, x \in \mathcal{X}$ and $\lambda_y, y \in \mathcal{Y}$ are the natural parameters, $\mathbb{I}(X = x), x \in \mathcal{X}$ and $\mathbb{I}(Y = y), y \in \mathcal{Y}$ are the sufficient statistics, and the normalization sum in the denominator is the partition function. We remark that Definition 3.4.3 of canonical exponential families was only for pdfs and had a single parameter. A more general definition includes $\mathcal{E}$ as a canonical exponential family as well. It is well-known that the optimizing $Q^*_{X,Y}$ of the i-projection exists, and is the unique distribution that is in both $\mathcal{E}$ and $\mathcal{L}$ [10]. Hence, $Q^*_{X,Y} \in \mathcal{L} \cap \mathcal{E}$. Since $Q^*_{X,Y} \in \mathcal{E}$, we have:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ Q^*_{X,Y}(x,y) = \frac{P_{X,Y}(x,y)e^{\tau_x + \lambda_y}}{\sum\limits_{a \in \mathcal{X}} \sum\limits_{b \in \mathcal{Y}} P_{X,Y}(a,b)e^{\tau_a + \lambda_b}}.$$

Furthermore, since $Q^*_{X,Y} \in \mathcal{L}$, we have:

$$\forall x \in \mathcal{X},\ \sum_{y \in \mathcal{Y}} Q^*_{X,Y}(x,y) = Q_X(x),$$

$$\forall y \in \mathcal{Y},\ \sum_{x \in \mathcal{X}} Q^*_{X,Y}(x,y) = Q_Y(y).$$

Imposing the marginal constraint on X gives:

$$\forall x \in \mathcal{X},\ \frac{e^{\tau_x} \sum\limits_{y \in \mathcal{Y}} P_{X,Y}(x,y)e^{\lambda_y}}{\sum\limits_{a \in \mathcal{X}} \sum\limits_{b \in \mathcal{Y}} P_{X,Y}(a,b)e^{\tau_a + \lambda_b}} = Q_X(x)$$

where we use the fact that $P_{X,Y}$ has marginal pmf Q_X to get:

$$\forall x \in \mathcal{X},\ \ e^{\tau_x} \sum_{y \in \mathcal{Y}} W_x(y) e^{\lambda_y} = \sum_{a \in \mathcal{X}} \sum_{b \in \mathcal{Y}} P_{X,Y}(a,b) e^{\tau_a + \lambda_b}.$$

Hence, the solution to the i-projection is:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ Q^*_{X,Y}(x,y) = \frac{P_{X,Y}(x,y) e^{\lambda_y}}{\sum_{z \in \mathcal{Y}} W_x(z) e^{\lambda_z}}$$

where the $\lambda_y,\ \ y \in \mathcal{Y}$ satisfy:

$$\forall y \in \mathcal{Y},\ \ \sum_{x \in \mathcal{X}} Q^*_{X,Y}(x,y) = \sum_{x \in \mathcal{X}} Q_X(x) \left(\frac{W_x(y) e^{\lambda_y}}{\sum_{z \in \mathcal{Y}} W_x(z) e^{\lambda_z}} \right) = Q_Y(y)$$

which imposes the marginal constraint on Y. Since $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ Q^*_{X,Y}(x,y) = V^*_x(y) Q_X(x)$, the optimal empirical conditional distributions are:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V^*_x(y) = \frac{W_x(y) e^{\lambda_y}}{\sum_{z \in \mathcal{Y}} W_x(z) e^{\lambda_z}}$$

where the $\lambda_y,\ \ y \in \mathcal{Y}$ satisfy:

$$\forall y \in \mathcal{Y},\ \ \sum_{x \in \mathcal{X}} Q_X(x) \left(\frac{W_x(y) e^{\lambda_y}}{\sum_{z \in \mathcal{Y}} W_x(z) e^{\lambda_z}} \right) = Q_Y(y).$$

This completes the proof. □

From this alternative proof of Theorem 4.3.1 using i-projections, we glean a better understanding of the form of V^*. The most probable empirical channel conditional pmfs, $V^*_x,\ \ x \in \mathcal{X}$, are exponentially tilted versions of the theoretical conditional pmfs, $W_x,\ \ x \in \mathcal{X}$, because we project the theoretical conditional pmfs along exponential families onto the linear family which represents the empirical observations. The geometric intuition behind Theorem 4.3.1 is further elaborated in subsection 4.3.2. It is rather dissatisfying that we cannot find $V^*_x,\ \ x \in \mathcal{X}$ explicitly by solving the conditions for $\lambda_y,\ \ y \in \mathcal{Y}$. Thus, the next two corollaries illustrate scenarios where the $\lambda_y,\ \ y \in \mathcal{Y}$, can be found explicitly.

Corollary 4.3.2 (Identical Conditional Distributions)**.** *If all the theoretical conditional pmfs of Y given X are identical:*

$$\forall x \in \mathcal{X},\ \ W_x = P_Y,$$

then the most probable empirical conditional pmfs of Y given X are also all identical:

$$\forall x \in \mathcal{X},\ \ V^*_x = Q_Y.$$

Proof.
From Theorem 4.3.1 and the assumption $\forall x \in \mathcal{X},\ W_x = P_Y$, we have:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = \frac{P_Y(y)e^{\lambda_y}}{\sum_{z\in\mathcal{Y}} P_Y(z)e^{\lambda_z}}$$

where the $\lambda_y,\ y \in \mathcal{Y}$ satisfy the marginal constraint:

$$\forall y \in \mathcal{Y},\ \sum_{x\in\mathcal{X}} Q_X(x) \left(\frac{P_Y(y)e^{\lambda_y}}{\sum_{z\in\mathcal{Y}} P_Y(z)e^{\lambda_z}} \right) = \frac{P_Y(y)e^{\lambda_y}}{\sum_{z\in\mathcal{Y}} P_Y(z)e^{\lambda_z}} = Q_Y(y).$$

This marginal constraint can be rearranged to:

$$\forall y \in \mathcal{Y},\ \sum_{z\in\mathcal{Y}} P_Y(z)e^{\lambda_z} = \frac{P_Y(y)e^{\lambda_y}}{Q_Y(y)}$$

which gives:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = Q_Y(y).$$

□

In Corollary 4.3.2, we essentially determine the $\lambda_y,\ y \in \mathcal{Y}$, by how much we exponentially tilt P_Y to get Q_Y. Then, as all the theoretical conditional distributions are identical to P_Y, we tilt them all by the same amount.

Corollary 4.3.3 (Exponentially Tilted Empirical Output PMF)**.** *Suppose we are given that Q_Y is exponentially tilted from P_Y:*

$$\forall y \in \mathcal{Y},\ Q_Y(y) = \frac{P_Y(y)\,(1 + l_y)}{1 + \sum_{z\in\mathcal{Y}} P_Y(z) l_z}.$$

If $l = \begin{bmatrix} l_1 \cdots l_{|\mathcal{Y}|} \end{bmatrix}^T = v + c1$ for any constant $c \in \mathbb{R}$ and any vector $v \in$ nullspace$\left(W^T\right)$, where 1 is the column vector with all entries equal to unity, then the most probable empirical channel conditional pmfs are:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = \frac{W_x(y)\,(1 + l_y)}{1 + \sum_{z\in\mathcal{Y}} W_x(z) l_z}.$$

Proof.
Letting $e^{\lambda_y} = 1 + \mu_y$ for each $y \in \mathcal{Y}$ in Theorem 4.3.1, we have:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = \frac{W_x(y)\,(1 + \mu_y)}{1 + \sum_{z\in\mathcal{Y}} W_x(z) \mu_z}$$

where the $\mu_y,\ y \in \mathcal{Y}$ satisfy the marginal constraint:

$$\forall y \in \mathcal{Y},\ \sum_{x\in\mathcal{X}} Q_X(x) \left(\frac{W_x(y)\,(1+\mu_y)}{1+\sum_{z\in\mathcal{Y}} W_x(z)\mu_z} \right) = Q_Y(y).$$

Since Q_Y is exponentially tilted from P_Y, we can write this marginal constraint as:

$$\begin{aligned} \forall y \in \mathcal{Y},\ \sum_{x\in\mathcal{X}} Q_X(x) \left(\frac{W_x(y)\,(1+\mu_y)}{1+\sum_{z\in\mathcal{Y}} W_x(z)\mu_z} \right) &= \frac{P_Y(y)\,(1+l_y)}{1+\sum_{z\in\mathcal{Y}} P_Y(z) l_z} \\ &= \sum_{x\in\mathcal{X}} Q_X(x) \left(\frac{W_x(y)\,(1+l_y)}{1+\sum_{a\in\mathcal{X}} Q_X(a) \sum_{b\in\mathcal{Y}} W_a(b) l_b} \right). \end{aligned}$$

We can clearly see from this relation that if we have:

$$\forall x \in \mathcal{X},\ \sum_{b\in\mathcal{Y}} W_x(b) l_b = c$$

for some constant $c \in \mathbb{R}$, then:

$$\sum_{a\in\mathcal{X}} Q_X(a) \sum_{b\in\mathcal{Y}} W_a(b) l_b = c$$

and so, $\forall y \in \mathcal{Y},\ \mu_y = l_y$ satisfies the marginal constraint:

$$\forall y \in \mathcal{Y},\ Q_Y(y) = \frac{P_Y(y) e^{l_y}}{\sum_{z\in\mathcal{Y}} P_Y(z) e^{l_z}}.$$

Let $\mu = \left[\mu_1 \cdots \mu_{|\mathcal{Y}|}\right]^T$ denote the vector of $\mu_y,\ y \in \mathcal{Y}$. We have shown that if $W^T l = c1$ for some constant $c \in \mathbb{R}$, then $\mu = l$ satisfies the marginal constraint. So, if $l = v + c1$ for any constant $c \in \mathbb{R}$ and any vector $v \in \text{nullspace}\left(W^T\right)$, then $W^T l = W^T v + cW^T 1 = c1$, which means $\mu = l$ satisfies the marginal constraint and:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ V_x^*(y) = \frac{W_x(y)\,(1+l_y)}{1+\sum_{z\in\mathcal{Y}} W_x(z) l_z}.$$

□

In Corollary 4.3.3, we assume that Q_Y is exponentially tilted from P_Y and write:

$$\forall y \in \mathcal{Y},\ Q_Y(y) = \frac{P_Y(y)\,(1+l_y)}{1+\sum_{z\in\mathcal{Y}} P_Y(z) l_z}.$$

While this equation does not have the form of an exponential tilt, letting $1+l_y = e^{\tau_y}$ for each $y \in \mathcal{Y}$ makes it clear that Q_Y is indeed exponentially tilted from P_Y. We must be careful and recognize that $\forall y \in \mathcal{Y},\ l_y > -1$. This ensures the $\tau_y,\ y \in \mathcal{Y}$ are well-defined and that $\forall y \in \mathcal{Y},\ Q_Y(y) > 0$. Corollary 4.3.3 illustrates that we can often explicitly find V^* when the dimension of the left nullspace of W is large. For example, if W is a tall matrix and $|\mathcal{Y}|$ is much larger than $|\mathcal{X}|$, Q_Y is more likely to satisfy the premise of Corollary 4.3.3.

Since we cannot generally find explicit global solutions to problem 4.53, we will also solve it using local approximations of KL divergence in subsection 4.3.3. For comparison with those forthcoming results, we compute local approximations of our global solutions directly. To this end, we reproduce the solutions in Theorem 4.3.1 with altered form:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x^*(y) = \frac{W_x(y)\left(1+\epsilon\mu_y\right)}{1+\epsilon\sum\limits_{z\in\mathcal{Y}} W_x(z)\mu_z} \tag{4.54}$$

where the $\mu_y,\ \ y \in \mathcal{Y}$, satisfy the marginal constraint:

$$\forall y \in \mathcal{Y},\ \sum_{x\in\mathcal{X}} Q_X(x)\left(\frac{W_x(y)\left(1+\epsilon\mu_y\right)}{1+\epsilon\sum\limits_{z\in\mathcal{Y}} W_x(z)\mu_z}\right) = Q_Y(y). \tag{4.55}$$

Note that we write $\forall y \in \mathcal{Y},\ e^{\lambda_y} = 1+\epsilon\mu_y$ for some small $\epsilon > 0$. The $\mu_y > -1,\ \ y \in \mathcal{Y}$, express the exponential tilting as multiplicative perturbations (without loss of generality). The ϵ factor enforces the assumption that the perturbed V^* is close to W. To complete the local approximation, we use the following substitution:

$$\frac{1}{1+\epsilon\sum\limits_{z\in\mathcal{Y}} W_x(z)\mu_z} = 1-\epsilon\sum_{z\in\mathcal{Y}} W_x(z)\mu_z + o\left(\epsilon\right)$$

where $o\left(\epsilon\right)$ denotes a function which satisfies $\lim\limits_{\epsilon\to 0^+}\dfrac{o\left(\epsilon\right)}{\epsilon} = 0$. This gives the locally approximated global solution:

$$\begin{aligned}\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x^*(y) &= W_x(y)\left(1+\epsilon\mu_y\right)\left(1-\epsilon\sum_{z\in\mathcal{Y}} W_x(z)\mu_z + o\left(\epsilon\right)\right)\\ &= W_x(y) + \epsilon W_x(y)\left(\mu_y - \sum_{z\in\mathcal{Y}} W_x(z)\mu_z\right) + o\left(\epsilon\right)\end{aligned} \tag{4.56}$$

where the $\mu_y,\ \ y \in \mathcal{Y}$, satisfy:

$$\begin{aligned}\forall y \in \mathcal{Y},\ \ \epsilon J_Y(y) &= \sum_{x\in\mathcal{X}} Q_X(x)W_x(y)\left(1+\epsilon\mu_y - \epsilon\sum_{z\in\mathcal{Y}} W_x(z)\mu_z\right) - P_Y(y) + o\left(\epsilon\right)\\ &= \epsilon\sum_{x\in\mathcal{X}} Q_X(x)W_x(y)\left(\mu_y - \sum_{z\in\mathcal{Y}} W_x(z)\mu_z\right) + o\left(\epsilon\right)\end{aligned} \tag{4.57}$$

because $Q_Y = P_Y + \epsilon J_Y$. Neglecting the ϵ factors and terms, equations 4.56 and 4.57 illustrate that the most probable channel perturbations, $J_x^* = V_x^* - W_x,\ \ x \in \mathcal{X}$ (when the global solutions are locally approximated) are:

$$\forall x \in \mathcal{X},\ \ J_x^* = [W_x]\left(I - 1W_x^T\right)\mu \tag{4.58}$$

where $\mu = \left[\mu_1 \cdots \mu_{|\mathcal{Y}|}\right]^T$ satisfies:

$$\left([P_Y] - W\,[Q_X]\,W^T\right)\mu = J_Y \tag{4.59}$$

where I is the identity matrix, and 1 denotes a column vector with all entries equal to 1. We will see that solving the locally approximated problem 4.53 will produce these same results.

4.3.2 Geometric Interpretation of Global Solution

Before expounding the solutions to problem 4.53 under local approximations, we take a moment to appreciate the geometry of its global solutions. From Theorem 4.3.1, we have:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ V_x^*(y) = \frac{W_x(y)e^{\lambda_y}}{\sum\limits_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}}$$

where the $\lambda_y,\ \ y \in \mathcal{Y}$ satisfy the marginal constraint:

$$\forall y \in \mathcal{Y},\ \ \sum_{x\in\mathcal{X}} Q_X(x)\left(\frac{W_x(y)e^{\lambda_y}}{\sum\limits_{z\in\mathcal{Y}} W_x(z)e^{\lambda_z}}\right) = Q_Y(y).$$

Figure 4.4 interprets these solutions by considering two spaces of pmfs on $\mathcal{Y}$. For simplicity, we assume that $\mathcal{X} = \{1, 2, 3\}$. On the theoretical space, we have the conditional pmfs W_1, W_2, and W_3, which average with respect to Q_X to produce the marginal pmf P_Y. On the empirical space, we have the empirical conditional pmfs V_1, V_2, and V_3, which average with respect to Q_X to produce some empirical marginal pmf. The most probable empirical channel conditional pmfs, $V_x^*,\ \ x \in \mathcal{X}$, are obtained by exponentially tilting the theoretical conditional pmfs, $W_x,\ \ x \in \mathcal{X}$, such that $V_x^*,\ \ x \in \mathcal{X}$ average to the observed empirical marginal pmf Q_Y. Hence, to get from the theoretical space to the empirical space, we construct parallel canonical exponential families starting at W_1, W_2, and W_3 (base distributions), and travel along these exponential families by the same amount $\lambda_y,\ \ y \in \mathcal{Y}$ (natural parameters), until the average of the $V_x,\ \ x \in \mathcal{X}$ with respect to Q_X becomes Q_Y. Therefore, we are essentially taking the entire theoretical space at P_Y and shifting it along a canonical exponential family to align it with Q_Y.

4.3.3 Local Solution using Lagrangian Optimization

To locally approximate problem 4.53, we once again assume that empirical distributions perturb very slightly from theoretical distributions, as in subsection 4.2.2. This means we have:

$$\begin{aligned} Q_Y &= P_Y + J_Y \\ \forall x \in \mathcal{X},\ \ V_x &= W_x + J_x \end{aligned}$$

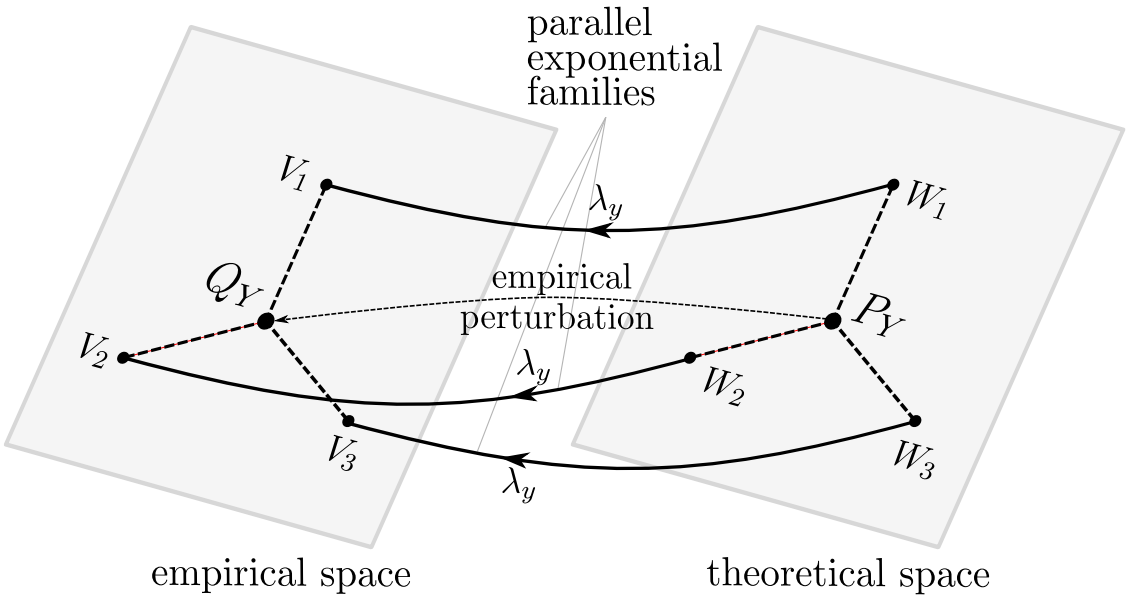

Figure 4.4: Geometric view of most probable empirical channel conditional pmfs.

where the additive perturbation vectors have elements that are small (or the ϵ factor which is understood to be in front of them is small). We also approximate KL divergences with χ^2-divergences:

$$\forall x \in \mathcal{X}, \;\; D(V_x||W_x) \overset{\text{local}}{=} \frac{1}{2} \|J_x\|^2_{W_x} \tag{4.60}$$

$$D(V||W|Q_X) \overset{\text{local}}{=} \frac{1}{2} \sum_{x \in \mathcal{X}} Q_X(x) \|J_x\|^2_{W_x} \tag{4.61}$$

where equations 4.60 and 4.61 are restatements of equations 4.30 and 4.31, respectively. Equation 4.61 can be used to recast the extremal problem in statement 4.53 in local form:

$$\begin{aligned} \min_{J_W = [J_1 \cdots J_{|\mathcal{X}|}]} \quad & \frac{1}{2} \sum_{x \in \mathcal{X}} Q_X(x) \|J_x\|^2_{W_x} \\ \text{subject to:} \quad & J_W^T 1 = 0, \\ \text{and} \quad & J_W Q_X = J_Y \end{aligned} \tag{4.62}$$

where 1 denotes the column vector of with all entries equal to 1, 0 denotes the column vector with all entries equal to 0, and the second constraint imposes $VQ_X = Q_Y$ as required in problem 4.53. The next theorem solves the optimization problem in statement 4.62.

Theorem 4.3.4 (Most Probable Local Channel Perturbations)**.** *The most probable local channel perturbations, which are optimizing arguments of the extremal problem in statement 4.62, are given by:*

$$\forall x \in \mathcal{X}, \;\; J_x^* = \left([W_x] - W_x W_x^T\right) \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger \left[\sqrt{P_Y}\right]^{-1} J_Y$$

where $B = \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{Q_X}\right]$ *is the divergence transition matrix, and* $\dagger$ *denotes the Moore-*

Penrose pseudoinverse. The optimal value of the objective function of problem 4.62 is:

$$\frac{1}{2}\sum_{x\in\mathcal{X}} Q_X(x)\left\|J_x^*\right\|_{W_x}^2 = \frac{1}{2}J_Y^T\left[\sqrt{P_Y}\right]^{-1}\left(I-BB^T\right)^\dagger\left[\sqrt{P_Y}\right]^{-1}J_Y.$$

Proof.
We first set up the Lagrangian, $\mathcal{L}\left(J_W,\lambda,\mu\right)$, of problem 4.62:

$$\mathcal{L} = \sum_{x\in\mathcal{X}} Q_X(x)\sum_{y\in\mathcal{Y}}\frac{J_x^2(y)}{W_x(y)} + \sum_{y\in\mathcal{Y}}\lambda_y\sum_{x\in\mathcal{X}}Q_X(x)J_x(y) + \sum_{x\in\mathcal{X}}\mu_x\sum_{y\in\mathcal{Y}}J_x(y)$$

where $\lambda = \left[\lambda_1\cdots\lambda_{|\mathcal{Y}|}\right]^T$ and $\mu = \left[\mu_1\cdots\mu_{|\mathcal{X}|}\right]^T$ are Lagrange multipliers. Note that we neglect the factor of $\frac{1}{2}$ in the objective function. Taking partial derivatives of $\mathcal{L}$ with respect to $J_x(y)$ and setting them equal to 0, we get:

$$\forall x\in\mathcal{X},\forall y\in\mathcal{Y},\ \ \frac{\partial\mathcal{L}}{\partial J_x(y)} = Q_X(x)\left(\frac{2J_x(y)}{W_x(y)}+\lambda_y\right)+\mu_x = 0$$

where we may appropriately absorb constants into the Lagrange multipliers to get:

$$\forall x\in\mathcal{X},\forall y\in\mathcal{Y},\ \ \frac{J_x(y)}{W_x(y)} = \lambda_y-\mu_x.$$

Solving these equations will produce the stationary points of the Lagrangian. Observe that our objective function is a strictly convex function of J_W, because $\forall x\in\mathcal{X},\ Q_X(x)>0$ and $\forall x\in\mathcal{X},\forall y\in\mathcal{Y},\ W_x(y)>0$. Furthermore, the constraints $J_W^T1=0$ and $J_WQ_X=J_Y$ are affine equality constraints. As in the first proof of Theorem 4.3.1, we may appeal to Theorem 2.2.8 in [27] which asserts that convex minimization problems with affine equality constraints attain their global minimum at the stationary point of their Lagrangian. Hence, we solve the equations derived from the first derivatives of the Lagrangian to find the global minimizing arguments of problem 4.62.

Recall from section 1.2 in chapter 1 that we can define log-likelihood perturbations:

$$\begin{aligned}\forall x\in\mathcal{X},\ \ L_x &= \left[W_x\right]^{-1}J_x\\ L_Y &= \left[P_Y\right]^{-1}J_Y\end{aligned}$$

using which we can rewrite the Lagrangian stationary conditions as:

$$\forall x\in\mathcal{X},\ \ L_x = \lambda-\mu_x1$$

where $\lambda = \left[\lambda_1\cdots\lambda_{|\mathcal{Y}|}\right]^T$, and 1 denotes the column vector with all entries equal to 1. Imposing the valid perturbation constraints, we have:

$$\forall x\in\mathcal{X},\ \ W_x^TL_x = W_x^T\lambda-\mu_xW_x^T1=0$$

$$\forall x\in\mathcal{X},\ \ \mu_x = W_x^T\lambda$$

which gives:

$$\forall x\in\mathcal{X},\ \ L_x = \lambda-\left(W_x^T\lambda\right)1 = \left(I-1W_x^T\right)\lambda \tag{4.63}$$

where I is the identity matrix. Imposing the marginal constraint, we have:

$$\sum_{x\in\mathcal{X}} Q_X(x)J_x = \sum_{x\in\mathcal{X}} Q_X(x)\left[W_x\right]L_x = \sum_{x\in\mathcal{X}} Q_X(x)\left[W_x\right]\left(I - 1W_x^T\right)\lambda = J_Y$$

which simplifies to:

$$\left(\left[P_Y\right] - W\left[Q_X\right]W^T\right)\lambda = J_Y. \tag{4.64}$$

Now recall from section 1.2 in chapter 1 that we can define normalized perturbations equivalent to the additive and log-likelihood perturbations.

$$\begin{aligned}\forall x\in\mathcal{X},\ K_x &= \left[\sqrt{W_x}\right]^{-1}J_x\\ K_Y &= \left[\sqrt{P_Y}\right]^{-1}J_Y\end{aligned}$$

We now transfer the marginal constraint into the space of normalized perturbations for convenience.

$$\begin{aligned}K_Y &= \left[\sqrt{P_Y}\right]^{-1}\left(\left[P_Y\right] - W\left[Q_X\right]W^T\right)\lambda\\ &= \left(I - \left[\sqrt{P_Y}\right]^{-1}W\left[\sqrt{Q_X}\right]\left[\sqrt{Q_X}\right]W^T\left[\sqrt{P_Y}\right]^{-1}\right)\left[\sqrt{P_Y}\right]\lambda\\ &= \left(I - BB^T\right)\left[\sqrt{P_Y}\right]\lambda\end{aligned}$$

where B is the DTM from Definition 1.3.4. This means the most probable normalized channel perturbations are:

$$\forall x\in\mathcal{X},\ K_x^* = \left[\sqrt{W_x}\right]\left(I - 1W_x^T\right)\lambda$$

where λ satisfies:

$$\left(I - BB^T\right)\left[\sqrt{P_Y}\right]\lambda = K_Y.$$

The matrix $I - BB^T$ is not invertible, because the matrix BB^T has an eigenvalue of 1 which means $I - BB^T$ has an eigenvalue of 0. Since we assumed that $\forall x\in\mathcal{X}, \forall y\in\mathcal{Y},\ P_{X,Y}(x,y) > 0$, every entry of B is strictly positive. So, every entry of BB^T is strictly positive. By the Perron-Frobenius theorem, BB^T has a unique largest real eigenvalue such that every other eigenvalue of BB^T is strictly smaller in magnitude. We know that this largest Perron-Frobenius eigenvalue of BB^T is 1, and all other eigenvalues are non-negative real numbers less than 1, as BB^T is positive semidefinite. Hence, the algebraic multiplicity of the eigenvalue 0 of $I - BB^T$ is 1. This means $I - BB^T$ has a nullity of 1. It is known from [4] that for BB^T, the eigenvector corresponding to the eigenvalue of 1 is $\sqrt{P_Y}$. This means that:

$$\text{nullspace}\left(I - BB^T\right) = \text{span}\left(\sqrt{P_Y}\right).$$

Thus, if we find any particular solution $\left[\sqrt{P_Y}\right]\lambda = x$ to the marginal constraint:

$$\left(I - BB^T\right)x = K_Y$$

then all other solutions are of the form $\left[\sqrt{P_Y}\right]\lambda = x + v$, where $v\in\text{nullspace}\left(I - BB^T\right)$. We first argue that such a particular solution x exists. Since K_Y is a valid normalized perturbation, K_Y is orthogonal to $\sqrt{P_Y}$: $\sqrt{P_Y}^T K_Y = 0$. This means we have:

$$K_Y\in\text{nullspace}\left(I - BB^T\right)^\perp = \text{range}\left(\left(I - BB^T\right)^T\right) = \text{range}\left(I - BB^T\right)$$

where the first equality holds by the fundamental theorem of linear algebra, and the second equality holds because $I - BB^T$ is symmetric. Hence, a particular solution x exists for $\left(I - BB^T\right) x = K_Y$. We can find a particular solution using the Moore-Penrose pseudoinverse:

$$x = \left(I - BB^T\right)^\dagger K_Y$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse. It can be shown that this x is actually the minimum 2-norm solution to the associated system of equations, which means that it does not contain any component in the nullspace of $I - BB^T$. So, solutions to the marginal constraint are of the form:

$$\begin{aligned} \left[\sqrt{P_Y}\right] \lambda &= \left(I - BB^T\right)^\dagger K_Y + c\sqrt{P_Y} \\ \lambda &= \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger K_Y + c1 \end{aligned}$$

where $c \in \mathbb{R}$ is an arbitrary constant, and 1 is a column vector with all entries equal to 1. Now notice that the most probable normalized channel perturbations are:

$$\begin{aligned} \forall x \in \mathcal{X},\ K_x^* &= \left[\sqrt{W_x}\right] \left(I - 1W_x^T\right) \left(\left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger K_Y + c1 \right) \\ &= \left[\sqrt{W_x}\right] \left(I - 1W_x^T\right) \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger K_Y + c \left[\sqrt{W_x}\right] \left(I - 1W_x^T\right) 1 \\ &= \left[\sqrt{W_x}\right] \left(I - 1W_x^T\right) \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger K_Y \end{aligned} \tag{4.65}$$

which shows that the component of λ corresponding to the nullspace of $I - BB^T$ vanishes. Therefore, we are left with a unique solution as we would expect from the strict convexity of the objective function of problem 4.62. Rewriting equation 4.65 using additive perturbations, we have:

$$\forall x \in \mathcal{X},\ J_x^* = \left([W_x] - W_x W_x^T\right) \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger \left[\sqrt{P_Y}\right]^{-1} J_Y$$

as claimed.

We must now find the optimal value of the objective function of problem 4.62:

$$\frac{1}{2} \sum_{x \in \mathcal{X}} Q_X(x) \left\| J_x^* \right\|_{W_x}^2 = \frac{1}{2} \sum_{x \in \mathcal{X}} Q_X(x) \left\| K_x^* \right\|^2.$$

Using equation 4.65, we have:

$$\sum_{x \in \mathcal{X}} Q_X(x) \left\| K_x^* \right\|^2 = \sum_{x \in \mathcal{X}} Q_X(x) K_x^{*T} K_x^* =$$

$$K_Y^T \left(I - BB^T\right)^\dagger \left[\sqrt{P_Y}\right]^{-1} \left(\sum_{x \in \mathcal{X}} Q_X(x) \left(I - W_x 1^T\right) [W_x] \left(I - 1W_x^T\right) \right) \left[\sqrt{P_Y}\right]^{-1} \left(I - BB^T\right)^\dagger K_Y$$

where we also use the fact that the Moore-Penrose pseudoinverse of a symmetric matrix is symmetric. The summation in the middle can be evaluated to be:

$$\begin{aligned}\sum_{x\in\mathcal{X}} Q_X(x)\left(I - W_x 1^T\right)[W_x]\left(I - 1W_x^T\right) &= [P_Y] - \sum_{x\in\mathcal{X}} Q_X(x) W_x W_x^T \\ &= [P_Y] - W[Q_X]W^T \\ &= \left[\sqrt{P_Y}\right]\left(I - BB^T\right)\left[\sqrt{P_Y}\right]\end{aligned}$$

which implies that:

$$\begin{aligned}\sum_{x\in\mathcal{X}} Q_X(x)\left\|K_x^*\right\|^2 &= K_Y^T\left(I - BB^T\right)^\dagger\left(I - BB^T\right)\left(I - BB^T\right)^\dagger K_Y \\ &= K_Y^T\left(I - BB^T\right)^\dagger K_Y\end{aligned}$$

where the second equality follows from the definition of the Moore-Penrose pseudoinverse. Therefore, we have:

$$\frac{1}{2}\sum_{x\in\mathcal{X}} Q_X(x)\left\|K_x^*\right\|^2 = \frac{1}{2}K_Y^T\left(I - BB^T\right)^\dagger K_Y \tag{4.66}$$

which is equivalent to:

$$\frac{1}{2}\sum_{x\in\mathcal{X}} Q_X(x)\left\|J_x^*\right\|_{W_x}^2 = \frac{1}{2}J_Y^T\left[\sqrt{P_Y}\right]^{-1}\left(I - BB^T\right)^\dagger\left[\sqrt{P_Y}\right]^{-1} J_Y.$$

This completes the proof. □

The local solutions in Theorem 4.3.4 match the local approximations to the global solutions of Theorem 4.3.1. This can be seen by comparing equations 4.63 and 4.64 with equations 4.58 and 4.59, respectively. The next section utilizes Theorem 4.3.4 to model channel perturbations as additive Gaussian noise.

4.4 Modeling Channel Perturbations as Gaussian Noise

In this section, we conform to the notation set forth in section 4.1. We are given a pair of jointly distributed discrete random variables (X, Y) with joint pmf $P_{X,Y}$, where the marginal pmfs of X and Y are P_X and $P_Y = WP_X$, respectively. We perceive X as the input to a discrete memoryless channel with output Y. Recall from section 4.1 that:

$$\hat{P}_{X_1^n} = P_X + J_X \tag{4.67}$$

$$\hat{P}_{Y_1^n} = P_Y + J_Y \tag{4.68}$$

$$\forall x \in \mathcal{X},\ V_x = W_x + J_x \tag{4.69}$$

where J_X, J_Y, and $J_x, x \in \mathcal{X}$, are additive perturbations of the empirical distributions from their corresponding theoretical distributions. The empirical distributions are computed from (X_1^n, Y_1^n), where $X_1^n \in \mathcal{X}^n$ is some input string which may not be i.i.d., and $Y_1^n \in \mathcal{Y}^n$ is the output of X_1^n

through the discrete memoryless channel. We will let $P^n_{X,Y}$ denote the probability distribution of (X_1^n, Y_1^n) in the ensuing derivations. In section 4.1, we derived the source-channel decomposition of the output perturbation in equation 4.9:

$$J_Y \overset{\text{local}}{=} WJ_X + J_W P_X \tag{4.70}$$

where we neglect all second order terms. Although we were assuming that X_1^n are i.i.d. in section 4.1, equation 4.70 holds without this assumption. Indeed, all we really need to derive it is that the empirical source and channel conditional distributions are close to the theoretical distributions (i.e. J_X and J_W have a common $\epsilon > 0$ factor in front of them). Observe that at the sender end, we can control the empirical source pmf, and hence, the source perturbation J_X. At the receiver end, we are interested in the output perturbation J_Y and the corresponding empirical output pmf. We neither have control over nor have much interest in the actual channel perturbations $J_W P_X$, besides their effect on J_Y. Hence, we now contemplate modeling the channel perturbations as additive noise. Specifically, we seek to model $J_W P_X$ in some meaningful sense as additive (jointly) Gaussian noise Z. This changes equation 4.70 into:

$$J_Y = WJ_X + Z \tag{4.71}$$

where we drop the $\overset{\text{local}}{=}$, and it is understood that we are operating under appropriate local approximations. It will be convenient to use normalized perturbations rather than additive perturbations in our calculations. So, equation 4.71 is transformed into:

$$K_Y = BK_X + Z \tag{4.72}$$

where $K_X = \left[\sqrt{P_X}\right]^{-1} J_X$, $K_Y = \left[\sqrt{P_Y}\right]^{-1} J_Y$, B is the DTM, and $Z = \left[\sqrt{P_Y}\right]^{-1} J_W P_X$ now denotes the additive Gaussian noise corresponding to the transformed (but equivalent) model. For some fixed large n, the source normalized perturbation K_X is a random vector which has a certain probability distribution associated to it. It can be perceived as the input to a multiple-input multiple-output (MIMO) channel with output K_Y, channel matrix B, and additive Gaussian noise Z. The next theorem characterizes Z by appropriately approximating the conditional probability of K_Y given K_X.

Theorem 4.4.1 (Gaussian MIMO Channel for Perturbations)**.** *For large n, the source-channel decomposition of the output perturbation can be modeled as a MIMO channel with additive Gaussian noise:*

$$K_Y = BK_X + Z$$

where B is the divergence transition matrix, $Z \sim \mathcal{N}(0, \Sigma)$ is Gaussian distributed with covariance matrix $\Sigma = \frac{1}{n}\left(I - BB^T\right)$, the input K_X is independent of Z, and the model holds under exponential and local approximations.

Proof.
We first appropriately approximate the conditional probability of K_Y given K_X. Consider any input normalized perturbation k_X and any output normalized perturbation k_Y. Let $p_Y = P_Y + \epsilon\left[\sqrt{P_Y}\right] k_Y$

and $p_X = P_X + \epsilon \left[\sqrt{P_X}\right] k_X$. Observe that:

$$\begin{aligned} P^n_{X,Y}\left(K_Y = k_Y \mid K_X = k_X\right) &= P^n_{X,Y}\left(\hat{P}_{Y_1^n} = p_Y \mid \hat{P}_{X_1^n} = p_X\right) \\ &= \sum_{v:\ vp_X = p_Y} P^n_{X,Y}\left(V = v \mid \hat{P}_{X_1^n} = p_X\right) \\ &\doteq \max_{v:\ vp_X = p_Y} P^n_{X,Y}\left(V = v \mid \hat{P}_{X_1^n} = p_X\right) \end{aligned}$$

where the summation indexes over a polynomial in n number of terms because the number of possible empirical channel matrices v which satisfy $vp_X = p_Y$ is polynomial in n for fixed n, and the third equation holds in an exponential approximation sense by the Laplace principle since each term in the sum is exponentially decaying. We recognize that the right hand side of this equation is precisely the problem of finding the most probable empirical channel conditional pmfs given the input and output empirical pmfs. The global case of this problem was identified in statements 4.43, 4.50, and 4.53 of section 4.3. Using these results, we have:

$$P^n_{X,Y}\left(K_Y = k_Y \mid K_X = k_X\right) \doteq \exp\left(-n \min_{v:\ vp_X = p_Y} D(v||W|p_X)\right).$$

The local case of the constrained minimum conditional KL divergence problem in the exponent of this equation is defined in problem 4.62, and the local optimal value of the exponent can be found in Theorem 4.3.4. We note that the constraint $vp_X = p_Y$ is indeed locally equivalent (neglecting second order terms) to the last constraint of problem 4.62 when we fix J_Y in the constraint appropriately. Hence, using Theorem 4.3.4, we get:

$$P^n_{X,Y}\left(K_Y = k_Y \mid K_X = k_X\right) \overset{\text{local}}{\doteq} \exp\left(-\frac{n}{2} \sum_{x \in \mathcal{X}} P_X(x) \left\|K_x^*\right\|^2\right)$$

where the notation $\overset{\text{local}}{\doteq}$ implies we apply the exponential approximation and then employ local approximations to solve the extremal problem in the exponent. The $K_x^*,\ x \in \mathcal{X}$, are given in equation 4.65 in the proof of Theorem 4.3.4:

$$\forall x \in \mathcal{X},\ K_x^* = \left[\sqrt{W_x}\right]\left(I - 1W_x^T\right)\left[\sqrt{P_Y}\right]^{-1}\left(I - BB^T\right)^{\dagger}\left(k_Y - Bk_X\right)$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse, and we fix K_Y (corresponding to J_Y) of problem 4.62 to be $k_Y - Bk_X$. Using equation 4.66 in the proof of Theorem 4.3.4, we have:

$$P^n_{X,Y}\left(K_Y = k_Y \mid K_X = k_X\right) \overset{\text{local}}{\doteq} \exp\left(-\frac{n}{2}\left(k_Y - Bk_X\right)^T\left(I - BB^T\right)^{\dagger}\left(k_Y - Bk_X\right)\right) \tag{4.73}$$

which is exactly in the form of a Gaussian pdf without the normalization.

We now illustrate that equation 4.73 also represents an approximation to the probability of Z. To this end, we argue that J_W is independent of J_X in equation 4.70, which will mean that Z is independent of K_X in equation 4.72. Notice that knowing J_X is equivalent to knowing $\hat{P}_{X_1^n}$. Given $\hat{P}_{X_1^n}$, we know that each V_x is constructed from $n\hat{P}_{X_1^n}(x)$ samples. However, the values of $V_x(y)$ are

determined by drawing independently (conditioned on knowing x) from W_x by the memorylessness of the channel. So, the only information J_X provides about each J_x is the number of samples used to construct each V_x. As $n \to \infty$, the number of samples used to construct each V_x also tends to ∞, which means J_W becomes independent of J_X. This means we can model the noise Z as independent of the source perturbation K_X. Hence, for any input normalized perturbation k_X and any output normalized perturbation k_Y:

$$P^n_{X,Y}\left(K_Y = k_Y \mid K_X = k_X\right) = P^n_{X,Y}\left(Z = k_Y - Bk_X \mid K_X = k_X\right) = P^n_{X,Y}\left(Z = k_Y - Bk_X\right)$$

where the first equality follows from equation 4.72, and the second equality follows from the (asymptotic) independence of Z and K_X. From equation 4.73, we get:

$$P^n_{X,Y}\left(Z = k_Y - Bk_X\right) \overset{\text{local}}{\doteq} \exp\left(-\frac{n}{2}\left(k_Y - Bk_X\right)^T \left(I - BB^T\right)^\dagger \left(k_Y - Bk_X\right)\right)$$

where we can let $z = k_Y - Bk_X$ to give:

$$P^n_{X,Y}\left(Z = z\right) \overset{\text{local}}{\doteq} \exp\left(-\frac{1}{2} z^T \left(\frac{1}{n}\left(I - BB^T\right)\right)^\dagger z\right). \tag{4.74}$$

From this equation and our earlier argument, we see that it is reasonable to model Z as a jointly Gaussian random vector independent of K_X. It is evident that the model is more accurate when n is large, and holds under exponential and local approximations. We let Z have jointly Gaussian distribution:

$$Z \sim \mathcal{N}\left(0, \Sigma\right)$$

with covariance matrix:

$$\Sigma = \frac{1}{n}\left(I - BB^T\right).$$

To characterize the set of values Z takes on, observe that by virtue of being valid normalized perturbations, every k_Y is orthogonal to $\sqrt{P_Y}$ and every k_X is orthogonal to $\sqrt{P_X}$. Since B has right singular vector $\sqrt{P_X}$ and left singular vector $\sqrt{P_Y}$ corresponding to singular value 1 (see proof of Theorem 3.2.4), Bk_X is orthogonal to $\sqrt{P_Y}$ as well. This means the set of values Z takes on is:

$$\left\{z \in \mathbb{R} : \sqrt{P_Y}^T z = 0\right\}$$

which implies:

$$\sqrt{P_Y}^T Z = 0.$$

We know from the proof of Theorem 4.3.4 that $I - BB^T$ is singular and has nullspace:

$$\text{nullspace}\left(I - BB^T\right) = \text{span}\left(\sqrt{P_Y}\right).$$

This is consistent with the linear dependence relation $\sqrt{P_Y}^T Z = 0$ that Z must satisfy to be a valid normalized perturbation of P_Y. It also means that Z is a degenerate jointly Gaussian random vector. So, we cannot construct a pdf for Z with respect to the Lebesgue measure on $\mathbb{R}^{|\mathcal{Y}|}$. However, Z does have a pdf with respect to the "equivalent" Lebesgue measure on a $(|\mathcal{Y}| - 1)$-dimensional

subspace of $\mathbb{R}^{|\mathcal{Y}|}$. The new measure can be defined using the disintegration theorem from measure theory, and the pdf of Z with respect to this measure is:

$$f_Z(z) = \frac{1}{\sqrt{(2\pi)^{|\mathcal{Y}|-1} \operatorname{pdet}(\Sigma)}} \exp\left(-\frac{1}{2} z^T \Sigma^\dagger z\right)$$

where $\operatorname{pdet}(\cdot)$ is the pseudo-determinant (the product of the non-zero eigenvalues). This pdf matches the form of equation 4.74. The appropriate normalization constant can be inserted into equation 4.74 to get:

$$P_{X,Y}^n(Z = z) \overset{\text{local}}{\doteq} \frac{1}{\sqrt{(2\pi)^{|\mathcal{Y}|-1} \operatorname{pdet}\left(\frac{1}{n}\left(I - BB^T\right)\right)}} \exp\left(-\frac{1}{2} z^T \left(\frac{1}{n}\left(I - BB^T\right)\right)^\dagger z\right)$$

because the exponential approximation causes the constant to vanish. Indeed, we have:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{1}{\sqrt{(2\pi)^{|\mathcal{Y}|-1} \operatorname{pdet}\left(\frac{1}{n}\left(I - BB^T\right)\right)}}\right) &= -\frac{1}{2} \lim_{n\to\infty} \frac{1}{n} \log\left(\operatorname{pdet}\left(\frac{1}{n}\left(I - BB^T\right)\right)\right) \\
&= \frac{|\mathcal{Y}| - 1}{2} \lim_{n\to\infty} \frac{\log(n)}{n} \\
&= 0
\end{aligned}$$

where the second equality holds because $I - BB^T$ has a nullity of 1. This completes the proof. □

Theorem 4.4.1 models channel perturbations as additive Gaussian noise Z under exponential and local approximations. We illustrate the elegant intuition behind this result with some modest calculations. Recall from the discussion at the onset of this chapter that drawing i.i.d. X_1^n engenders a uniform KL divergence ball of perturbations, K_X, around P_X. By the memorylessness of the channel, Y_1^n are also i.i.d. and the perturbations, K_Y, form a uniform KL divergence ball around P_Y. From equations 4.10 and 4.11, we have for any input normalized perturbation k_X and any output normalized perturbation k_Y:

$$P_{X,Y}^n(K_X = k_X) \overset{\text{local}}{\doteq} \exp\left(-\frac{n}{2} k_X^T \left(I - \sqrt{P_X}\sqrt{P_X}^T\right)^\dagger k_X\right) \tag{4.75}$$

$$P_{X,Y}^n(K_Y = k_Y) \overset{\text{local}}{\doteq} \exp\left(-\frac{n}{2} k_Y^T \left(I - \sqrt{P_Y}\sqrt{P_Y}^T\right)^\dagger k_Y\right) \tag{4.76}$$

where we apply local approximations to the KL divergences which normally reside in the exponents. Note that the Moore-Penrose pseudoinverses in the exponents can be replaced by identity matrices because $\sqrt{P_X}^T k_X = 0$ and $\sqrt{P_Y}^T k_Y = 0$. We model K_X and K_Y as jointly Gaussian random vectors:

$$K_X \sim \mathcal{N}\left(0, \frac{1}{n}\left(I - \sqrt{P_X}\sqrt{P_X}^T\right)\right) \text{ and } K_Y \sim \mathcal{N}\left(0, \frac{1}{n}\left(I - \sqrt{P_Y}\sqrt{P_Y}^T\right)\right).$$

The nullspace of the covariance matrix of K_X is the span of $\sqrt{P_X}$ and the nullspace of the covariance matrix of K_Y is the span of $\sqrt{P_Y}$, because $\sqrt{P_X}^T K_X = 0$ and $\sqrt{P_Y}^T K_Y = 0$. Hence, the rank 1

subtractions in the covariances ensure that K_X and K_Y are valid normalized perturbations. On the other hand, the identity matrices in the covariances portray that K_X and K_Y lie on spherical level sets around P_X and P_Y, respectively. Our Gaussian MIMO channel model for normalized perturbations is:

$$K_Y = BK_X + Z$$

where $Z \sim \mathcal{N}\left(0, \frac{1}{n}\left(I - BB^T\right)\right)$ is the independent additive Gaussian noise from Theorem 4.4.1. We verify that:

$$\begin{aligned}
\mathbb{E}\left[K_Y K_Y^T\right] &= B\mathbb{E}\left[K_X K_X^T\right] B^T + \mathbb{E}\left[ZZ^T\right] \\
&= \frac{1}{n} B\left(I - \sqrt{P_X}\sqrt{P_X}^T\right) B^T + \frac{1}{n}\left(I - BB^T\right) \\
&= \frac{1}{n}\left(I - B\sqrt{P_X}\sqrt{P_X}^T B^T\right) \\
&= \frac{1}{n}\left(I - \sqrt{P_Y}\sqrt{P_Y}^T\right)
\end{aligned}$$

where the first equality holds by independence of K_X and Z, the second equality holds by substituting in the covariances, and the final equality holds because B has right singular vector $\sqrt{P_X}$ and left singular vector $\sqrt{P_Y}$ corresponding to singular value 1 (see proof of Theorem 3.2.4). Hence, when we consider the Gaussian MIMO channel model for normalized perturbations, we see that spherical source perturbation level sets pass through the channel and are warped into ellipsoids governed by the spectral decomposition of $BB^T - \sqrt{P_Y}\sqrt{P_Y}^T$. The noise due to channel perturbations then takes these ellipsoidal level sets and transforms them back into spherical output perturbation level sets. This is the source-channel decomposition of the large deviation behavior of the output perturbation when i.i.d. X_1^n is sent through a discrete memoryless channel.

As a final remark, we briefly examine how the results of this chapter, which culminate in Theorem 4.4.1, can be used for communications purposes. Envision a scenario where we are given a discrete memoryless channel which has a random permutation attached to it. In other words, codewords are randomly permuted before being passed through a classical discrete memoryless channel. In traditional channel coding, it is known that fixed composition codes achieve capacity. So, traditional codebooks typically try to fix the empirical distribution of codewords (at least approximately), and embed information through the permutations (ordering) of symbols within the codewords. Such coding schemes would fail lamentably in a channel with random permutation, because all information associated with the ordering of symbols in a codeword is lost. In such a setting, information must be communicated by varying the empirical distributions of the codewords. It is conceivable that we would want to use codeword empirical distributions around some fixed source distribution P_X (perhaps the capacity achieving distribution). Then, the normalized perturbations, K_X, carry the information from the messages. Theorem 4.4.1 allows us to view the effect of any discrete memoryless channel on normalized perturbations as an additive Gaussian noise MIMO channel. Therefore, we can select normalized perturbations which maximize the information sent down the MIMO channel based on the noise statistics. We encourage the interested reader to refer to [8] for further details regarding the utility of Theorem 4.4.1 in communications.

Chapter 5

Spectral Decomposition of Infinite Alphabet Channels

In chapter 1, we delineated a method of analyzing linear information coupling problems on discrete and finite channels using the singular value decomposition (SVD) of the DTM [4]. We then explored several aspects of this framework in chapters 3 and 4. A natural next step is to consider channels whose input and output random variables have infinite range. Since theory from linear algebra specifies how the SVD of a matrix may be calculated, and numerical linear algebra provides elegant algorithms that carry out this task, finding the SVD of the DTM is a trivial consideration for discrete and finite channels. However, finding the SVD for more general channels requires powerful analytical tools from functional analysis. Indeed, the spectral theory of linear operators (which subsumes the SVD) in functional spaces is a deep and subtle subject. In this final discourse, we turn our attention to the many intricacies of computing the SVD for channels which are not discrete and finite. Much of the measure theory, integration theory, and functional analysis used in this chapter may be found in texts on functional analysis like [28] and [29].

Due to the dearth of algorithms which readily compute SVDs in functional spaces, the SVD for more general channels must be computed on a case by case basis. [23] derives it for AWGN channels with Gaussian input, and [8] derives it for Poisson channels with exponential input. In this chapter, we reintroduce the appropriate perturbation spaces and notation for general channels, and then present a criterion to verify whether the singular vectors of such channels (after transformation) are orthogonal polynomials. This criterion is used to generalize the spectral decomposition results for Poisson channels, and also find SVDs for other channels.

5.1 Preliminary Definitions and Notation

In this section, we reiterate and generalize several definitions pertaining to local perturbation spaces and the divergence transition matrix (DTM) from sections 1.2 and 1.3 in chapter 1. Some of these generalizations were briefly mentioned in section 2.1 in chapter 2, and implicitly used in section 3.4 in chapter 3. Throughout this chapter, we will be primarily interested in channels whose input and output random variables have infinite range. We now define the notion of an "infinite alphabet channel" more precisely.

Definition 5.1.1 (Infinite Alphabet Channel)**.** Given a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, an infinite alphabet channel consists of a discrete or continuous input random variable $X : \Omega \to \mathcal{X}$ with infinite range $\mathcal{X} \subseteq \mathbb{R}$, a discrete or continuous output random variable $Y : \Omega \to \mathcal{Y}$ with infinite range $\mathcal{Y} \subseteq \mathbb{R}$, and conditional probability distributions $P_{Y|X}(\cdot|x), \ \ x \in \mathcal{X}$.

The AWGN and Poisson channels we mentioned earlier are both infinite alphabet channels. They will be formally defined later as we proceed through our discussion. We will use the notation $\mathbb{P}_X$ and $\mathbb{P}_Y$ to denote the probability laws of X and Y, respectively. These are both push-forward measures of $\mathbb{P}$. Moreover, we will always restrict ourselves to situations where $\mathbb{P}_X$ is absolutely continuous with respect to a σ-finite measure λ, and $\mathbb{P}_Y$ is absolutely continuous with respect to a σ-finite measure μ. The Radon-Nikodym derivative of $\mathbb{P}_X$ with respect to λ is denoted P_X, and the Radon-Nikodym derivative of $\mathbb{P}_Y$ with respect to μ is denoted P_Y. In particular, we will consider two special cases of λ and μ. If X (respectively Y) is a discrete random variable and $\mathcal{X}$ (respectively $\mathcal{Y}$) is countably infinite, then $\mathbb{P}_X$ (respectively $\mathbb{P}_Y$) is a probability measure and λ (respectively μ) is the counting measure on the measurable space $\left(\mathcal{X}, 2^{\mathcal{X}}\right)$ (respectively $\left(\mathcal{Y}, 2^{\mathcal{Y}}\right)$), so that P_X (respectively P_Y) is a pmf with infinite non-zero mass points. If X (respectively Y) is a continuous random variable and $\mathcal{X}$ (respectively $\mathcal{Y}$) is uncountably infinite, then $\mathbb{P}_X$ (respectively $\mathbb{P}_Y$) is a probability measure and λ (respectively μ) is the Lebesgue measure on the measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra, so that P_X (respectively P_Y) is a pdf. The conditional distributions $P_{Y|X}(\cdot|x), \ \ x \in \mathcal{X}$, and $P_{X|Y}(\cdot|y), \ \ y \in \mathcal{Y}$, will be regarded as pmfs or pdfs with respect to μ and λ, respectively. Finally, we note that all integrals in this chapter will refer to abstract Lebesgue integrals with respect to general measures. For example, if λ is the counting measure, then for any integrable measurable function $f : \mathcal{X} \to \mathbb{R}$, we have:

$$\mathbb{E}\left[f(X)\right] = \int_{\mathcal{X}} f \ \ d\mathbb{P}_X = \int_{\mathcal{X}} f P_X \ \ d\lambda = \sum_{x \in \mathcal{X}} P_X(x) f(x).$$

Likewise, if λ is the Lebesgue measure, then for any integrable measurable function $f : \mathbb{R} \to \mathbb{R}$, we have:

$$\mathbb{E}\left[f(X)\right] = \int_{\mathcal{X}} f \ \ d\mathbb{P}_X = \int_{\mathcal{X}} f P_X \ \ d\lambda$$

where the second integral is a standard Lebesgue integral.

We now create an analogous picture of perturbation spaces (as in the discrete and finite channel case) for infinite alphabet channels. Fix a reference pmf or pdf, P_X, satisfying:

$$\lambda\left(\{x \in \mathcal{X} : P_X(x) = 0\}\right) = 0$$

which means P_X is not at the edge of the stochastic manifold of distributions on $\mathcal{X}$, and has a well-defined neighborhood so that local perturbations can exist in all directions around it. The perturbed input pmf or pdf, Q_X, can be written as:

$$\forall x \in \mathcal{X}, \ \ Q_X(x) = P_X(x) + \epsilon J_X(x) \tag{5.1}$$

where the (measurable) additive perturbation function, J_X, satisfies:

$$\int_{\mathcal{X}} J_X \ \ d\lambda = 0 \tag{5.2}$$

to ensure that Q_X is normalized, and $\epsilon > 0$ is chosen small enough to ensure that:

$$\lambda\left(\{x \in \mathcal{X} : Q_X(x) < 0\}\right) = 0.$$

As in the discrete and finite channel case, we may also define the (measurable) normalized and log-likelihood perturbation functions, K_X and L_X, respectively:

$$\forall x \in \mathcal{X}, \;\; K_X(x) \triangleq \frac{J_X(x)}{\sqrt{P_X(x)}} \tag{5.3}$$

$$\forall x \in \mathcal{X}, \;\; L_X(x) \triangleq \frac{J_X(x)}{P_X(x)} \tag{5.4}$$

which are well-defined because $\lambda\left(\{x \in \mathcal{X} : P_X(x) = 0\}\right) = 0$. Note that K_X satisfies the valid normalized perturbation constraint:

$$\int_{\mathcal{X}} \sqrt{P_X} K_X \; d\lambda = 0 \tag{5.5}$$

and L_X satisfies the valid log-likelihood perturbation constraint:

$$\int_{\mathcal{X}} L_X \; d\mathbb{P}_X = \int_{\mathcal{X}} P_X L_X \; d\lambda = 0. \tag{5.6}$$

The infinite alphabet channel transforms the input marginal pmfs or pdfs, P_X and Q_X, into output marginal pmfs or pdfs, P_Y and Q_Y, respectively. Formally, we have:

$$\forall y \in \mathcal{Y}, \;\; P_Y(y) = \int_{\mathcal{X}} P_{Y|X}(y|x) P_X(x) \; d\lambda(x) \tag{5.7}$$

$$\forall y \in \mathcal{Y}, \;\; Q_Y(y) = \int_{\mathcal{X}} P_{Y|X}(y|x) Q_X(x) \; d\lambda(x) \tag{5.8}$$

where we write the variable x inside the integrals to clarify what we are integrating over. We can also define output perturbations analogously to input perturbations. The (measurable) additive output perturbation function, J_Y, is defined by:

$$\forall y \in \mathcal{Y}, \;\; Q_Y(y) = P_Y(y) + \epsilon J_Y(y) \tag{5.9}$$

where $\epsilon > 0$ is chosen small enough to ensure that:

$$\mu\left(\{y \in \mathcal{Y} : Q_Y(x) < 0\}\right) = 0.$$

Note that the infinite alphabet channel transforms the input additive perturbation J_X into the output additive perturbation J_Y:

$$\forall y \in \mathcal{Y}, \;\; J_Y(y) = \int_{\mathcal{X}} P_{Y|X}(y|x) J_X(x) \; d\lambda(x). \tag{5.10}$$

Finally, the (measurable) normalized and log-likelihood perturbation functions, K_Y and L_Y, corresponding to J_Y are:

$$\forall y \in \mathcal{Y}, \;\; K_Y(y) \triangleq \frac{J_Y(y)}{\sqrt{P_Y(y)}} \tag{5.11}$$

$$\forall y \in \mathcal{Y}, \;\; L_Y(y) \triangleq \frac{J_Y(y)}{P_Y(y)} \tag{5.12}$$

where we assume that $\mu(\{y \in \mathcal{Y} : P_Y(y) = 0\}) = 0$. The perturbations functions J_Y, K_Y, and L_Y satisfy:

$$\int_{\mathcal{Y}} J_Y \ d\mu = 0 \tag{5.13}$$

$$\int_{\mathcal{Y}} \sqrt{P_Y} K_Y \ d\mu = 0 \tag{5.14}$$

$$\int_{\mathcal{Y}} L_Y \ d\mathbb{P}_Y = \int_{\mathcal{Y}} P_Y L_Y \ d\mu = 0 \tag{5.15}$$

to ensure that Q_Y normalizes to 1. It is easily verified that the input and output additive, normalized, and log-likelihood perturbations form separate vector spaces.

From equation 5.10, we see that the channel (conditional distributions $P_{Y|X}$) transforms any J_X into J_Y. We may also define operators which transform input normalized and log-likelihood perturbations into output normalized and log-likelihood perturbations. The operator which transforms K_X to K_Y is called the Divergence Transition Map (DTM), because 2-norms of normalized perturbations are proportional to local KL divergences. The operator which transforms L_X to L_Y is called the Log-likelihood Transition Map (LTM). We formally define these operators below.

Definition 5.1.2 (Perturbation Transition Maps)**.** Consider an infinite alphabet channel with input random variable X, output random variable Y, and conditional probability distributions $P_{Y|X}(\cdot|x), \ x \in \mathcal{X}$. Suppose the marginal distributions of X and Y satisfy $\lambda(\{x \in \mathcal{X} : P_X(x) = 0\}) = 0$ and $\mu(\{y \in \mathcal{Y} : P_Y(y) = 0\}) = 0$, respectively. Then, the channel, denoted A, transforms additive input perturbation functions into additive output perturbation functions:

$$\forall y \in \mathcal{Y}, \ \ A(J_X)(y) = \int_{\mathcal{X}} P_{Y|X}(y|x) J_X(x) \ d\lambda(x) = J_Y(y),$$

the divergence transition map (DTM), denoted B, transforms normalized input perturbation functions into normalized output perturbation functions:

$$\forall y \in \mathcal{Y}, \ \ B(K_X)(y) = \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x) \sqrt{P_X(x)} K_X(x) \ d\lambda(x) = K_Y(y),$$

and the log-likelihood transition map (LTM), denoted C, transforms log-likelihood input perturbation functions into log-likelihood output perturbation functions:

$$\begin{aligned} \forall y \in \mathcal{Y}, \ \ C(L_X)(y) &= \frac{1}{P_Y(y)} \int_{\mathcal{X}} P_{Y|X}(y|x) L_X(x) \ d\mathbb{P}_X(x) \\ &= \frac{1}{P_Y(y)} \int_{\mathcal{X}} P_{Y|X}(y|x) P_X(x) L_X(x) \ d\lambda(x) \\ &= \int_{\mathcal{X}} P_{X|Y}(x|y) L_X(x) \ d\lambda(x) \\ &= L_Y(y). \end{aligned}$$

From Definition 5.1.2, it is easily verified that A (channel), B (DTM), and C (LTM) are linear operators which transform vectors in an input perturbation vector space to vectors in an output

perturbation vector space. We are interested in computing the SVD of B. Indeed, this was part of the analysis in chapter 1 for discrete and finite channels. The first step towards computing an SVD is to find the spectral decomposition of the Gramian operator; this is in fact the majority of the computation of the SVD in the matrix case. The Gramian of a linear operator A is A^*A, where A^* denotes the adjoint operator of A. We note that referring to A^*A as the Gramian operator is non-standard usage. Typically, this terminology is reserved for matrices, but we use it for operators on functional spaces for lack of a better name. Regularity conditions which ensure that the spectral decompositions of self-adjoint Gramian operators exist come from the theory of Hilbert spaces in functional analysis. We will state and explain the pertinent regularity conditions in the theorem statements that follow, but a complete discussion of this topic is omitted for the sake of brevity. Gramian operators for the channel, DTM, and LTM are derived next. Technically, to find the adjoint operator of a given operator, we must first define the input and output Hilbert spaces and their inner products precisely. We must also assume that the original operator is bounded so that Riesz' representation theorem can be used to guarantee a unique adjoint. (In fact, the term "operator" is usually reserved for bounded linear maps on Banach spaces.) We will omit the proofs of boundedness in the next theorem despite the loss in rigor, because the Gramian operators of the channel and LTM will not be used in our main results. The derivation of the Gramian operator of B will be done rigorously before it is used.

Theorem 5.1.1 (Gramian Operators of Perturbation Transition Maps)**.** *Consider an infinite alphabet channel with input random variable X, output random variable Y, and conditional probability distributions $P_{Y|X}(\cdot|x)$, $x \in \mathcal{X}$. Suppose the marginal distributions of X and Y satisfy $\lambda(\{x \in \mathcal{X} : P_X(x) = 0\}) = 0$ and $\mu(\{y \in \mathcal{Y} : P_Y(y) = 0\}) = 0$, respectively. Then, under appropriate regularity conditions:*

1. *The Gramian operator of the channel, A^*A, is given by:*

$$(A^*A)(J_X)(x) = \int_{\mathcal{X}} \Lambda_A(x,t) J_X(t)\ d\lambda(t) \quad \text{a.e. with respect to } \lambda$$

where equality holds for all $x \in \mathcal{X}$ except a set with measure 0, and the kernel of the operator is: $\Lambda_A(x,t) = \displaystyle\int_{\mathcal{Y}} P_{Y|X}(y|x) P_{Y|X}(y|t)\ d\mu(y)$.

2. *The Gramian operator of the DTM, B^*B, is given by:*

$$(B^*B)(K_X)(x) = \int_{\mathcal{X}} \Lambda_B(x,t) K_X(t)\ d\lambda(t) \quad \text{a.e. with respect to } \lambda$$

where equality holds for all $x \in \mathcal{X}$ except a set with measure 0, and the kernel of the operator is: $\Lambda_B(x,t) = \sqrt{\dfrac{P_X(t)}{P_X(x)}} \displaystyle\int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t)\ d\mu(y)$.

3. *The Gramian operator of the LTM, C^*C, is given by:*

$$(C^*C)(L_X)(x) = \int_{\mathcal{X}} \Lambda_C(x,t) L_X(t)\ d\lambda(t) \quad \text{a.e. with respect to } \lambda$$

where equality holds for all $x \in \mathcal{X}$ except a set with measure 0, and the kernel of the operator is: $\Lambda_C(x,t) = \displaystyle\int_{\mathcal{Y}} P_{X|Y}(x|y) P_{X|Y}(t|y)\ d\mu(y)$.

Proof.
We only derive the Gramian operator of A. The Gramian operators B and C can be found similarly. Suppose we are given the Hilbert space $\mathcal{H}_1$ of measurable, real, λ-square integrable functions on $\mathcal{X}$, and the Hilbert space $\mathcal{H}_2$ of measurable, real, μ-square integrable functions on $\mathcal{Y}$. Let the associated inner product of $\mathcal{H}_1$ be:

$$\forall f_1, f_2 \in \mathcal{H}_1, \;\; \langle f_1, f_2 \rangle_1 \triangleq \int_{\mathcal{X}} f_1 f_2 \;\; d\lambda,$$

and the associated inner product of $\mathcal{H}_2$ be:

$$\forall g_1, g_2 \in \mathcal{H}_2, \;\; \langle g_1, g_2 \rangle_2 \triangleq \int_{\mathcal{Y}} g_1 g_2 \;\; d\mu.$$

To find the Gramian of the channel, $A : \mathcal{H}_1 \to \mathcal{H}_2$, we first find the adjoint operator, $A^* : \mathcal{H}_2 \to \mathcal{H}_1$. By definition of the adjoint operator, for any $f \in \mathcal{H}_1$ and any $g \in \mathcal{H}_2$, we have:

$$\begin{aligned}
\langle A(f), g \rangle_2 &= \langle f, A^*(g) \rangle_1 \\
\int_{\mathcal{Y}} A(f)\, g \;\; d\mu &= \int_{\mathcal{X}} f A^*(g) \;\; d\lambda \\
\int_{\mathcal{Y}} \left(\int_{\mathcal{X}} P_{Y|X}(y|x) f(x) \;\; d\lambda(x) \right) g(y) \;\; d\mu(y) &= \int_{\mathcal{X}} f(x) \left(A^*(g) \right)(x) \;\; d\lambda(x) \\
\int_{\mathcal{X}} f(x) \left(\int_{\mathcal{Y}} P_{Y|X}(y|x) g(y) \;\; d\mu(y) \right) \;\; d\lambda(x) &= \int_{\mathcal{X}} f(x) \left(A^*(g) \right)(x) \;\; d\lambda(x)
\end{aligned}$$

where the second line follows from the definitions of the inner products, the third line follows from Definition 5.1.2, and the final line follows from the Fubini-Tonelli theorem because $A(f)g$ must be μ-integrable by the Cauchy-Schwarz-Bunyakovsky inequality in $\mathcal{H}_2$. Since the final equality is true for all $f \in \mathcal{H}_1$ and all $g \in \mathcal{H}_2$, the unique adjoint operator (assuming A is bounded or continuous) A^* is:

$$\forall x \in \mathcal{X}, \;\; A^*(g)(x) = \int_{\mathcal{Y}} P_{Y|X}(y|x) g(y) \;\; d\mu(y) \tag{5.16}$$

for every function $g \in \mathcal{H}_2$. The Gramian operator, $A^*A : \mathcal{H}_1 \to \mathcal{H}_1$, is straightforward to derive once we have the adjoint. For any $f \in \mathcal{H}_1$, we have:

$$\begin{aligned}
\forall x \in \mathcal{X}, \;\; \left(A^*A \right)(f)(x) &= \int_{\mathcal{Y}} P_{Y|X}(y|x) \left(\int_{\mathcal{X}} P_{Y|X}(y|t) f(t) \;\; d\lambda(t) \right) \;\; d\mu(y) \\
\left(A^*A \right)(f)(x) &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} P_{Y|X}(y|x) P_{Y|X}(y|t) \;\; d\mu(y) \right) f(t) \;\; d\lambda(t) \;\; a.e.
\end{aligned}$$

where the second equality holds almost everywhere with respect to λ, and it can be justified by the Fubini-Tonelli theorem. Since $A^*A(f) \in \mathcal{H}_1$, we have $A^*A(f) < \infty \;\; a.e.$ with respect to λ. This means the iterated integral in the first line is finite almost everywhere, which implies:

$$\int_{\mathcal{Y}} P_{Y|X}(y|x) \left(\int_{\mathcal{X}} P_{Y|X}(y|t) \left| f(t) \right| \;\; d\lambda(t) \right) \;\; d\mu(y) < \infty \;\; a.e.$$

with respect to λ. Applying the Fubini-Tonelli theorem to the iterated integral in the first line produces the second equality. This defines the Gramian operator of A. □

The Gramian operators A^*A, B^*B, and C^*C are self-adjoint. It can be checked that their kernels are symmetric. This is evident for A^*A and C^*C:

$$\forall x,t \in \mathcal{X},\ \ \Lambda_A(x,t) = \Lambda_A(t,x), \tag{5.17}$$
$$\forall x,t \in \mathcal{X},\ \ \Lambda_C(x,t) = \Lambda_C(t,x). \tag{5.18}$$

It holds for B^*B with a little more algebra using Bayes' rule:

$$\begin{aligned}
\forall x,t \in \mathcal{X},\ \ \Lambda_B(x,t) &= \sqrt{\frac{P_X(t)}{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t)\ d\mu(y) \\
&= \sqrt{\frac{P_X(t)}{P_X(x)}} \int_{\mathcal{Y}} \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} \frac{P_{X|Y}(t|y)P_Y(y)}{P_X(t)}\ d\mu(y) \\
&= \sqrt{\frac{P_X(x)}{P_X(t)}} \int_{\mathcal{Y}} P_{X|Y}(t|y) P_{Y|X}(y|x)\ d\mu(y) \\
&= \Lambda_B(t,x).
\end{aligned} \tag{5.19}$$

These observations parallel the real matrix case, where the Gramian matrix of a matrix is always symmetric. In the next section, we study the spectral decomposition of compact self-adjoint Gramian operators, and then specialize these results to find SVDs of DTMs. In particular, we concentrate on polynomial eigenfunctions because this unveils deeper insights into the fortuitous results of [23] and [8].

5.2 Polynomial Spectral Decomposition

Before we delve into spectral decompositions of compact self-adjoint operators, we pause to elucidate our interest in polynomial eigenfunctions. The importance of the spectral decomposition itself is evident from our discussion regarding linear information coupling problems for discrete and finite channels; this can be naturally extended to infinite alphabet channels. For infinite alphabet channels, the type of eigenfunction (for example, decaying exponential, sinusoid, or polynomial) can make a difference to its practicality in real-world applications. It is a worthwhile endeavor to find polynomial eigenfunctions because polynomials are easy to evaluate using computers, and this makes data analysis methods which use the local approximation technique computationally more efficient. [23] and [8] find that the eigenfunctions of Gramian operators of the DTM of the AWGN and Poisson channels are (appropriately weighted) orthogonal polynomials. Such results propel us to try and characterize the class of infinite alphabet channels which have (appropriately weighted) orthogonal polynomials as eigenfunctions of the Gramian operators of their DTMs.

5.2.1 Orthogonal Polynomial Eigenbasis for Compact Self-Adjoint Operators

The literature on spectral decomposition methods in Hilbert spaces does not offer any systematic means of computing spectral decompositions of compact self-adjoint operators. So, it is difficult to methodically derive polynomial spectral decompositions of Gramian operators corresponding to infinite alphabet channels. In the next theorem, we provide an elementary and intuitive condition which can be tested to determine if the orthogonal eigenbasis of a compact self-adjoint operator

in a Hilbert space consists of orthogonal polynomials. To present this theorem clearly, we first precisely state our assumptions.

Suppose we are given a measurable space $(X, \mathcal{G}, \nu)$ where $X \subseteq \mathbb{R}$ is an infinite set and ν is a σ-finite measure, and a separable Hilbert space $\mathcal{H}$ over $\mathbb{R}$:

$$\mathcal{H} = \mathcal{L}^2(X, \nu) \triangleq \{f : X \to \mathbb{R} : f \text{ measurable and } \nu\text{-square integrable}\}$$

which is the space of square integrable functions with domain $(X, \mathcal{G})$ and codomain $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra. The Hilbert space has inner product:

$$\forall f, g \in \mathcal{H}, \ \langle f, g \rangle \triangleq \int_X fg \ d\nu \tag{5.20}$$

and induced norm:

$$\forall f \in \mathcal{H}, \ \|f\| \triangleq \sqrt{\langle f, f \rangle}. \tag{5.21}$$

In such a Hilbert space of functions, equality is defined as equality almost everywhere with respect to the measure ν. Each vector in this space is really an equivalence class of functions that are all equal almost everywhere. In the remainder of our discussion in this chapter, equality of functions should be correctly interpreted in this manner. The σ-finiteness of ν ensures that several convenient results like the Radon-Nikodym theorem, the Fubini-Tonelli theorem, and Carathéodory's extension theorem are valid. Such results will be useful in ensuing derivations. For example, the Fubini-Tonelli theorem was already used to prove Theorem 5.1.1. On a different note, if X is a compact interval in $\mathbb{R}$ and ν is absolutely continuous with respect to the Lebesgue measure on X, then the Radon-Nikodym derivative of ν with respect to the Lebesgue measure can be thought of as a weight function of the inner product on $\mathcal{H}$. This portrays that equation 5.20 generalizes familiar weighted inner products which are used to define well-known orthogonal polynomials. The separability of $\mathcal{H}$ is equivalent to the existence of a countable complete orthonormal basis of $\mathcal{H}$. Here, "complete" refers to the denseness of the span of the vectors in the orthonormal basis. We only consider separable Hilbert spaces which have a unique countable complete orthonormal basis of polynomials. Assume that this unique complete orthonormal basis is $P = \{p_0, p_1, p_2, \dots\} \subseteq \mathcal{H}$, where $p_k : X \to \mathbb{R}$ is a polynomial with degree k. We can find P by starting with the monomials $\{1, x, x^2, \dots\}$ and applying the Gram-Schmidt algorithm. Note that when we say P is unique, we still allow arbitrary sign changes for each orthonormal polynomial. The existence of P implies that:

$$\int_X |x|^n \ d\nu(x) < \infty \tag{5.22}$$

for every $n \in \mathbb{N} = \{0, 1, 2, \dots\}$. We now define two properties of bounded linear operators on $\mathcal{H}$ (endomorphisms) which will be useful in presenting our results.

Definition 5.2.1 (Closure over Polynomials)**.** A bounded linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ between two Hilbert spaces is closed over polynomials if for any polynomial $p \in \mathcal{H}_1$, $T(p) \in \mathcal{H}_2$ is also a polynomial.

Definition 5.2.2 (Degree Preservation)**.** A bounded linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ between two Hilbert spaces which is closed over polynomials is degree preserving if for any polynomial $p \in \mathcal{H}_1$ with degree k, $T(p) \in \mathcal{H}_2$ is also a polynomial with degree at most k. T is strictly degree preserving if for any polynomial $p \in \mathcal{H}_1$ with degree k, $T(p) \in \mathcal{H}_2$ is also a polynomial with degree exactly k.

Using the assumptions stated earlier and these definitions, we present an equivalent condition to determine if the orthogonal eigenbasis of a compact self-adjoint operator consists of orthogonal polynomials.

Theorem 5.2.1 (Condition for Orthogonal Polynomial Eigenbasis)**.** *Let $T : \mathcal{H} \to \mathcal{H}$ be a compact self-adjoint linear operator. Then, the orthonormal eigenbasis of T is the orthonormal basis of polynomials $P = \{p_0, p_1, p_2, \dots\}$ of $\mathcal{H}$ if and only if T is closed over polynomials and degree preserving.*

Proof.
We first check that there exists a complete orthonormal eigenbasis of T. Indeed, by the spectral theorem for compact self-adjoint operators on a separable Hilbert space [28], T has a countable complete orthonormal eigenbasis with real eigenvalues. Let this complete orthonormal eigenbasis be $Q = \{q_0, q_1, q_2, \dots\} \subseteq \mathcal{H}$. Hence, we have:

$$T(q_i) = \alpha_i q_i$$

for every $i \in \mathbb{N}$, where $\alpha_i \in \mathbb{R}$ are the real eigenvalues.

If $Q = P$, then T is trivially closed over polynomials and degree preserving. This is because any polynomial in $\mathcal{H}$ with degree k is a linear combination of $\{p_0, \dots, p_k\}$. So, it suffices to prove the converse direction.

We prove the converse by strong induction. Suppose T is closed over polynomials and degree preserving. We need to show that $Q = P$. The first eigenfunction of T is the constant function (which is an orthonormal polynomial) $q_0 = p_0 \neq 0$, because T is closed over polynomials and degree preserving. This is the base case. Assume that the first $k+1$ eigenfunctions are the first $k+1$ orthonormal polynomials:

$$q_i = p_i, \text{ for } i \in \{0, \dots, k\}.$$

This is the inductive hypothesis. We now prove that $q_{k+1} = p_{k+1}$, which is the inductive step.

Since p_{k+1} is orthogonal to $\operatorname{span}(p_0, \dots, p_k) = \operatorname{span}(q_0, \dots, q_k)$, where the equality holds by the inductive hypothesis, we may write:

$$p_{k+1} = \sum_{j=k+1}^{\infty} \langle p_{k+1}, q_j \rangle q_j$$

which means that the partial sums converge to p_{k+1} in the sense of the induced norm in $\mathcal{H}$:

$$\lim_{m \to \infty} \left\| p_{k+1} - \sum_{j=k+1}^{m} \langle p_{k+1}, q_j \rangle q_j \right\| = 0.$$

Since T is a compact operator, it is bounded by definition. This means $\exists C > 0$ such that:

$$\forall f \in \mathcal{H}, \ \|T(f)\| \leq C \|f\|.$$

Hence, we have for any $m \geq k+1$:

$$\begin{aligned}
\left\| T\left(p_{k+1} - \sum_{j=k+1}^{m} \langle p_{k+1}, q_j \rangle q_j \right) \right\| &\leq C \left\| p_{k+1} - \sum_{j=k+1}^{m} \langle p_{k+1}, q_j \rangle q_j \right\| \\
\lim_{m\to\infty} \left\| T(p_{k+1}) - \sum_{j=k+1}^{m} \langle p_{k+1}, q_j \rangle T(q_j) \right\| &\leq C \lim_{m\to\infty} \left\| p_{k+1} - \sum_{j=k+1}^{m} \langle p_{k+1}, q_j \rangle q_j \right\| \\
\lim_{m\to\infty} \left\| T(p_{k+1}) - \sum_{j=k+1}^{m} \alpha_j \langle p_{k+1}, q_j \rangle q_j \right\| &= 0
\end{aligned}$$

where the second line follows from the linearity of T, and the third line uses the fact that q_j are eigenfunctions of T. We have shown that:

$$T(p_{k+1}) = \sum_{j=k+1}^{\infty} \alpha_j \langle p_{k+1}, q_j \rangle q_j$$

where the equality holds in the sense of the induced norm in $\mathcal{H}$ (as derived). Note that our argument is effectively a continuity argument, but we can use boundedness because the two notions are equivalent for linear operators on Banach spaces. Now observe that for any $m \geq k+1$:

$$\left\langle \sum_{j=k+1}^{m} \alpha_j \langle p_{k+1}, q_j \rangle q_j \ , \ p_i \right\rangle = 0 \ \text{ for } \ i \in \{0, \ldots, k\}$$

which holds by the inductive hypothesis. Hence, for any $i \in \{0, \ldots, k\}$ and any $m \geq k+1$:

$$\begin{aligned}
\left| \langle T(p_{k+1}), p_i \rangle - \left\langle \sum_{j=k+1}^{m} \alpha_j \langle p_{k+1}, q_j \rangle q_j \ , \ p_i \right\rangle \right| &= \left| \left\langle T(p_{k+1}) - \sum_{j=k+1}^{m} \alpha_j \langle p_{k+1}, q_j \rangle q_j \ , \ p_i \right\rangle \right| \\
&\leq \left\| T(p_{k+1}) - \sum_{j=k+1}^{m} \alpha_j \langle p_{k+1}, q_j \rangle q_j \right\| \|p_i\|
\end{aligned}$$

by the Cauchy-Schwarz-Bunyakovsky inequality. This means we can let $m \to \infty$ to get:

$$\langle T(p_{k+1}), p_i \rangle = 0 \ \text{ for } \ i \in \{0, \ldots, k\}.$$

We have effectively used the continuity of the inner product to prove this. As T is closed over polynomials and degree preserving, $T(p_{k+1})$ is a polynomial with degree at most $k+1$ that is orthogonal to all polynomials of degree k or less. Hence, $T(p_{k+1})$ must be a scaled version of the orthonormal polynomial with degree $k+1$:

$$T(p_{k+1}) = \beta p_{k+1}$$

for some $\beta \in \mathbb{R}$ which is possibly zero. This implies (without loss of generality) that $\alpha_{k+1} = \beta$ and $q_{k+1} = p_{k+1}$. Therefore, by strong induction, the complete orthonormal eigenbasis of T is $Q = P$. This proves the converse direction. □

Some remarks regarding Theorem 5.2.1 are in order. Firstly, we explain the compactness assumption on T. A bounded linear operator T on a separable Hilbert space is compact if the closure of the image of the closed unit ball is compact [28]. Intuitively, such operators are a natural extension of finite rank operators. We impose the compactness assumption on T, because the spectral theorem asserts that a complete orthonormal eigenbasis exists for compact self-adjoint operators. Secondly, we note that although Theorem 5.2.1 holds for any compact self-adjoint operator, we are only interested in Gramian operators of perturbation transition maps in this thesis. Gramian operators are positive (or positive semidefinite) self-adjoint operators, and perturbation transition maps are integral operators (using Theorem 5.1.1). Hence, the compact, positive, self-adjoint operators that we consider have the form:

$$\forall x \in X, \;\; T\left(f\right)\left(x\right) \triangleq \int_X \Lambda(x,s) f(s) \;\; d\nu(s)$$

where $\Lambda : X \times X \to \mathbb{R}$ is the symmetric kernel of the integral operator.

Finally, we elaborate on the separable Hilbert spaces, $\mathcal{H} = \mathcal{L}^2(X,\nu)$, by offering particular insight on why countable complete polynomial bases exist for such spaces. Let us restrict ourselves to the case where $X = [a,b]$ for some $-\infty < a < b < \infty$, its associated σ-algebra $\mathcal{G} = \mathcal{B}\left([a,b]\right)$ is the Borel σ-algebra on the compact interval, and ν is the (uniform) Lebesgue measure on $(X,\mathcal{G})$. Observe that polynomials are dense in the space of all continuous functions on X in the sup-norm sense (uniform convergence) by the Stone-Weierstrass theorem [30]. Since convergence in sup-norm implies convergence in $\mathcal{L}^2(X,\nu)$, polynomials are dense in the space of all continuous functions in the $\mathcal{L}^2(X,\nu)$ norm sense as well. Continuous functions on X (a compact interval) are in $\mathcal{L}^2(X,\nu)$ because they are bounded by the boundedness theorem. In fact, $\mathcal{L}^2(X,\nu)$ is the completion of the space of continuous functions with respect to the $\mathcal{L}^2(X,\nu)$ norm. So, continuous functions are dense in $\mathcal{L}^2(X,\nu)$. This means polynomials (which are dense in the continuous functions) are also dense in $\mathcal{L}^2(X,\nu)$ in the $\mathcal{L}^2(X,\nu)$ norm sense. Hence, orthogonal polynomials form a complete basis of $\mathcal{H} = \mathcal{L}^2(X,\nu)$. As a concrete example, letting $a = -1$ and $b = 1$ leads to a complete basis of Legendre polynomials. On the other hand, when we consider an infinite (unbounded and hence, non-compact) interval X, polynomials are no longer square integrable with respect to the Lebesgue measure as the boundedness theorem fails to hold. So, the inner product of $\mathcal{H}$ must be weighted, or more generally, the measure ν must be altered to include polynomials in the Hilbert space and permit them to form a complete basis. For example, letting $L = \mathbb{R}$ and ν be the Gaussian measure (Gaussian pdf weight with Lebesgue measure) leads to a complete basis of Hermite polynomials.

Since we have Theorem 5.2.1 at our disposal, our next affair is to determine which Gramian operator (A^*A, B^*B, or C^*C) it applies to. The operator in Theorem 5.2.1 must have the constant function as an eigenfunction. We check whether this is true for any of our Gramian operators. Let $\gamma \neq 0$ be the constant function on $\mathcal{X}$:

$$\forall x \in \mathcal{X}, \;\; \gamma(x) = \gamma \neq 0$$

with slight abuse of notation. Using Theorem 5.1.1, the Gramian operator of the channel, A^*A, applied to γ produces:

$$(A^*A)(\gamma)(x) = \gamma \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{Y|X}(y|x) P_{Y|X}(y|t) \;\; d\mu(y) \;\; d\lambda(t) \quad a.e. \text{ with respect to } \lambda$$

which means A^*A does not necessarily have a constant eigenfunction. Using Theorem 5.1.1, the Gramian operator of the DTM, B^*B, applied to γ produces:

$$(B^*B)(\gamma)(x) = \frac{\gamma}{\sqrt{P_X(x)}} \int_{\mathcal{X}} \sqrt{P_X(t)} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t) \; d\mu(y) \; d\lambda(t) \quad a.e. \text{ with respect to } \lambda$$

which means B^*B does not necessarily have a constant eigenfunction. Using Theorem 5.1.1, the Gramian operator of the LTM, C^*C, applied to γ produces:

$$\begin{aligned}
(C^*C)(\gamma)(x) &= \gamma \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{X|Y}(t|y) \; d\mu(y) \; d\lambda(t) \quad a.e. \text{ with respect to } \lambda \\
&= \gamma \int_{\mathcal{Y}} P_{X|Y}(x|y) \int_{\mathcal{X}} P_{X|Y}(t|y) \; d\lambda(t) \; d\mu(y) \quad a.e. \text{ with respect to } \lambda \\
&= \gamma \int_{\mathcal{Y}} P_{X|Y}(x|y) \; d\mu(y) \quad a.e. \text{ with respect to } \lambda
\end{aligned}$$

where the second equality uses Tonelli's theorem, as all functions involved are non-negative and measurable, and all measures involved are σ-finite. This means C^*C does not necessarily have a constant eigenfunction. Hence, none of the Gramian operators we are considering lend themselves to the application of Theorem 5.2.1. The next subsection remedies this dilemma.

5.2.2 Construction of Transformed Gramian Operator of DTM

Propelled by the importance of the DTM in the analysis of discrete and finite channels, we focus our attention on the DTM B. Recall that we are considering an infinite alphabet channel with input random variable X, output random variable Y, and conditional probability distributions $P_{Y|X}(\cdot|x), \;\; x \in \mathcal{X}$. Moreover, we are assuming that the marginal distributions of X and Y satisfy $\lambda\left(\{x \in \mathcal{X} : P_X(x) = 0\}\right) = 0$ and $\mu\left(\{y \in \mathcal{Y} : P_Y(y) = 0\}\right) = 0$, respectively, where λ and μ are typically Lebesgue or counting measures. Suppose we are given the separable Hilbert space $\mathcal{L}^2(\mathcal{X}, \lambda)$ of measurable, real, λ-square integrable functions on $\mathcal{X}$, and the separable Hilbert space $\mathcal{L}^2(\mathcal{Y}, \mu)$ of measurable, real, μ-square integrable functions on $\mathcal{Y}$. Let the associated inner product of $\mathcal{L}^2(\mathcal{X}, \lambda)$ be:

$$\forall f_1, f_2 \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; \langle f_1, f_2 \rangle_{\mathcal{X}} \triangleq \int_{\mathcal{X}} f_1 f_2 \; d\lambda \tag{5.23}$$

with induced norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; \|f\|_{\mathcal{X}} \triangleq \sqrt{\langle f, f \rangle_{\mathcal{X}}} \;, \tag{5.24}$$

and the associated inner product of $\mathcal{L}^2(\mathcal{Y}, \mu)$ be:

$$\forall g_1, g_2 \in \mathcal{L}^2(\mathcal{Y}, \mu), \;\; \langle g_1, g_2 \rangle_{\mathcal{Y}} \triangleq \int_{\mathcal{Y}} g_1 g_2 \; d\mu \tag{5.25}$$

with induced norm:

$$\forall g \in \mathcal{L}^2(\mathcal{Y}, \mu), \;\; \|g\|_{\mathcal{Y}} \triangleq \sqrt{\langle g, g \rangle_{\mathcal{Y}}} \;. \tag{5.26}$$

The DTM, $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$, is a bounded linear operator from $\mathcal{L}^2(\mathcal{X}, \lambda)$ to $\mathcal{L}^2(\mathcal{Y}, \mu)$. This requires a proof. Recall from Definition 5.1.2 that for any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$:

$$\forall y \in \mathcal{Y}, \;\; B(f)(y) = \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x) \sqrt{P_X(x)} f(x) \; d\lambda(x). \tag{5.27}$$

The next lemma uses this definition to verify the codomain and boundedness of B.

Lemma 5.2.2 (DTM is a Bounded Linear Operator)**.** *The DTM, $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$, is a bounded linear operator.*

Proof. The linearity of B follows from its definition as an integral. It suffices to prove that the operator norm of B is 1:

$$\|B\| \triangleq \sup_{f \in \mathcal{L}^2(\mathcal{X},\lambda)} \frac{\|B(f)\|_{\mathcal{Y}}}{\|f\|_{\mathcal{X}}} = 1. \tag{5.28}$$

This guarantees that $B(f) \in \mathcal{L}^2(\mathcal{Y}, \mu)$ if $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$ (which justifies why the codomain of B is $\mathcal{L}^2(\mathcal{Y}, \mu)$), and proves that B is bounded. First observe that $\sqrt{P_X} \in \mathcal{L}^2(\mathcal{X}, \lambda)$ because:

$$\left\|\sqrt{P_X}\right\|_{\mathcal{X}}^2 = \int_{\mathcal{X}} P_X \; d\lambda = 1.$$

Furthermore, we have:

$$\forall y \in \mathcal{Y}, \;\; B\left(\sqrt{P_X}\right)(y) = \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x) P_X(x) \; d\lambda(x) = \sqrt{P_Y(y)}$$

which means $B\left(\sqrt{P_X}\right) = \sqrt{P_Y} \in \mathcal{L}^2(\mathcal{Y}, \mu)$ because:

$$\left\|\sqrt{P_Y}\right\|_{\mathcal{Y}}^2 = \int_{\mathcal{Y}} P_Y \; d\mu = 1.$$

Hence, $\|B\| \geq 1$. Now fix any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$. Then, we have:

$$\begin{aligned}
\|B(f)\|_{\mathcal{Y}}^2 &= \left\| \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x) \sqrt{P_X(x)} f(x) \; d\lambda(x) \right\|_{\mathcal{Y}}^2 \\
&= \left\| \sqrt{P_Y(y)} \int_{\mathcal{X}} P_{X|Y}(x|y) \frac{f(x)}{\sqrt{P_X(x)}} \; d\lambda(x) \right\|_{\mathcal{Y}}^2 \\
&= \int_{\mathcal{Y}} P_Y(y) \left(\mathbb{E}\left[\frac{f(X)}{\sqrt{P_X(X)}} \middle| Y = y \right] \right)^2 \; d\mu(y) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{f(X)}{\sqrt{P_X(X)}} \middle| Y \right]^2 \right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\frac{f^2(X)}{P_X(X)} \middle| Y \right] \right] \\
&= \mathbb{E}\left[\frac{f^2(X)}{P_X(X)} \right] \\
&= \int_{\mathcal{X}} f^2 \; d\lambda
\end{aligned}$$

where the second equality holds by Bayes' rule, the inequality follows from the Cauchy-Schwarz inequality, and the second to last equality holds by the tower property. This gives us:

$$\|B(f)\|_{\mathcal{Y}}^2 \leq \|f\|_{\mathcal{X}}^2 .$$

Since we have already derived $\|B\| \geq 1$, this proves that $\|B\| = 1$. □

We now derive the Gramian operator of B (already stated in Theorem 5.1.1) by following the proof of Theorem 5.1.1. To find $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$, we first find the unique adjoint operator, $B^* : \mathcal{L}^2(\mathcal{Y}, \mu) \to \mathcal{L}^2(\mathcal{X}, \lambda)$, where the uniqueness of the adjoint is guaranteed by Riesz' representation theorem as B is bounded. By definition of the adjoint operator, for any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$ and any $g \in \mathcal{L}^2(\mathcal{Y}, \mu)$, we have:

$$\begin{aligned}
\langle B(f), g\rangle_{\mathcal{Y}} &= \langle f, B^*(g)\rangle_{\mathcal{X}} \\
\int_{\mathcal{Y}} \left(\frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x)\sqrt{P_X(x)} f(x) \; d\lambda(x) \right) g(y) \; d\mu(y) &= \int_{\mathcal{X}} f(x) \left(B^*(g)\right)(x) \; d\lambda(x) \\
\int_{\mathcal{Y}} \left(\sqrt{P_Y(y)} \int_{\mathcal{X}} P_{X|Y}(x|y) \frac{f(x)}{\sqrt{P_X(x)}} \; d\lambda(x) \right) g(y) \; d\mu(y) &= \int_{\mathcal{X}} f(x) \left(B^*(g)\right)(x) \; d\lambda(x) \\
\int_{\mathcal{X}} f(x) \left(\frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y)\sqrt{P_Y(y)} g(y) \; d\mu(y) \right) \; d\lambda(x) &= \int_{\mathcal{X}} f(x) \left(B^*(g)\right)(x) \; d\lambda(x)
\end{aligned}$$

where the second line follows from the definitions of the inner products and equation 5.27, the third line follows from Bayes' rule, and the final line follows from the Fubini-Tonelli theorem because $B(f)g$ must be μ-integrable by the Cauchy-Schwarz-Bunyakovsky inequality in $\mathcal{L}^2(\mathcal{Y}, \mu)$. Since the final equality is true for all $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$ and all $g \in \mathcal{L}^2(\mathcal{Y}, \mu)$, the unique bounded linear adjoint operator (by Riesz' representation theorem and the boundedness of B in Lemma 5.2.2) B^* is:

$$\forall x \in \mathcal{X}, \;\; B^*(g)(x) = \frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y)\sqrt{P_Y(y)} g(y) \; d\mu(y) \tag{5.29}$$

for every function $g \in \mathcal{L}^2(\mathcal{Y}, \mu)$. B^* is clearly the DTM of the reverse channel from Y to X. The Gramian operator of the DTM, $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$, is defined as:

$$\forall x \in \mathcal{X}, \;\; (B^*B)(f)(x) = \frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y) \int_{\mathcal{X}} P_{Y|X}(y|t)\sqrt{P_X(t)} f(t) \; d\lambda(t) \; d\mu(y) \tag{5.30}$$

for any input $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$. We can use the Fubini-Tonelli theorem (as in the proof of Theorem 5.1.1) at this point to get an elegant expression for B^*B as an integral operator. However, we will work with equation 5.30 for convenience.

Observe that $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ is a bounded, positive, self-adjoint operator. The boundedness stems from:

$$\|B^*B\| \leq \|B^*\| \, \|B\| = 1$$

because the operator norm of any bounded operator and its adjoint are equal, and equation 5.28 gives $\|B\| = \|B^*\| = 1$. Note that $\|B\| = \|B^*\|$ can be proven using Riesz' representation theorem in a Hilbert space setting, or by the Hahn-Banach theorem in the more general Banach space setting. The next lemma proves that B^*B has an eigenvalue of 1 with corresponding eigenvector $\sqrt{P_X} \in \mathcal{L}^2(\mathcal{X}, \lambda)$.

Lemma 5.2.3 (Largest Eigenvalue and Eigenvector of Gramian of DTM)**.** *The largest eigenvalue of $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ is $\|B^*B\| = 1$, and the corresponding eigenvector is $\sqrt{P_X} \in \mathcal{L}^2(\mathcal{X}, \lambda)$:*

$$B^*B\left(\sqrt{P_X}\right) = 1\sqrt{P_X} \;\; .$$

Proof.
First note that $\sqrt{P_X} \in \mathcal{L}^2(\mathcal{X}, \lambda)$ because $\left\|\sqrt{P_X}\right\|_{\mathcal{X}} = 1$. From equation 5.30, we have:

$$\begin{aligned}\forall x \in \mathcal{X},\ \ (B^*B)\left(\sqrt{P_X}\right)(x) &= \frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y) \int_{\mathcal{X}} P_{Y|X}(y|t) P_X(t)\ d\lambda(t)\ d\mu(y) \\ &= \frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_Y(y)\ d\mu(y) \\ &= \sqrt{P_X(x)}\end{aligned}$$

using the total probability law. This means B^*B has an eigenvalue of 1 with corresponding eigenvector $\sqrt{P_X}$. Since the operator norm of B^*B is bounded by 1, $\|B^*B\| \leq 1$, we must have that $\|B^*B\| = 1$. So, $\|B^*B\| = 1$ is the largest eigenvalue of B^*B. □

The results of Lemma 5.2.3 parallel those presented in the proof of Theorem 3.2.4 in the discrete and finite channel case. Indeed, B is derived from a Markov operator, so its largest eigenvalue should intuitively be 1 (in analogy with Perron-Frobenius theory for Markov matrices).

Since we have formally established the Gramian operator of the DTM, B^*B, we turn to transforming this Gramian operator into a self-adjoint operator which has a constant eigenfunction. For any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$, manipulating the definition of B^*B in equation 5.30 produces:

$$\forall x \in \mathcal{X},\ \ \frac{(B^*B)(f)(x)}{\sqrt{P_X(x)}} = \frac{1}{P_X(x)} \int_{\mathcal{Y}} P_{X|Y}(x|y) \int_{\mathcal{X}} P_{Y|X}(y|t) P_X(t) \frac{f(t)}{\sqrt{P_X(t)}}\ d\lambda(t)\ d\mu(y).$$

As $B^*B(f) \in \mathcal{L}^2(\mathcal{X}, \lambda)$, we have $B^*B(f) < \infty$ *a.e.* with respect to λ. This means the iterated integral on the right hand side is finite almost everywhere, which implies:

$$\int_{\mathcal{Y}} P_{X|Y}(x|y) \int_{\mathcal{X}} P_{Y|X}(y|t) P_X(t) \frac{|f(t)|}{\sqrt{P_X(t)}}\ d\lambda(t)\ d\mu(y) < \infty\ \ a.e.$$

with respect to λ. Applying the Fubini-Tonelli theorem to the iterated integral produces:

$$\frac{(B^*B)(f)(x)}{\sqrt{P_X(x)}} = \int_{\mathcal{X}} \left(\frac{1}{P_X(x)} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t)\ d\mu(y) \right) \frac{f(t)}{\sqrt{P_X(t)}}\ P_X(t) d\lambda(t)\ \ a.e. \tag{5.31}$$

where the equality holds almost everywhere with respect to λ. Prompted by this equation, consider the new separable Hilbert space $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ of measurable, real, $\mathbb{P}_X$-square integrable functions on $\mathcal{X}$ with associated inner product:

$$\forall f, g \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X),\ \ \langle f, g \rangle_{P_X} \triangleq \int_{\mathcal{X}} fg\ d\mathbb{P}_X = \int_{\mathcal{X}} fg P_X\ d\lambda \tag{5.32}$$

and induced norm:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X),\ \ \|f\|_{P_X} \triangleq \sqrt{\langle f, f \rangle_{P_X}}\ . \tag{5.33}$$

We now define a transformed Gramian operator of the DTM on this new Hilbert space based on equation 5.31. The ensuing lemma relates this new operator to the Gramian of the DTM.

Definition 5.2.3 (Transformed Gramian Operator of DTM)**.** The transformed Gramian operator of the DTM (TGDTM), $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, is defined for any function $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ by:

$$\forall x \in \mathcal{X}, \;\; D(f)(x) \triangleq \int_{\mathcal{X}} \Lambda_D(x,t) f(t) \;\; d\mathbb{P}_X(t)$$

where the kernel, $\Lambda_D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, of the integral operator is:

$$\Lambda_D(x,t) = \frac{1}{P_X(x)} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t) \;\; d\mu(y).$$

Lemma 5.2.4 (Isomorphism between Hilbert Spaces)**.** *The separable Hilbert spaces $\mathcal{L}^2(\mathcal{X}, \lambda)$ and $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ are isometrically isomorphic, and the isomorphism between them is given by the map $T : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ which is defined as:*

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; T(f) = \frac{f}{\sqrt{P_X}}.$$

Furthermore, $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ and $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ are equivalent operators on the isomorphic spaces in the sense that:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; T(B^*B(f)) = D(T(f)) \;\; a.e.$$

with respect to $\mathbb{P}_X$.

Proof.
First observe that $f \in \mathcal{L}^2(\mathcal{X}, \lambda) \Leftrightarrow T\left(\sqrt{P_X}\right) \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ because:

$$\|f\|_{\mathcal{X}} = \|T(f)\|_{P_X} = \left\| \frac{f}{\sqrt{P_X}} \right\|_{P_X}$$

which means T is a well-defined map between $\mathcal{L}^2(\mathcal{X}, \lambda)$ and $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$. T is clearly linear and bijective. Moreover, we have:

$$\forall f, g \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; \langle f, g \rangle_{\mathcal{X}} = \langle T(f), T(g) \rangle_{P_X}$$

from the definitions of the inner products. Hence, T is an isometric isomorphism between $\mathcal{L}^2(\mathcal{X}, \lambda)$ and $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, and the two spaces are isometrically isomorphic to each other by definition. Note that this is expected since any two infinite-dimensional separable Hilbert spaces are isometrically isomorphic to each other. Finally, observe that:

$$\forall f \in \mathcal{L}^2(\mathcal{X}, \lambda), \;\; T(B^*B(f)) = \frac{B^*B(f)}{\sqrt{P_X}} \overset{a.e.}{=} D\left(\frac{f}{\sqrt{P_X}}\right) = D(T(f))$$

where the second equality holds almost everywhere with respect to $\mathbb{P}_X$ by equation 5.31 (and the absolute continuity of $\mathbb{P}_X$ with respect to λ). Note that T guarantees that the codomain of D is indeed $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ since the codomain of B^*B is $\mathcal{L}^2(\mathcal{X}, \lambda)$. □

For readers unfamiliar with the concept of isomorphisms between Hilbert spaces, we remark that Lemma 5.2.4 can be interpreted as saying that B^*B and D are similarity transformations of each other (analogous to the matrix sense). This lemma will allow us to translate spectral decomposition results proven for D to results regarding B^*B. We also note that any vector in $\mathcal{L}^2(\mathcal{X}, \lambda)$ that is orthogonal to $\sqrt{P_X}$ can be considered a normalized perturbation, because $\mathcal{L}^2(\mathcal{X}, \lambda)$ is the domain of the DTM. From equations 5.3 and 5.4, it is evident that T takes normalized perturbations to log-likelihood perturbations. Indeed, any vector in $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ that is orthogonal to the constant function can considered a log-likelihood perturbation. We next present some properties of the TGDTM which follow from Lemma 5.2.4.

Corollary 5.2.5 (Properties of TGDTM)**.** *The TGDTM, $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, is a positive, self-adjoint, bounded, linear operator with operator norm $\|D\| = 1$. Moreover, the largest eigenvalue of D is 1 with corresponding eigenvector $\gamma \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, which is a constant function such that $\forall x \in \mathcal{X},\ \gamma(x) = \gamma \neq 0$:*

$$D(\gamma) = \gamma.$$

Proof.
The linearity of D is obvious. Since B^*B is a positive, self-adjoint, bounded, linear operator with operator norm $\|B\| = 1$, the isomorphism T from Lemma 5.2.4 easily implies that D is also a positive, self-adjoint, bounded, linear operator with operator norm $\|D\| = 1$. That D has eigenvalue 1 with corresponding eigenvector γ also follows from the isomorphism T and Lemma 5.2.3. We elaborate this part of the proof to provide an example of how T is used. From Lemma 5.2.4, we have:

$$\begin{aligned} T\left(B^*B\left(\sqrt{P_X}\right)\right) &= D\left(T\left(\sqrt{P_X}\right)\right) \\ T\left(\sqrt{P_X}\right) &= D\left(T\left(\sqrt{P_X}\right)\right) \\ 1 &= D(1) \end{aligned}$$

where the second line follows from $B^*B\left(\sqrt{P_X}\right) = \sqrt{P_X}$ in Lemma 5.2.3, and 1 denotes the constant function with value 1 for all $x \in \mathcal{X}$ in the final line. This proves the eigenvector equation in the statement of the corollary. Note that $\gamma \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ because $\|\gamma\|^2_{P_X} = \gamma^2 < \infty$. □

That D is self-adjoint can also be inferred from the symmetry of its kernel. Indeed, observe using Bayes' rule that:

$$\begin{aligned} \forall x, t \in \mathcal{X},\ \Lambda_D(x,t) &= \frac{1}{P_X(x)} \int_{\mathcal{Y}} P_{X|Y}(x|y) P_{Y|X}(y|t)\ d\mu(y) \\ &= \frac{1}{P_X(x)} \int_{\mathcal{Y}} \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} \frac{P_{X|Y}(t|y)P_Y(y)}{P_X(t)}\ d\mu(y) \\ &= \frac{1}{P_X(t)} \int_{\mathcal{Y}} P_{X|Y}(t|y) P_{Y|X}(y|x)\ d\mu(y) \\ &= \Lambda_D(t,x). \end{aligned} \tag{5.34}$$

Since the transformed Gramian operator of the DTM, $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, has an eigenvector that is the constant function, it retains the possibility of being closed over polynomials and degree preserving. So, we can apply Theorem 5.2.1 to it with additional assumptions.

5.2.3 Singular Value Decomposition of Divergence Transition Map

This subsection finds the singular value decomposition of the DTM B by first finding the spectral decomposition of D. In order to apply Theorem 5.2.1 to D, we will assume that $\mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ has a countable complete orthonormal basis of polynomials that is unique up to sign changes, and D is a compact operator. We will denote the orthonormal basis of polynomials as $P = \{p_0, p_1, p_2, \dots\} \subseteq \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ where p_k is the orthonormal polynomial with degree k. As mentioned in subsection 5.2.1, the first assumption implies that:

$$\mathbb{E}\left[|X|^n\right] = \int_{\mathcal{X}} |x|^n \ d\mathbb{P}_X(x) < \infty \tag{5.35}$$

for every $n \in \mathbb{N}$. So, we only consider channels with input random variable X such that all moments of X exist. A sufficient condition for all moments existing is if the moment generating function of X is finite on any open interval that contains 0. On the other hand, a sufficient condition which ensures D is compact is given in the next lemma. This condition can be found in [31] as well, where analysis on conditional expectation operators is performed using weak convergence arguments.

Lemma 5.2.6 (Hilbert-Schmidt Condition for Compactness)**.** *If the kernel of $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ is square integrable:*

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

then $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$, its Gramian $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$, and its transformed Gramian $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ are all compact.

Proof.
Notice that we only need to find a sufficient condition for the compactness of B. If B is compact, then its adjoint B^* is also compact by Schauder's theorem [29]; a simple proof of this theorem for endomorphisms on Hilbert spaces can be found in [28]. So, the Gramian B^*B is also compact, because the composition of compact operators is compact. This means D is compact using Lemma 5.2.4. We now derive the sufficient condition on the compactness of B. Recall from equation 5.27 that for any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$:

$$\forall y \in \mathcal{Y}, \ \ B(f)(y) = \int_{\mathcal{X}} \frac{P_{Y|X}(y|x)\sqrt{P_X(x)}}{\sqrt{P_Y(y)}} f(x) \ d\lambda(x).$$

Such integral operators are called Hilbert-Schmidt operators when their kernel is square integrable with respect to the product measure of the input and output measures. So, B is a Hilbert-Schmidt operator if:

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \left| \frac{P_{Y|X}(y|x)\sqrt{P_X(x)}}{\sqrt{P_Y(y)}} \right|^2 \ d\mu(y) \ d\lambda(x) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{P_{X,Y}^2(x,y)}{P_X(x) P_Y(y)} \ d\mu(y) \ d\lambda(x) < \infty.$$

Since Hilbert-Schmidt operators are compact [28], the above condition implies that B is compact. □

In section 5.3, we will assume that B is compact in each example without explicitly proving it. These assumptions can be proven using notions like Lemma 5.2.6. We now return to finding the spectral decomposition of D. The next lemma specifies a criterion using conditional moments which ensures D is closed over polynomials and degree preserving, and hence its orthonormal eigenbasis is P by Theorem 5.2.1. To present this lemma, we define $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ as the separable Hilbert space of measurable, real, $\mathbb{P}_Y$-square integrable functions on $\mathcal{Y}$ with associated inner product:

$$\forall f, g \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y), \ \langle f, g \rangle_{P_Y} \triangleq \int_{\mathcal{Y}} fg \ d\mathbb{P}_Y = \int_{\mathcal{Y}} fgP_Y \ d\mu \tag{5.36}$$

with induced norm:

$$\forall f \in \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y), \ \|f\|_{P_Y} \triangleq \sqrt{\langle f, f \rangle_{P_Y}} \ . \tag{5.37}$$

We also assume that $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ has a countable complete orthonormal basis of polynomials that is unique up to sign changes. We denote this basis as $Q = \{q_0, q_1, q_2, \dots\} \subseteq \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, where q_k is the orthonormal polynomial with degree k. Now note that the conditional expectation operators:

$$\mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \ \text{ and } \ \mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$$

are bounded linear operators. The linearity of these operators trivially holds, and the boundedness can be justified by the Cauchy-Schwarz inequality and the tower property. For example, for any $\forall f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$:

$$\|\mathbb{E}[f(X)|Y]\|_{P_Y}^2 = \mathbb{E}\left[\mathbb{E}[f(X)|Y]^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[f^2(X)|Y\right]\right] = \mathbb{E}\left[f^2(X)\right] = \|f\|_{P_X}^2$$

which shows that $\mathbb{E}[\cdot|Y]$ is bounded. We now present the conditions on the conditional moments which ensure that D has a spectral decomposition with polynomial eigenvectors.

Theorem 5.2.7 (Conditional Moment Conditions for Spectral Decomposition)**.** *Suppose the DTM $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ is compact. If the conditional expectation operators, $\mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $\mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, are closed over polynomials and degree preserving, then $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ has an orthonormal eigenbasis $P = \{p_0, p_1, p_2, \dots\} \subseteq \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ of orthonormal polynomials, and $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ has an orthonormal eigenbasis $\sqrt{P_X}P = \left\{\sqrt{P_X}p_0, \sqrt{P_X}p_1, \sqrt{P_X}p_2, \dots\right\} \subseteq \mathcal{L}^2(\mathcal{X}, \lambda)$:*

$$\begin{aligned} \forall k \in \mathbb{N}, \ D(p_k) &= \alpha_k p_k \\ \forall k \in \mathbb{N}, \ B^*B\left(\sqrt{P_X}p_k\right) &= \alpha_k \sqrt{P_X}p_k \end{aligned}$$

where the eigenvalues α_k are real, $\forall k \in \mathbb{N}, \ 0 \leq \alpha_k \leq 1$, and $\alpha_0 = 1$.

Proof.
Recall from Definition 5.2.3 that for any function $f \in \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, we have:

$$\begin{aligned} \forall x \in \mathcal{X}, \ D(f)(x) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{P_{X|Y}(x|y)P_{Y|X}(y|t)}{P_X(x)} f(t) \ d\mu(y) \ d\mathbb{P}_X(t) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{Y|X}(y|x)P_{X|Y}(t|y)f(t) \ d\mu(y) \ d\lambda(t) \end{aligned}$$

where the second equality holds using Bayes' rule. Using the Fubini-Tonelli theorem (whose validity was justified in the derivation of equation 5.31), we have:

$$D(f)(x) = \int_{\mathcal{Y}} P_{Y|X}(y|x) \int_{\mathcal{X}} P_{X|Y}(t|y) f(t) \; d\lambda(t) \; d\mu(y) \quad a.e.$$

where the equality holds almost everywhere with respect to λ, and hence, almost everywhere with respect to $\mathbb{P}_X$ (as $\mathbb{P}_X$ is absolutely continuous with respect to λ). Equivalently, we have:

$$D(f)(x) = \mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]|X = x\right] \quad a.e. \tag{5.38}$$

where the equality holds almost everywhere with respect to λ or $\mathbb{P}_X$. Since for every $n \in \mathbb{N}$, $\mathbb{E}\left[X^n|Y = y\right]$ is a polynomial in y with degree at most n and $\mathbb{E}\left[Y^n|X = x\right]$ is a polynomial in x with degree at most n, if f is a polynomial with degree k, then $\mathbb{E}\left[f(X)|Y\right]$ is a polynomial in Y with degree at most k and $\mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]|X = x\right]$ is a polynomial in x with degree at most k. Hence, if f is a polynomial with degree k, then $D(f)$ is a polynomial with degree at most k. This means D is closed over polynomials and degree preserving. Since B is compact, D must be compact from the proof of Lemma 5.2.6. D is also positive and self-adjoint from Corollary 5.2.5. Hence, using Theorem 5.2.1, the spectral decomposition of D has orthonormal eigenbasis P. The eigenvalues of D are all non-negative real numbers, because D is positive and self-adjoint. Moreover, Corollary 5.2.5 asserts that all the eigenvalues are bounded above by 1 since $\|D\| = 1$, and $\alpha_0 = 1$. The spectral decomposition of B^*B then follows from the isomorphism in Lemma 5.2.4. □

We clarify that Theorem 5.2.7 indeed provides conditions on the conditional moments because the conditional expectation operators are closed over polynomials and degree preserving if and only if for every $n \in \mathbb{N}$, $\mathbb{E}\left[X^n|Y = y\right]$ is a polynomial in y with degree at most n and $\mathbb{E}\left[Y^n|X = x\right]$ is a polynomial in x with degree at most n. We now use this theorem to prove the pivotal result of this chapter. The next theorem presents equivalent conditions on the conditional moments to assure the existence of the singular value decomposition (SVD) of the DTM B where the singular vectors are orthogonal polynomials.

Theorem 5.2.8 (Singular Value Decomposition of Divergence Transition Map)**.** *Suppose the DTM of the channel $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ is compact. Then, the conditional expectation operators, $\mathbb{E}\left[\cdot|Y\right] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $\mathbb{E}\left[\cdot|X\right] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, are closed over polynomials and strictly degree preserving if and only if the DTM of the channel $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ has right singular vector basis $\sqrt{P_X}P = \left\{\sqrt{P_X}p_0, \sqrt{P_X}p_1, \sqrt{P_X}p_2, \dots\right\} \subseteq \mathcal{L}^2(\mathcal{X}, \lambda)$ and left singular vector basis $\sqrt{P_Y}Q = \left\{\sqrt{P_Y}q_0, \sqrt{P_Y}q_1, \sqrt{P_Y}q_2, \dots\right\} \subseteq \mathcal{L}^2(\mathcal{Y}, \mu)$, and the DTM of the reverse channel $B^* : \mathcal{L}^2(\mathcal{Y}, \mu) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ has right singular vector basis $\sqrt{P_Y}Q$ and left singular vector basis $\sqrt{P_X}P$:*

$$\begin{aligned} \forall k \in \mathbb{N}, \;\; B\left(\sqrt{P_X}p_k\right) &= \sqrt{\alpha_k}\sqrt{P_Y}q_k \\ \forall k \in \mathbb{N}, \;\; B^*\left(\sqrt{P_Y}q_k\right) &= \sqrt{\alpha_k}\sqrt{P_X}p_k \end{aligned}$$

where the singular values $\sqrt{\alpha_k}$ satisfy $\forall k \in \mathbb{N}, \; 0 < \sqrt{\alpha_k} \leq 1$ and $\sqrt{\alpha_0} = 1$, and α_k are the eigenvalues of $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$.

Proof.
Recall from Definition 5.1.2 that for any $f \in \mathcal{L}^2(\mathcal{X}, \lambda)$:

$$\forall y \in \mathcal{Y},\ \ B(f)(y) = \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x)\sqrt{P_X(x)} f(x)\ \ d\lambda(x)$$

and from equation 5.29 that for any $g \in \mathcal{L}^2(\mathcal{Y}, \mu)$:

$$\forall x \in \mathcal{X},\ \ B^*(g)(x) = \frac{1}{\sqrt{P_X(x)}} \int_{\mathcal{Y}} P_{X|Y}(x|y)\sqrt{P_Y(y)} g(y)\ \ d\mu(y).$$

So, B^* is indeed the DTM of the reverse channel from Y to X as we noted earlier.

We first prove the forward direction. If the conditional expectation operators are closed over polynomials and strictly degree preserving, then Theorem 5.2.7 guarantees that B^*B has orthonormal eigenbasis $\sqrt{P_X}P$:

$$\forall k \in \mathbb{N},\ \ B^*B\left(\sqrt{P_X}p_k\right) = \alpha_k \sqrt{P_X}p_k$$

where the α_k are real, $\forall k \in \mathbb{N},\ 0 < \alpha_k \leq 1$, and $\alpha_0 = 1$. Note that the strict degree preservation of the conditional expectation operators ensures that α_k are strictly positive. This is because $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ becomes strictly degree preserving by equation 5.38, which means it cannot map any polynomial to 0. As is typical in the matrix case, the orthonormal eigenbasis of the Gramian is the right (input) singular vector basis of the original map. Hence, we input these vectors into B and analyze the outputs. For any $k \in \mathbb{N}$, we have:

$$\begin{aligned}
\forall y \in \mathcal{Y},\ \ B\left(\sqrt{P_X}p_k\right)(y) &= \frac{1}{\sqrt{P_Y(y)}} \int_{\mathcal{X}} P_{Y|X}(y|x) P_X(x) p_k(x)\ \ d\lambda(x) \\
&= \sqrt{P_Y(y)} \int_{\mathcal{X}} P_{X|Y}(x|y) p_k(x)\ \ d\lambda(x) \\
&= \sqrt{P_Y(y)}\mathbb{E}\left[p_k(X)|Y = y\right]
\end{aligned}$$

where the second equality follows from Bayes' rule, and $\mathbb{E}\left[p_k(X)|Y = y\right]$ must be a polynomial in y with degree k. Moreover, for any $j, k \in \mathbb{N}$, we have:

$$\left\langle B\left(\sqrt{P_X}p_j\right), B\left(\sqrt{P_X}p_k\right)\right\rangle_{\mathcal{Y}} = \left\langle \sqrt{P_X}p_j, B^*B\left(\sqrt{P_X}p_k\right)\right\rangle_{\mathcal{Y}} = \alpha_k \left\langle \sqrt{P_X}p_j, \sqrt{P_X}p_k\right\rangle_{\mathcal{Y}} = \alpha_k \delta_{jk}$$

where δ_{jk} is the Kronecker delta (which equals 1 if $j = k$ and 0 otherwise), the first equality follows from the definition of adjoints, and the second equality uses Theorem 5.2.7. This means that for any $j, k \in \mathbb{N}$:

$$\left\langle B\left(\sqrt{P_X}p_j\right), B\left(\sqrt{P_X}p_k\right)\right\rangle_{\mathcal{Y}} = \int_{\mathcal{Y}} \mathbb{E}\left[p_j(X)|Y = y\right]\mathbb{E}\left[p_k(X)|Y = y\right]\ \ d\mathbb{P}_Y(y) = \alpha_k \delta_{jk}.$$

Hence, $\forall k \in \mathbb{N},\ \mathbb{E}\left[p_k(X)|Y = y\right] = \sqrt{\alpha_k} q_k$ are orthogonal polynomials with respect to $\mathbb{P}_Y$, where $q_k \in Q$ is the orthonormal polynomial with degree k. (Note that the sign of each q_k is chosen to keep $\sqrt{\alpha_k} > 0$.) This is essentially the SVD of the conditional expectation operator. So, we must have:

$$\forall k \in \mathbb{N},\ \ B\left(\sqrt{P_X}p_k\right) = \sqrt{\alpha_k}\sqrt{P_Y}q_k$$

such that the right singular vector basis of B is $\sqrt{P_X}P$ and the left singular vector basis is $\sqrt{P_Y}Q$. Now observe that the entire analysis of this chapter can be performed mutatis mutandis for the reverse channel from Y to X which has DTM B^*, where B^* is compact because B is compact (by Schauder's theorem). So, we must also have:

$$\forall k \in \mathbb{N}, \;\; B^*\left(\sqrt{P_Y}q_k\right) = \sqrt{\alpha_k}\sqrt{P_X}p_k$$

such that the right singular vector basis of B^* is $\sqrt{P_Y}Q$ and the left singular vector basis is $\sqrt{P_X}P$. The form of the singular vectors of B^* follows from arguments similar to those given for B, while the equivalence of the singular values is well-known in linear algebra. Indeed, if $B\left(\sqrt{P_X}p_k\right) = \sqrt{\alpha_k}\sqrt{P_Y}q_k$ and $B^*\left(\sqrt{P_Y}q_k\right) = \beta\sqrt{P_X}p_k$ with $\sqrt{\alpha_k} \neq \beta \in \mathbb{R}$, then by the definition of adjoints:

$$\begin{aligned}
\left\langle B\left(\sqrt{P_X}p_k\right), \sqrt{P_Y}q_k\right\rangle_{\mathcal{Y}} &= \left\langle \sqrt{P_X}p_k, B^*\left(\sqrt{P_Y}q_k\right)\right\rangle_{\mathcal{X}} \\
\sqrt{\alpha_k}\left\langle \sqrt{P_Y}q_k, \sqrt{P_Y}q_k\right\rangle_{\mathcal{Y}} &= \beta\left\langle \sqrt{P_X}p_k, \sqrt{P_X}p_k\right\rangle_{\mathcal{X}} \\
\sqrt{\alpha_k} &= \beta
\end{aligned}$$

and we get a contradiction. This justifies the equivalence of the singular values, and completes the proof of the forward direction.

To prove the converse direction, we assume that:

$$\begin{aligned}
\forall k \in \mathbb{N}, \;\; B\left(\sqrt{P_X}p_k\right) &= \sqrt{\alpha_k}\sqrt{P_Y}q_k \\
\forall k \in \mathbb{N}, \;\; B^*\left(\sqrt{P_Y}q_k\right) &= \sqrt{\alpha_k}\sqrt{P_X}p_k
\end{aligned}$$

where the singular values $\sqrt{\alpha_k}$ satisfy $\forall k \in \mathbb{N}, \; 0 < \sqrt{\alpha_k} \leq 1$ and $\sqrt{\alpha_0} = 1$. From our earlier derivation, for any polynomial p with degree n, we have:

$$\begin{aligned}
\forall y \in \mathcal{Y}, \;\; B\left(\sqrt{P_X}p\right)(y) &= \frac{1}{\sqrt{P_Y(y)}}\int_{\mathcal{X}} P_{Y|X}(y|x)P_X(x)p(x) \; d\lambda(x) \\
&= \sqrt{P_Y(y)}\mathbb{E}\left[p(X)|Y=y\right]
\end{aligned}$$

which means $\mathbb{E}\left[p(X)|Y=y\right]$ must be a polynomial in y with degree n. Hence, $\mathbb{E}\left[\cdot|Y\right]$ is closed over polynomials and strictly degree preserving. An analogous argument using B^* conveys that $\mathbb{E}\left[\cdot|X\right]$ is also closed over polynomials and strictly degree preserving. This completes the proof. □

The careful reader should observe that the spectral decomposition (eigendecomposition) results only require degree preservation of conditional expectation operators, while the the singular value decomposition results require strict degree preservation. We now briefly recapitulate our scheme for proving the polynomial SVD of the DTM. It closely parallels the proof of the SVD in the matrix case; this is particularly true in the proof of Theorem 5.2.8. We first define appropriate Hilbert spaces and the bounded linear operator B (the DTM) on these spaces. Then, we find conditions which ensure the spectral decomposition of B^*B has an orthonormal eigenbasis of polynomials

(weighted by the square root of the marginal density). Finally, we use the spectral decomposition results of B^*B to prove similar results about the SVD of B.

There are several noteworthy features of Theorem 5.2.8 which are worth commenting on. Firstly, although Theorem 5.2.8 explicitly characterizes the singular vectors of B, it does not explicitly calculate the singular values. The singular values can be computed on a case by case basis by inputting a unit norm right singular vector into B and verifying how much the corresponding unit norm left singular vector has been scaled. More generally, it can be justified by the spectral theorem for compact self-adjoint operators that $\alpha_k \to 0$ as $k \to \infty$, but we will not explicitly prove this. Secondly, we can identify the space of valid normalized perturbations from the SVD of B; this is why we pursued the SVD in the first place. Recall from equations 5.5 and 5.14 that valid normalized input perturbations $K_X \in \mathcal{L}^2(\mathcal{X}, \lambda)$ satisfy:

$$\int_{\mathcal{X}} \sqrt{P_X} K_X \ d\lambda = 0$$

and valid normalized output perturbations $K_Y \in \mathcal{L}^2(\mathcal{Y}, \mu)$ satisfy:

$$\int_{\mathcal{Y}} \sqrt{P_Y} K_Y \ d\mu = 0.$$

This means the orthonormal basis of valid normalized input perturbations is:

$$\left\{ \sqrt{P_X} p_1, \sqrt{P_X} p_2, \sqrt{P_X} p_3, \dots \right\} \subseteq \mathcal{L}^2(\mathcal{X}, \lambda) \tag{5.39}$$

and the orthonormal basis of valid normalized output perturbations is:

$$\left\{ \sqrt{P_Y} q_1, \sqrt{P_Y} q_2, \sqrt{P_Y} q_3, \dots \right\} \subseteq \mathcal{L}^2(\mathcal{Y}, \mu). \tag{5.40}$$

Finally, we remark that the proof of Theorem 5.2.8 also computes the SVD of the conditional expectation operators, $\mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $\mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$. Definition 5.1.2 states that the LTM C is precisely the conditional expectation operator $\mathbb{E}[\cdot|Y]$. Hence, the SVD of $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is:

$$\forall k \in \mathbb{N}, \ \ C(p_k) = \sqrt{\alpha_k} q_k. \tag{5.41}$$

and the orthonormal polynomials $P \backslash \{p_0\}$ and $Q \backslash \{q_0\}$ are the bases for the input and output log-likelihood perturbations, respectively. This might appear to be puzzling. In our previous calculations, the Gramian operator of C (in Theorem 5.1.1) did not necessarily have a constant eigenfunction. This discrepancy can be clarified by realizing that Theorem 5.1.1 computed the Gramian operator of $C : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$, whereas we have now found the SVD of $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$. Indeed, P and Q are not orthonormal sets in $\mathcal{L}^2(\mathcal{X}, \lambda)$ and $\mathcal{L}^2(\mathcal{Y}, \mu)$, respectively; they may not even belong in these Hilbert spaces. Furthermore, it is straightforward to show that $\mathbb{E}[\cdot|Y]$ and $\mathbb{E}[\cdot|X]$ are adjoints of each other, and therefore, the adjoint of $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is $\mathbb{E}[\cdot|X]$. So, the Gramian operator of $C : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is precisely $D : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ from equation 5.38 (which is not the same as $C^*C : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ calculated in Theorem 5.1.1).

Theorem 5.2.8 was proven for infinite alphabet channels. However, the result holds mutatis mutandis when X or Y is discrete and finite. For the sake of completeness, we present the result for the case where $\mathcal{X}$ is an infinite set as before, but $|\mathcal{Y}| = m < \infty$ for some $m \in \mathbb{Z}^+$ i.e. Y is a discrete and finite random variable. The Hilbert spaces and operators we defined all remain the same. However, we note that $\mathcal{L}^2(\mathcal{Y}, \mu)$ and $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ become finite dimensional with dimension m. Thus, the unique orthonormal basis of polynomials for $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ becomes $Q = \{q_0, \ldots, q_{m-1}\} \subseteq \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$, because we are only specifying m points, which means we only require up to degree $m-1$ polynomials to complete the space $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$. We also note that the DTM $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ is now a finite rank operator because its range is finite dimensional. Hence, B must be a compact operator and we do not need to additionally assume compactness anymore. The next theorem states the analogous SVD result to Theorem 5.2.8.

Theorem 5.2.9 (SVD of DTM in Discrete and Finite Case)**.** *Suppose we have that $|\mathcal{Y}| = m \in \mathbb{Z}^+$ and Y is a discrete and finite random variable. Then, the conditional expectation operators, $\mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $\mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$, are closed over polynomials and strictly degree preserving if and only if the DTM of the channel $B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{Y}, \mu)$ has right singular vector basis $\sqrt{P_X}P = \left\{\sqrt{P_X}p_0, \sqrt{P_X}p_1, \sqrt{P_X}p_2, \ldots\right\} \subseteq \mathcal{L}^2(\mathcal{X}, \lambda)$ and left singular vector basis $\sqrt{P_Y}Q = \left\{\sqrt{P_Y}q_0, \ldots, \sqrt{P_Y}q_{m-1}\right\} \subseteq \mathcal{L}^2(\mathcal{Y}, \mu)$, and the DTM of the reverse channel $B^* : \mathcal{L}^2(\mathcal{Y}, \mu) \to \mathcal{L}^2(\mathcal{X}, \lambda)$ has right singular vector basis $\sqrt{P_Y}Q$ and left singular vector basis $\left\{\sqrt{P_X}p_0, \ldots, \sqrt{P_X}p_{m-1}\right\}$:*

$$\begin{aligned}
\forall k \in \{0, \ldots, m-1\}, \ B\left(\sqrt{P_X}p_k\right) &= \sqrt{\alpha_k}\sqrt{P_Y}q_k \\
\forall k \in \mathbb{N}\backslash\{0, \ldots, m-1\}, \ B\left(\sqrt{P_X}p_k\right) &= 0 \\
\forall k \in \{0, \ldots, m-1\}, \ B^*\left(\sqrt{P_Y}q_k\right) &= \sqrt{\alpha_k}\sqrt{P_X}p_k
\end{aligned}$$

where the singular values $\sqrt{\alpha_k}$ satisfy $\forall k \in \{0, \ldots, m-1\}, \ 0 < \sqrt{\alpha_k} \leq 1$ and $\sqrt{\alpha_0} = 1$, and α_k are the eigenvalues of $B^*B : \mathcal{L}^2(\mathcal{X}, \lambda) \to \mathcal{L}^2(\mathcal{X}, \lambda)$.

Proof.
First note that the interpretation of strict degree preservation slightly changes as $\mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ is finite dimensional. $\mathbb{E}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X) \to \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y)$ and $\mathbb{E}[\cdot|X] : \mathcal{L}^2(\mathcal{Y}, \mathbb{P}_Y) \to \mathcal{L}^2(\mathcal{X}, \mathbb{P}_X)$ are closed over polynomials and strictly degree preserving if and only if for every $n \in \{0, \ldots, m-1\}$, $\mathbb{E}[X^n|Y = y]$ is a polynomial in y with degree n and $\mathbb{E}[Y^n|X = x]$ is a polynomial in x with degree n, and for every $n \in \mathbb{N}\backslash\{0, \ldots, m-1\}$, $\mathbb{E}[X^n|Y = y]$ is a polynomial in y with degree at most n and $\mathbb{E}[Y^n|X = x]$ is a polynomial in x with degree at most n.

We now prove the forward direction. Theorem 5.2.7 turns out to hold without any changes (as it only requires degree preservation). Hence, if the conditional expectation operators are closed over polynomials and strictly degree preserving, then Theorem 5.2.7 guarantees that B^*B has orthonormal eigenbasis $\sqrt{P_X}P$:

$$\forall k \in \mathbb{N}, \ B^*B\left(\sqrt{P_X}p_k\right) = \alpha_k\sqrt{P_X}p_k$$

where the α_k are real, $\forall k \in \mathbb{N}, \ 0 \leq \alpha_k \leq 1$, and $\alpha_0 = 1$. Following the proof of Theorem 5.2.8, for any $k \in \mathbb{N}$, we have:

$$\forall y \in \mathcal{Y}, \ B\left(\sqrt{P_X}p_k\right)(y) = \sqrt{P_Y(y)}\mathbb{E}[p_k(X)|Y = y]$$

where $\mathbb{E}\left[p_k(X)|Y=y\right]$ is a polynomial in y with degree k if $k < m$, or a polynomial in y with degree at most k if $k \geq m$. Moreover, for any $j, k \in \mathbb{N}$, we get:

$$\left\langle B\left(\sqrt{P_X}p_j\right), B\left(\sqrt{P_X}p_k\right)\right\rangle_{\mathcal{Y}} = \int_{\mathcal{Y}} \mathbb{E}\left[p_j(X)|Y=y\right]\mathbb{E}\left[p_k(X)|Y=y\right] \; d\mathbb{P}_Y(y) = \alpha_k\delta_{jk}.$$

Hence, $\forall k \in \{0,\ldots,m-1\}$, $\mathbb{E}\left[p_k(X)|Y=y\right] = \sqrt{\alpha_k}q_k$ are orthogonal polynomials with respect to $\mathbb{P}_Y$, where $\forall k \in \{0,\ldots,m-1\}$, $\alpha_k > 0$. For any $k \geq m$, $\mathbb{E}\left[p_k(X)|Y=y\right]$ is orthogonal to every $q_j \in Q$ (where $j < m$), which implies that $\mathbb{E}\left[p_k(X)|Y=y\right] = 0$ as Q is an orthonormal basis of $\mathcal{L}^2\left(\mathcal{Y}, \mathbb{P}_Y\right)$. This also implies that $\alpha_k = 0$ for $k \geq m$. So, we must have:

$$\begin{aligned} \forall k \in \{0,\ldots,m-1\}, \; B\left(\sqrt{P_X}p_k\right) &= \sqrt{\alpha_k}\sqrt{P_Y}q_k \\ \forall k \in \mathbb{N}\backslash\{0,\ldots,m-1\}, \; B\left(\sqrt{P_X}p_k\right) &= 0 \end{aligned}$$

such that the right singular vector basis of B is $\sqrt{P_X}P$ and the left singular vector basis is $\sqrt{P_Y}Q$. In a similar manner, we can prove that:

$$\forall k \in \{0,\ldots,m-1\}, \; B^*\left(\sqrt{P_Y}q_k\right) = \sqrt{\alpha_k}\sqrt{P_X}p_k$$

such that the right singular vector basis of B^* is $\sqrt{P_Y}Q$ and the left singular vector basis is $\left\{\sqrt{P_X}p_0,\ldots,\sqrt{P_X}p_{m-1}\right\}$. This completes the proof of the forward direction. The converse direction can be proven using exactly the same approach as in Theorem 5.2.8. □

In prior work such as in [8] and [23], polynomial SVDs of DTMs are calculated on a case by case basis through a myriad of clever mathematical techniques. To circumvent this cumbersome process, Theorems 5.2.8 and 5.2.9 offer elementary conditions on the conditional moments which can be checked to deduce whether or not an SVD with orthogonal polynomial singular vectors exists. Furthermore, when such an SVD exists, the orthogonal polynomials which form the singular vectors can be readily identified. It is typically much simpler to determine if the conditional moment conditions are satisfied (through a survey of the literature or some manageable calculations) rather than directly computing the SVD of a DTM.

5.3 Exponential Families, Conjugate Priors, and their Orthogonal Polynomial SVDs

In this section, we illustrate the applications of Theorems 5.2.8 and 5.2.9 by deriving elegant orthogonal polynomial SVDs for DTMs of channels. The setup of exponential families and their conjugate priors turns out to be a convenient method of generating such examples. So, we introduce exponential families and conjugate priors in subsection 5.3.1. We then define the pertinent orthogonal polynomial families in subsection 5.3.2. The remaining subsections are devoted to computing SVDs for different channels.

5.3.1 Exponential Families and their Conjugate Priors

We first define the notion of (one-parameter) exponential families more generally than in Definition 3.4.3 in chapter 3. Since any exponential family can be transformed into canonical form, we only present the canonical exponential family below [25]. The ensuing definition can be further generalized to include families with any finite number of parameters, but we will not require this level of generality in our analysis.

Definition 5.3.1 (Canonical Exponential Family)**.** Given a measurable space $(\mathcal{Y}, \mathcal{F})$ with $\mathcal{Y} \subseteq \mathbb{R}$, and a σ-finite measure μ on this space, the parametrized family of distributions, $\{P_Y(\cdot; x) : x \in \mathcal{X}\}$, with respect to μ is called a regular canonical exponential family when the support of the distributions do not depend on x, and each distribution in the family has the form:

$$\forall y \in \mathcal{Y}, \;\; P_Y(y; x) = \exp\left[xt(y) - \alpha(x) + \beta(y)\right]$$

where the measurable function $t : (\mathcal{Y}, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is the sufficient statistic of the distribution, the non-negative function $P_Y(y; 0) = \exp\left[\beta(y)\right]$ is a valid distribution with respect to μ known as the base distribution, and:

$$\forall x \in \mathcal{X}, \;\; \exp\left[\alpha(x)\right] = \int_{\mathcal{Y}} \exp\left[xt(y) + \beta(y)\right] \;\; d\mu(y)$$

is the partition function with $\alpha(0) = 0$ without loss of generality. The parameter x is called the natural parameter, and it takes values from the natural parameter space $\mathcal{X} \subseteq \mathbb{R}$, defined as:

$$\mathcal{X} \triangleq \{x \in \mathbb{R} : \alpha(x) < \infty\}$$

which ensures that $P_Y(\cdot; x)$ is a valid distribution when $x \in \mathcal{X}$.

In Definition 5.3.1, $\mathcal{B}$ denotes the Borel σ-algebra and $\mathcal{F}$ is the appropriate σ-algebra corresponding to $\mathcal{Y}$. So, if $\mathcal{Y} = \mathbb{R}$ then $\mathcal{F} = \mathcal{B}$, and if $\mathcal{Y}$ is countable then $\mathcal{F} = 2^{\mathcal{Y}}$. Likewise, the measure μ is typically the Lebesgue measure when $\mathcal{Y} = \mathbb{R}$ and a counting measure when $\mathcal{Y}$ is countable. Definition 5.3.1 generalizes Definition 3.4.3 by incorporating larger classes of distributions like discrete pmfs into the exponential family framework. We now consider a specific class of canonical exponential families known as natural exponential families with quadratic variance functions (NEFQVF). This class of exponential families is studied in [32] where the author asserts that the wide applicability and utility of certain distributions like Gaussian, Poisson, and binomial stems from their characterization as NEFQVFs. The next definition formally presents the NEFQVF [32].

Definition 5.3.2 (Natural Exponential Family with Quadratic Variance Function)**.** Given a measurable space $(\mathcal{Y}, \mathcal{F})$ with $\mathcal{Y} \subseteq \mathbb{R}$, and a σ-finite measure μ on this space, a canonical exponential family, $\{P_Y(\cdot; x) : x \in \mathcal{X}\}$, with respect to μ is called a natural exponential family (NEF) if the sufficient statistic $t : \mathcal{Y} \to \mathbb{R}$ is the identity function $t(y) = y$:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_Y(y; x) = \exp\left[xy - \alpha(x) + \beta(y)\right].$$

For such a natural exponential family, the expected value is given by:

$$\forall x \in \mathcal{X}, \;\; \mathbb{E}\left[Y; x\right] = \int_{\mathcal{Y}} y P_Y(y; x) \;\; d\mu(y) \;\; \triangleq \;\; \gamma$$

where γ is a function of x. Let $\mathcal{M}$ be the set of values γ can take. Then, the variance function, $V : \mathcal{M} \to \mathbb{R}^+$, is defined as:

$$V(\gamma) \triangleq \mathbb{VAR}(Y; x) = \int_{\mathcal{Y}} (y - \gamma)^2 \, P_Y(y; x) \; d\mu(y)$$

where $x \in \mathcal{X}$ corresponding to γ is used in the definition. A natural exponential family is said to have a quadratic variance function (QVF) if V is a quadratic function of γ.

At first glance, the definition of the variance function seems questionable. It is not obvious that V can be reproduced only as a function of γ. However, it turns out that γ is an injective function of x [32], and this ensures that the variance function V is well-defined. The channel conditional distributions of every example we will consider in this section will be NEFQVFs. Although we will not explicitly use the NEFQVF form in our calculations, a discussion of these concepts permits the reader to develop deeper insights into how our examples were constructed. Before we present the pertinent examples, we introduce the notion of conjugate priors. Conjugate priors are attractive in Bayesian inference because they lead to computationally tractable algorithms. In Bayesian inference, the goal is to estimate some random variable X from some observation Y, which is related to X through likelihoods (conditional distributions) $P_{Y|X}$. The key insight is to associate a prior to X such that the posterior distribution of X given Y is in the same distribution family as the prior distribution. This can allow more efficient computation of the posterior distribution every time new information is observed. Such distribution families are known as conjugate priors. It is well-known in statistics that (well-behaved) exponential families have conjugate priors that are themselves exponential families. The next definition presents such conjugate priors for natural exponential families.

Definition 5.3.3 (Conjugate Prior for NEF)**.** Suppose we are given a measurable space $(\mathcal{Y}, \mathcal{F})$ with $\mathcal{Y} \subseteq \mathbb{R}$, a σ-finite measure μ on this space, and a natural exponential family, $\{P_Y(\cdot; x) : x \in \mathcal{X}\}$, with respect to μ:

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \;\; P_Y(y; x) = \exp\left[xy - \alpha(x) + \beta(y)\right].$$

Suppose further that $\mathcal{X} \subseteq \mathbb{R}$ is a non-empty open interval and $(\mathcal{X}, \mathcal{G})$ be a measurable space with σ-finite measure λ. Then, the conjugate prior family of distributions, $\{P_X(\cdot; y', n)\}$, with respect to λ which are parametrized by the hyper-parameters $y' \in \mathbb{R}$ and $n \in \mathbb{R}$ is given by:

$$\forall x \in \mathcal{X}, \;\; P_X(x; y', n) = \exp\left[y'x - n\alpha(x) + \tau(y', n)\right]$$

where $\exp\left[-\tau(y', n)\right]$ is the partition function:

$$\exp\left[-\tau(y', n)\right] = \int_{\mathcal{X}} \exp\left[y'x - n\alpha(x)\right] \;\; d\lambda(x)$$

and we only consider conjugate prior distributions where the partition function is finite.

In Definition 5.3.3, we defined conjugate priors for natural exponential families assuming $\mathcal{X}$ is a non-empty open interval. This means $\mathcal{G}$ is the Borel σ-algebra on this interval, and λ will typically be the Lebesgue measure. Conjugate priors can be defined more generally as well, but Definition

5.3.3 suffices for our purposes. The alluring property of conjugate priors is the following. Suppose we have a channel where the input X is distributed according to the conjugate prior of an NEF which defines the conditional distributions:

$$\forall x \in \mathcal{X},\ \ P_X(x) = \exp\left[y'x - n\alpha(x) + \tau(y', n)\right] = P_X(x; y', n), \tag{5.42}$$

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y},\ \ P_{Y|X}(y|x) = \exp\left[xy - \alpha(x) + \beta(y)\right] = P_Y(y; x). \tag{5.43}$$

Then, the posterior distribution of X given Y is:

$$\begin{aligned} \forall y \in \mathcal{Y}, \forall x \in \mathcal{X},\ \ P_{X|Y}(x|y) &= \exp\left[(y' + y)x - (n+1)\alpha(x) + \tau(y' + y, n+1)\right] \\ &= P_X(x; y' + y, n+1). \end{aligned} \tag{5.44}$$

This property of conjugate priors will be at the heart of our analysis. We will consider channels with input X and output Y whose conditional distributions (likelihoods), $P_{Y|X}(\cdot|x),\ \ x \in \mathcal{X}$, are NEFQVFs and input distributions (priors) of X belong to the corresponding conjugate prior families. This will mean that the posterior distributions, $P_{X|Y}(\cdot|y),\ \ y \in \mathcal{Y}$, will also belong to the conjugate prior families as shown in equation 5.44. Observe that for any $m \in \mathbb{N}$:

$$\mathbb{E}\left[Y^m | X = x\right] = \int_{\mathcal{Y}} y^m P_Y(y; x)\ \ d\mu(y)$$

using equation 5.43, and:

$$\mathbb{E}\left[X^m | Y = y\right] = \int_{\mathcal{X}} x^m P_X(x; y' + y, n+1)\ \ d\lambda(x)$$

using equation 5.44. Hence, to verify that the conditional moments are strictly degree preserving polynomials (the condition of Theorems 5.2.8 and 5.2.9), we simply need to check that the moments of the NEFQVF and its conjugate prior are strictly degree preserving polynomials of the parameters of the families. This straightforward verification procedure in terms of the parameters (or stochastic primitives) of classes of distributions is what makes the exponential family and conjugate prior structure conducive for generating illustrations of Theorems 5.2.8 and 5.2.9.

We now present the channels which will be explored in subsections 5.3.3, 5.3.4, and 5.3.5. The agenda of these subsections will largely be to establish that the NEFQVF and conjugate prior families have moments that are degree preserving polynomials of their parameters. A comprehensive list of NEFQVFs can be found in [32], and a list of several corresponding conjugate priors can be found in [33]. Many of our definitions have been derived from these sources.

Gaussian Likelihood and Gaussian Conjugate Prior

Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$, so that $\mathcal{G} = \mathcal{B}$ and $\mathcal{F} = \mathcal{B}$ are the Borel σ-algebra, and λ and μ are the Lebesgue measure. The channel conditional pdfs (likelihoods) are Gaussian distirbutions:

$$\forall x, y \in \mathbb{R},\ \ P_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) = P_Y(y; x) \tag{5.45}$$

where x is the expectation parameter of the Gaussian NEFQVF and $\nu > 0$ is some fixed variance. The conjugate prior of this NEFQVF is also Gaussian:

$$\forall x \in \mathbb{R},\ \ P_X(x) = \frac{1}{\sqrt{2\pi p}} \exp\left(-\frac{(x-r)^2}{2p}\right) = P_X(x; r, p) \tag{5.46}$$

where the hyper-parameters $r \in \mathbb{R}$ and $p \in (0, \infty)$ represent the mean and variance of X, respectively. The posterior distribution $P_{X|Y}$ and the output marginal distribution P_Y are also Gaussian. If we let $r = 0$ to begin with, then equations 5.45 and 5.46 resemble an AWGN channel (Definition 3.4.1) with capacity achieving input distribution. Subsection 5.3.3 will prove the conditional moment conditions for the AWGN channel to re-derive the SVD results of [23]. Note that the conjugate prior in equation 5.46 is not in the form of Definition 5.3.3. It is easy to write the prior in this form, but we will not require this because subsection 5.3.3 will prove the conditional moment conditions using the alternative notion of translation invariant kernels.

Poisson Likelihood and Gamma Conjugate Prior

Let $\mathcal{X} = (0, \infty)$ and $\mathcal{Y} = \mathbb{N} = \{0, 1, 2, \dots\}$, so that $\mathcal{G} = \mathcal{B}$ is the Borel σ-algebra and $\mathcal{F} = 2^{\mathcal{Y}}$, and λ is the Lebesgue measure and μ is the counting measure. The channel conditional pdfs (likelihoods) are Poisson distributions:

$$\forall x > 0, \forall y \in \mathbb{N},\ \ P_{Y|X}(y|x) = \frac{x^y e^{-x}}{y!} = P_Y(y; x) \tag{5.47}$$

where x is the rate parameter of the Poisson NEFQVF. The conjugate prior of this NEFQVF is the gamma distribution:

$$\forall x > 0,\ \ P_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} = P_X(x; \alpha, \beta) \tag{5.48}$$

where the hyper-parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$ represent the shape and rate of the distribution respectively, and the gamma function, $\Gamma : \{z \in \mathbb{C} : \Re(z) > 0\} \to \mathbb{C}$, which is a well-known generalization of the factorial function, is defined as:

$$\Gamma(z) \triangleq \int_0^\infty x^{z-1} e^{-x}\ dx. \tag{5.49}$$

Moreover, for any $m \in \mathbb{Z}^+,\ \Gamma(m) = (m-1)!$. The posterior distribution of X given Y is also gamma:

$$\forall y \in \mathbb{N}, \forall x > 0,\ \ P_{X|Y}(x|y) = P_X(x; \alpha + y, \beta + 1). \tag{5.50}$$

When α is a positive integer, the gamma distribution becomes the Erlang distribution. Hence, the gamma distribution generalizes the Erlang distribution (and also the exponential distribution) to non-integer α values. We next introduce the negative binomial distribution with parameters $p \in (0, 1)$ and $s \in (0, \infty)$ representing the success probability and number of failures, respectively:

$$\forall y \in \mathbb{N},\ \ P(y) = \frac{\Gamma(s+y)}{\Gamma(s)y!} p^y (1-p)^s. \tag{5.51}$$

When s is a positive integer, the negative binomial distribution is essentially the sum of s geometric distributions. It models the probability distribution of the number of successes in a Bernoulli process

with s failures. In fact, when $s \in \mathbb{Z}^+$, the coefficient with gamma functions in equation 5.51 can be written as a binomial coefficient:

$$\frac{\Gamma(s+y)}{\Gamma(s)y!} = \binom{s+y-1}{y}.$$

The marginal distribution of Y for a Poisson channel with gamma input is negative binomial with parameters $p = \frac{1}{\beta+1}$ and $s = \alpha$:

$$\forall y \in \mathbb{N},\ \ P_Y(y) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{1}{\beta+1}\right)^y\left(\frac{\beta}{\beta+1}\right)^\alpha. \tag{5.52}$$

Subsection 5.3.4 proves that the conditional moments of the Poisson and gamma distributions are polynomials in the their parameters in order to derive the SVD of the DTM of the Poisson channel.

Binomial Likelihood and Beta Conjugate Prior

Let $\mathcal{X} = (0,1)$ and $\mathcal{Y} = \{0,\dots,n\}$ for some fixed $n \in \mathbb{Z}^+$, so that $\mathcal{G} = \mathcal{B}$ is the Borel σ-algebra and $\mathcal{F} = 2^{\mathcal{Y}}$, and λ is the Lebesgue measure and μ is the counting measure. The channel conditional pdfs (likelihoods) are binomial distributions:

$$\forall x \in (0,1), \forall y \in \{0,\dots,n\},\ \ P_{Y|X}(y|x) = \binom{n}{y}x^y(1-x)^{n-y} = P_Y(y;x) \tag{5.53}$$

where x is the success probability parameter of the binomial NEFQVF. The conjugate prior of this NEFQVF is the beta distribution:

$$\forall x \in (0,1),\ \ P_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} = P_X(x;\alpha,\beta) \tag{5.54}$$

where the hyper-parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$ are shape parameters, and the beta function, $\mathrm{B} : \left\{(z_1,z_2) \in \mathbb{C}^2 : \Re(z_1) > 0, \Re(z_2) > 0\right\} \to \mathbb{C}$, is defined as:

$$\mathrm{B}(z_1,z_2) \triangleq \int_0^1 x^{z_1-1}(1-x)^{z_2-1}\ dx = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}. \tag{5.55}$$

We note that when $\alpha = \beta = 1$, the beta distribution becomes the uniform distribution. The posterior distribution of X given Y is also beta:

$$\forall y \in \{0,\dots,n\}, \forall x \in (0,1),\ \ P_{X|Y}(x|y) = P_X(x;\alpha+y,\beta+n-y). \tag{5.56}$$

The marginal distribution of Y is called the beta-binomial distribution:

$$\forall y \in \{0,\dots,n\},\ \ P_Y(y) = \binom{n}{y}\frac{\mathrm{B}(\alpha+y,\beta+n-y)}{\mathrm{B}(\alpha,\beta)} \tag{5.57}$$

with parameters $n \in \mathbb{Z}^+$ (which is fixed in this case), $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$. Note that the update of the conjugate prior in equation 5.56 differs from that in equation 5.44. However, the updated parameters in equation 5.56 are both linear in y. Hence, it suffices to prove that the conditional moments of the binomial and beta distributions are polynomials in the their parameters in order to derive the SVD of the DTM of the binomial channel. This is carried out in subsection 5.3.5.

We note that all the examples presented here use the Lebesgue and counting measures, and deal with well-behaved functions. Therefore, in the remainder of this chapter, we will omit the measure theoretic technicalities and simply use summations and Riemann integrals without any loss of rigor.

5.3.2 Orthogonal Polynomials

In this subsection, we introduce the orthogonal polynomials that are pertinent to the three examples we are considering. Much of the theory on orthogonal polynomials we will present (with some modifications) can be found in [34], which offers a concise introduction to the subject, and [35], which offers a more comprehensive coverage. We will use these sources in the ensuing discussion without tediously referring back to them in every definition. The overarching idea of orthogonal polynomials is fairly simple. Suppose we are given a Hilbert space of functions over a common support equipped with an inner product such that polynomials are valid vectors in this Hilbert space. Typically, the monomials $\{1, x, x^2, \dots\}$ are linearly independent in this space. So, the Gram-Schmidt algorithm can be applied to them to produce orthogonal polynomials. These orthogonal polynomials are unique up to scaling. This uniqueness is potentially surprising, because one may conceive of initiating the Gram-Schmidt algorithm with a different linearly independent set of polynomials. However, a set of orthogonal polynomials is formally defined as a set of polynomials $\{p_0, p_1, p_2, \dots\}$ such that each p_k has degree k, and every polynomial in the set is orthogonal to every other polynomial in the set. This definition ensures the uniqueness property. If we vary the support of the functions and the weight function of the inner product, we can derive several classes of orthogonal polynomials. Such classes of polynomials have been studied extensively, because they appear in various areas of applied mathematics like perturbation theory, quantum mechanics, and stochastic processes.

For each channel we stated earlier, we will require two sets of orthogonal polynomials. The first set includes the polynomials which are orthogonal with respect to the input distribution, and the second set includes the polynomials which are orthogonal with respect to the output distribution. Since the Gaussian channel has Gaussian input and output, we only need to define one class of orthogonal polynomials for it. So, five classes of polynomials are presented next.

Hermite Polynomials

The Hermite polynomials form a family of polynomials on the real line that are orthogonal with respect to the standard Gaussian pdf. In particular, the Hermite polynomial with degree $k \in \mathbb{N}$, denoted $H_k : \mathbb{R} \to \mathbb{R}$, is defined as:

$$\forall x \in \mathbb{R}, \;\; H_k(x) \triangleq (-1)^k e^{\frac{x^2}{2}} \frac{d^k}{dx^k} \left(e^{-\frac{x^2}{2}} \right). \tag{5.58}$$

Such a formula to generate orthogonal polynomials using repeated derivatives (or repeated finite differences in the discrete case) is called a Rodrigues formula. The orthogonality relation for Hermite polynomials is:

$$\forall j, k \in \mathbb{N}, \;\; \int_{-\infty}^{\infty} H_j(x) H_k(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \; dx = k! \delta_{jk}. \tag{5.59}$$

Since we will require polynomials that are orthonormal with respect to Gaussian pdfs with arbitrary variances, we define the normalized Hermite polynomials for any variance $p > 0$ as:

$$\forall k \in \mathbb{N}, \forall x \in \mathbb{R}, \;\; H_k^{(p)}(x) \triangleq \frac{1}{\sqrt{k!}} H_k \left(\frac{x}{\sqrt{p}} \right). \tag{5.60}$$

It can be checked that these polynomials are orthonormal with respect to the Gaussian pdf for any variance $p > 0$:

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

The normalized Hermite polynomials will be used in subsection 5.3.3.

Generalized Laguerre Polynomials

The generalized Laguerre polynomials form a family of polynomials on the positive real line that are orthogonal with respect to the gamma pdf. In particular, the generalized Laguerre polynomial with degree $k \in \mathbb{N}$, denoted $L_k^{(a)} : (0, \infty) \to \mathbb{R}$, is defined by the Rodrigues formula:

$$\forall x > 0, \ \ L_k^{(a)}(x) \triangleq \frac{x^{-a} e^x}{k!} \frac{d^k}{dx^k} \left(x^{k+a} e^{-x} \right) \tag{5.61}$$

where the parameter $a \in (-1, \infty)$. The orthogonality relation for generalized Laguerre polynomials is:

$$\forall j, k \in \mathbb{N}, \ \int_0^{\infty} L_j^{(a)}(x) L_k^{(a)}(x) x^a e^{-x} \ dx = \frac{\Gamma(k + a + 1)}{k!} \delta_{jk}. \tag{5.62}$$

The special case when $a = 0$ begets the Laguerre polynomials, which are orthogonal with respect to the exponential pdf. Since we will require polynomials that are orthonormal with respect to gamma pdfs with arbitrary parameters $\alpha > 0$ and $\beta > 0$, we define the normalized generalized Laguerre polynomials with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$ as:

$$\forall k \in \mathbb{N}, \forall x > 0, \ \ L_k^{(\alpha,\beta)}(x) \triangleq \sqrt{\frac{k! \Gamma(\alpha)}{\Gamma(k + \alpha)}} L_k^{(\alpha - 1)}(\beta x) \, . \tag{5.63}$$

It can be checked that these polynomials are orthonormal with respect to the gamma pdf for any $\alpha > 0$ and $\beta > 0$:

$$\forall j, k \in \mathbb{N}, \ \int_0^{\infty} L_j^{(\alpha,\beta)}(x) L_k^{(\alpha,\beta)}(x) \frac{\beta^\alpha x^{\alpha - 1} e^{-\beta x}}{\Gamma(\alpha)} \ dx = \delta_{jk}.$$

The normalized generalized Laguerre polynomials will be used in subsection 5.3.4.

Jacobi Polynomials

The Jacobi polynomials form a family of polynomials on $(-1, 1)$ that are orthogonal with respect to the beta pdf on $(-1, 1)$. Together, the Hermite, generalized Laguerre, and Jacobi polynomials are often called the classical orthogonal polynomials because they share many common properties. For example, within a linear change of variables, they are the only orthogonal polynomials whose derivatives are also orthogonal polynomials [34]. The Jacobi polynomial with degree $k \in \mathbb{N}$, denoted $\bar{J}_k^{(a,b)} : (-1, 1) \to \mathbb{R}$, is defined by the Rodrigues formula:

$$\forall x \in (-1, 1), \ \ \bar{J}_k^{(a,b)}(x) \triangleq (1 - x)^{-a} (1 + x)^{-b} \frac{(-1)^k}{2^k k!} \frac{d^k}{dx^k} \left((1 - x)^{k+a} (1 + x)^{k+b} \right) \tag{5.64}$$

where the parameters $a, b \in (-1, \infty)$. The orthogonality relation for Jacobi polynomials is:

$$\forall j,k \in \mathbb{N},\ \int_{-1}^{1} \bar{J}_j^{(a,b)}(x)\bar{J}_k^{(a,b)}(x)(1-x)^a(1+x)^b\ dx = \frac{2^{a+b+1}\Gamma(k+a+1)\Gamma(k+b+1)}{(2k+a+b+1)\Gamma(k+a+b+1)k!}\delta_{jk}. \tag{5.65}$$

The Jacobi polynomials generalize several other orthogonal polynomial families like the Legendre and Chebyshev polynomials. In particular, letting $a = b = 0$ produces the Legendre polynomials which are orthogonal with respect to the uniform pdf. Since we will require polynomials that are orthonormal with respect to beta pdfs on $(0,1)$ with arbitrary parameters $\alpha > 0$ and $\beta > 0$, we define the normalized (shifted) Jacobi polynomials with parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$ as:

$$\forall k \in \mathbb{N}, \forall x \in (0,1),\ \ J_k^{(\alpha,\beta)}(x) \triangleq \sqrt{\frac{(2k+\alpha+\beta-1)\mathrm{B}(\alpha,\beta)\Gamma(k+\alpha+\beta-1)k!}{\Gamma(k+\alpha)\Gamma(k+\beta)}}\bar{J}_k^{(\beta-1,\alpha-1)}\left(2x-1\right). \tag{5.66}$$

It can be checked that these polynomials are orthonormal with respect to the beta pdf for any $\alpha > 0$ and $\beta > 0$:

$$\forall j,k \in \mathbb{N},\ \int_0^1 J_j^{(\alpha,\beta)}(x)J_k^{(\alpha,\beta)}(x)\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\ dx = \delta_{jk}.$$

The normalized Jacobi polynomials will be used in subsection 5.3.5.

Meixner Polynomials

The Meixner polynomials form a family of polynomials on the natural numbers $\mathbb{N} = \{0,1,2,\dots\}$ that are orthogonal with respect to the negative binomial pmf. This is the first class of polynomials we have encountered that has discrete (countable) support. Such orthogonal polynomials are often conveniently defined using hypergeometric series. For any $p, q \in \mathbb{Z}^+$, the hypergeometric series is defined as:

$${}_pF_q\left(\begin{array}{c} a_1,\dots,a_p \\ b_1,\dots,b_q \end{array};t\right) \triangleq \sum_{k=0}^{\infty}\frac{(a_1)_k\dots(a_p)_k t^k}{(b_1)_k\dots(b_q)_k k!} \tag{5.67}$$

where the argument $t \in \mathbb{C}$ and the parameters $a_1,\dots,a_p,b_1,\dots,b_q \in \mathbb{C}$ are such that the series is well-defined. Moreover, for any $k \in \mathbb{N}$ and any $z \in \mathbb{C}$, $(z)_k$ is the (non-standard) Pochhammer symbol for the rising factorial, which is defined for $k \geq 1$ as:

$$(z)_k \triangleq z(z+1)(z+2)\dots(z+k-1) = \frac{\Gamma(z+k)}{\Gamma(z)} \tag{5.68}$$

and for $k = 0$ as $(z)_0 = 1$. The Meixner polynomial with degree $k \in \mathbb{N}$, denoted $\bar{M}_k^{(s,p)} : \mathbb{N} \to \mathbb{R}$, can be defined using the hypergeometric series as:

$$\forall y \in \mathbb{N},\ \ \bar{M}_k^{(s,p)}(y) \triangleq {}_2F_1\left(\begin{array}{c} -k,-y \\ s \end{array};1-\frac{1}{p}\right) \tag{5.69}$$

where the parameters $s \in (0,\infty)$ and $p \in (0,1)$. The orthogonality relation for Meixner polynomials is:

$$\forall j,k \in \mathbb{N},\ \sum_{y=0}^{\infty}\bar{M}_j^{(s,p)}(y)\bar{M}_k^{(s,p)}(y)\frac{\Gamma(s+y)}{\Gamma(s)y!}p^y(1-p)^s = \frac{p^{-k}k!}{(s)_k}\delta_{jk}. \tag{5.70}$$

These polynomials are already orthogonal with respect to negative binomial pmfs with arbitrary parameters $s > 0$ and $p \in (0,1)$. So, we define the normalized Meixner polynomials with parameters $s \in (0,\infty)$ and $p \in (0,1)$ as:

$$\forall k \in \mathbb{N}, \forall y \in \mathbb{N}, \;\; M_k^{(s,p)}(y) \triangleq \sqrt{\frac{(s)_k}{p^{-k}k!}} \bar{M}_k^{(s,p)}(y)\,. \tag{5.71}$$

Obviously, these polynomials are orthonormal with respect to the negative binomial pmf for any $s > 0$ and $p \in (0,1)$:

$$\forall j,k \in \mathbb{N}, \;\; \sum_{y=0}^{\infty} M_j^{(s,p)}(y) M_k^{(s,p)}(y) \frac{\Gamma(s+y)}{\Gamma(s)y!} p^y (1-p)^s = \delta_{jk}.$$

The normalized Meixner polynomials will be used in subsection 5.3.4.

Hahn Polynomials

Finally, the Hahn polynomials form a family of polynomials on the discrete and finite set $\{0,\ldots,n\}$ that are orthogonal with respect to the beta-binomial pmf. In particular, the Hahn polynomial with degree $k \in \{0,\ldots,n\}$, denoted $\bar{Q}_k^{(a,b)} : \{0,\ldots,n\} \to \mathbb{R}$, can be defined using the hypergeometric series as:

$$\forall y \in \{0,\ldots,n\}\,, \;\; \bar{Q}_k^{(a,b)}(y) \triangleq {}_3F_2\left(\begin{matrix} -k, k+a+b+1, -y \\ a+1, -n \end{matrix} ; 1 \right) \tag{5.72}$$

where the parameters $a, b \in (-1,\infty)$. The orthogonality relation for Hahn polynomials is:

$$\forall j,k \in \{0,\ldots,n\}\,, \;\; \sum_{y=0}^{n} \bar{Q}_j^{(a,b)}(y) \bar{Q}_k^{(a,b)}(y) \frac{(a+1)_y (b+1)_{n-y}}{y!(n-y)!} = \frac{(k+a+b+1)_{n+1}(b+1)_k}{\binom{n}{k}(2k+a+b+1)(a+1)_k n!} \delta_{jk}. \tag{5.73}$$

The Hahn polynomials 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 pmfs, respectively [34]. Since we will require polynomials that are orthonormal with respect to beta-binomial pmfs with arbitrary parameters $\alpha > 0$ and $\beta > 0$, we define the normalized Hahn polynomials with parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$ as:

$$\forall k, y \in \{0,\ldots,n\}\,, \;\; Q_k^{(\alpha,\beta)}(y) \triangleq \sqrt{\binom{n}{k} \frac{(2k+\alpha+\beta-1)(\alpha)_k \Gamma(n+\alpha+\beta)}{(k+\alpha+\beta-1)_{n+1}(\beta)_k \Gamma(\alpha+\beta)}} \bar{Q}_k^{(\alpha-1,\beta-1)}(y)\,. \tag{5.74}$$

It can be checked that these polynomials are orthonormal with respect to the beta-binomial pmf for any $\alpha > 0$ and $\beta > 0$:

$$\forall j,k \in \{0,\ldots,n\}\,, \;\; \sum_{y=0}^{n} Q_j^{(\alpha,\beta)}(y) Q_k^{(\alpha,\beta)}(y) \binom{n}{y} \frac{\mathrm{B}(\alpha+y, \beta+n-y)}{\mathrm{B}(\alpha,\beta)} = \delta_{jk}.$$

The normalized Hahn polynomials will be used in subsection 5.3.5. This completes our introduction to the relevant orthogonal polynomial families.

5.3.3 Gaussian Input, Gaussian Channel, and Hermite Polynomials

The SVD of the DTM of the additive white Gaussian noise (AWGN) channel was computed in [23] in the context of attacking network information theory problems which use additive Gaussian noise models. This subsection is devoted to providing an alternative derivation of this SVD. We will find the singular vectors of the DTM of the AWGN channel by directly employing Theorem 5.2.8. To this end, recall Definition 3.4.1 of the (single letter) AWGN channel. The AWGN channel has jointly distributed input random variable X and output random variable Y, where X and Y are related by the equation:

$$Y = X + W, \;\; X \perp\!\!\!\perp W \sim \mathcal{N}(0, \nu) \tag{5.75}$$

such that X is independent of the Gaussian noise $W \sim \mathcal{N}(0, \nu)$ with variance $\nu > 0$, and X satisfies the average power constraint, $\mathbb{E}\left[X^2\right] \leq p$, for some given power $p > 0$. We assume that we use the capacity achieving input distribution $X \sim \mathcal{N}(0, p)$. Then, the marginal pdf of X is:

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

which coincides with equation 5.46 in subsection 5.3.1 when $r = 0$, and the conditional pdfs of Y given X are:

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

which coincide with equation 5.45 in subsection 5.3.1. Moreover, the marginal pdf of Y is:

$$\forall y \in \mathbb{R}, \;\; P_Y(y) = \frac{1}{\sqrt{2\pi(p+\nu)}} \exp\left(-\frac{y^2}{2(p+\nu)}\right) \tag{5.76}$$

because $Y \sim \mathcal{N}(0, p+\nu)$. We next define the notion of translation invariant kernels, which will play a considerable role in computing the SVD of the DTM of the AWGN channel.

Definition 5.3.4 (Translation Invariant Kernel)**.** Suppose we are given an integral operator T, which is defined for an input function $f : \mathbb{R} \to \mathbb{R}$ as:

$$\forall x \in \mathbb{R}, \;\; T(f)(x) = \int_{-\infty}^{\infty} K(x,t) f(t) \;\; dt$$

where $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the kernel of the integral operator. The kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called a translation invariant kernel if there exists a function $\phi : \mathbb{R} \to \mathbb{R}$ such that:

$$\forall x, t \in \mathbb{R}, \;\; K(x,t) = \phi(x + vt)$$

for some constant $v \in \mathbb{R} \backslash \{0\}$.

We remark that the term "translation invariant" is often reserved for kernels where $v = -1$ in the above definition. Such kernels with $v = -1$ are also known as difference kernels and correspond to the convolution operation. Indeed, the convolution of $\phi : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ is given by:

$$\forall x \in \mathbb{R}, \;\; (\phi \star f)(x) = \int_{-\infty}^{\infty} \phi(x-t) f(t) \;\; dt \tag{5.77}$$

where $\star$ denotes the convolution operation. This operation plays a vital role in the theory of linear time-invariant (LTI) systems. In Definition 5.3.4, we refer to the kernel $K(x,t) = \phi(x+vt)$ as translation invariant because it can be physically interpreted as describing a wave traveling at constant phase velocity v with a fixed profile. Functions of the form $\phi(x+vt)$ are often encountered while solving partial differential equations involving the d'Alembert operator (which appears in the study of wave phenomena). The next lemma presents a crucial observation regarding translation invariant kernels.

Lemma 5.3.1 (Closure and Degree Preservation of Translation Invariant Kernels)**.** *An integral operator T with a translation invariant kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that $\forall x, t \in \mathbb{R},\ K(x,t) = \phi(x+vt)$ for some $v \in \mathbb{R}\setminus\{0\}$, is closed over polynomials and degree preserving. Furthermore, T is strictly degree preserving if and only if:*

$$0 < \left| \int_{-\infty}^{\infty} \phi(z)\ dz \right| < \infty.$$

Proof.
Suppose T is an integral operator with translation invariant kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $\forall x, t \in \mathbb{R},\ K(x,t) = \phi(x+vt)$ for some $v \in \mathbb{R}\setminus\{0\}$. Consider any polynomial $p : \mathbb{R} \to \mathbb{R}$ with degree $n \in \mathbb{N}$ defined by $p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n$, which means $a_n \neq 0$. Applying T to p, we have:

$$\forall x \in \mathbb{R},\ T(p)(x) = \int_{-\infty}^{\infty} K(x,t)p(t)\ dt = \int_{-\infty}^{\infty} \phi(x+vt)\left(a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n\right)\ dt.$$

We seek to show that $T(p)$ is a polynomial with degree at most n. Observe that:

$$\begin{aligned}
\forall x \in \mathbb{R},\ T(p)(x) &= \int_{-\infty}^{\infty} \phi(x+vt)\left(a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n\right)\ dt \\
&= \sum_{i=0}^{n} a_i \int_{-\infty}^{\infty} t^i \phi(x+vt)\ dt \\
&= \operatorname{sign}(v) \sum_{i=0}^{n} \frac{a_i}{v^{i+1}} \int_{-\infty}^{\infty} (z-x)^i\, \phi(z)\ dz \\
&= \operatorname{sign}(v) \sum_{i=0}^{n} \frac{a_i}{v^{i+1}} \int_{-\infty}^{\infty} \left(\sum_{j=0}^{i} (-1)^j \binom{i}{j} z^{i-j} x^j \right) \phi(z)\ dz \\
&= \operatorname{sign}(v) \sum_{i=0}^{n} \frac{a_i}{v^{i+1}} \sum_{j=0}^{i} (-1)^j \binom{i}{j} x^j \int_{-\infty}^{\infty} z^{i-j} \phi(z)\ dz \\
&= \operatorname{sign}(v) \sum_{j=0}^{n} x^j \left((-1)^j \sum_{i=j}^{n} \binom{i}{j} \frac{a_i}{v^{i+1}} \int_{-\infty}^{\infty} z^{i-j} \phi(z)\ dz \right)
\end{aligned}$$

where the third equality follows from a change of variables $z = x + vt$, the fourth equality follows from the binomial theorem, and the final equality follows from swapping the order of summations (which is valid as the summations are finite). Clearly, $T(p)$ is a polynomial with degree at most n.

Hence, T is closed over polynomials and degree preserving. Now notice that the coefficient of x^n in the expression for $T(p)(x)$ is:

$$\operatorname{sign}(v)(-1)^n \sum_{i=n}^{n} \binom{i}{n} \frac{a_i}{v^{i+1}} \int_{-\infty}^{\infty} z^{i-n} \phi(z) \ dz = \operatorname{sign}(v) \frac{(-1)^n a_n}{v^{n+1}} \int_{-\infty}^{\infty} \phi(z) \ dz$$

where $\frac{(-1)^n a_n}{v^{n+1}}$ is some finite non-zero constant. Therefore, $T(p)$ has degree n if and only if:

$$0 < \left| \int_{-\infty}^{\infty} \phi(z) \ dz \right| < \infty,$$

which means T is strictly degree preserving if and only if the above condition holds. This completes the proof. □

The SVD of the DTM of the AWGN channel is presented in the next theorem. Its proof employs the closure over polynomials and degree preservation properties of translation invariant kernels.

Theorem 5.3.2 (AWGN Channel SVD)**.** *Given the AWGN channel:*

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

with input random variable $X \sim \mathcal{N}(0, p)$ that is Gaussian distributed with variance $p > 0$, the SVD of the DTM, $B : \mathcal{L}^2(\mathbb{R}, \lambda) \to \mathcal{L}^2(\mathbb{R}, \mu)$, is given by:

$$\forall k \in \mathbb{N}, \ \ B\left(\sqrt{P_X} H_k^{(p)}\right) = \left(\frac{p}{p+\nu}\right)^{\frac{k}{2}} \sqrt{P_Y} H_k^{(p+\nu)}$$

where the singular vectors are normalized Hermite polynomials weighted by the square root of the corresponding Gaussian pdfs.

Proof.
It suffices to demonstrate that the conditional expectation operators are strictly degree preserving. Theorem 5.2.8 will then ensure that the singular vectors have the form given in the theorem statement, and the singular values are strictly positive and bounded by 1, with the first singular value being exactly equal to 1. The explicit form of the singular values is proven in [23], and we do not prove it here.

We first show that the conditional expectation operator $\mathbb{E}[\cdot|X]$ is closed over polynomials and strictly degree preserving. For any real (Borel measurable) function $f : \mathbb{R} \to \mathbb{R}$, we have:

$$\mathbb{E}[f(Y)|X = x] = \int_{\mathbb{R}} f(y) P_{Y|X}(y|x) \ d\mu(y) = \int_{\mathbb{R}} f(y) \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) \ d\mu(y)$$

where μ is the Lebesgue measure and the integrals are Lebesgue integrals. So, $\mathbb{E}[\cdot|X]$ is an integral operator with translation invariant kernel (with $v = -1$). Moreover, we have:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{z^2}{2\nu}\right) \ dz = 1$$

and hence, $\mathbb{E}[\cdot|X]$ is closed over polynomials and strictly degree preserving using Lemma 5.3.1.

We next prove that the conditional expectation operator $\mathbb{E}[\cdot|Y]$ is closed over polynomials and strictly degree preserving. For zero mean jointly Gaussian random variables (X, Y), the conditional expectation of X given $Y = y \in \mathbb{R}$ is:

$$\mathbb{E}[X|Y=y] = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}y = \frac{\mathbb{E}\left[X^2\right] + \mathbb{E}[X]\mathbb{E}[W]}{\mathbb{E}[Y^2]}y = \frac{p}{p+\nu}y \tag{5.78}$$

as $Y = X + W$ and $X \perp\!\!\!\perp W$. This is also the minimum mean-square error (MMSE) estimator of X given Y. The conditional variance of X given $Y = y \in \mathbb{R}$ is:

$$\mathbb{VAR}(X|Y=y) = \mathbb{E}\left[X^2\right] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} = \mathbb{E}\left[X^2\right] - \frac{\left(\mathbb{E}\left[X^2\right] + \mathbb{E}[X]\mathbb{E}[W]\right)^2}{\mathbb{E}[Y^2]} = \frac{p\nu}{p+\nu}. \tag{5.79}$$

Hence, the conditional pdfs of X given Y are:

$$\forall y, x \in \mathbb{R},\ \ P_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\left(\frac{p\nu}{p+\nu}\right)}} \exp\left(-\frac{\left(x - \frac{p}{p+\nu}y\right)^2}{2\left(\frac{p\nu}{p+\nu}\right)}\right) \tag{5.80}$$

where the conditional distributions must be Gaussian because (X, Y) are jointly Gaussian. This means for any real (Borel measurable) function $f : \mathbb{R} \to \mathbb{R}$, we have:

$$\mathbb{E}[f(X)|Y=y] = \int_{\mathbb{R}} f(x) P_{X|Y}(x|y)\ d\lambda(x) = \int_{\mathbb{R}} f(x) \frac{1}{\sqrt{2\pi\left(\frac{p\nu}{p+\nu}\right)}} \exp\left(-\frac{\left(x - \frac{p}{p+\nu}y\right)^2}{2\left(\frac{p\nu}{p+\nu}\right)}\right)\ d\lambda(x)$$

where λ is the Lebesgue measure and the integrals are Lebesgue integrals. So, $\mathbb{E}[\cdot|Y]$ is also an integral operator with translation invariant kernel (with $v = -\frac{p}{p+\nu}$). Moreover, we again have:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\left(\frac{p\nu}{p+\nu}\right)}} \exp\left(-\frac{z^2}{2\left(\frac{p\nu}{p+\nu}\right)}\right)\ dz = 1$$

which implies that $\mathbb{E}[\cdot|Y]$ is closed over polynomials and strictly degree preserving using Lemma 5.3.1. This completes the proof. □

As mentioned earlier, Theorem 5.3.2 concurs with the SVD result in [23]. The right and left singular vectors derived in the theorem form complete orthonormal bases of $\mathcal{L}^2(\mathbb{R}, \lambda)$ and $\mathcal{L}^2(\mathbb{R}, \mu)$, respectively. Moreover, the normalized Hermite polynomials with parameter p form a complete orthonormal basis of $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_X)$, and the normalized Hermite polynomials with parameter $p + \nu$ form a complete orthonormal basis of $\mathcal{L}^2(\mathbb{R}, \mathbb{P}_Y)$. Theorem 5.3.2 illustrates how Theorem 5.2.8 may be used to swiftly and efficiently derive SVDs of DTMs of channels. This is possible because Theorem 5.2.8 essentially transfers the technical intricacies of computing the SVD in a functional space into blindly verifying some conditions on the conditional moments.

5.3.4 Gamma Input, Poisson Channel, and Laguerre and Meixner Polynomials

In this subsection, we find the SVD of the DTM of the Poisson channel introduced in subsection 5.3.1. Recall that the Poisson channel has input random variable X with range $\mathcal{X} = (0, \infty)$, and output random variable Y with range $\mathcal{Y} = \mathbb{N}$. The channel conditional distributions of Y given X are Poisson distributions:

$$\forall x > 0, \forall y \in \mathbb{N}, \;\; P_{Y|X}(y|x) = \frac{x^y e^{-x}}{y!}.$$

We assume that X is gamma distributed with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$:

$$\forall x > 0, \;\; P_X(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},$$

and Y is therefore negative binomial distributed with parameters $p = \frac{1}{\beta+1}$ and $s = \alpha$:

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

The next theorem derives the SVD of the DTM of this channel using Theorem 5.2.8.

Theorem 5.3.3 (Poisson Channel SVD)**.** *Given the Poisson channel with input random variable X that is gamma distributed with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$, the SVD of the DTM, $B : \mathcal{L}^2((0, \infty), \lambda) \to \mathcal{L}^2(\mathbb{N}, \mu)$, is given by:*

$$\forall k \in \mathbb{N}, \;\; B\left(\sqrt{P_X} L_k^{(\alpha,\beta)}\right) = \sigma_k \sqrt{P_Y} M_k^{\left(\alpha, \frac{1}{\beta+1}\right)}$$

for some singular values σ_k such that $\forall k \in \mathbb{N}, \;\; 0 < \sigma_k \leq 1$ and $\sigma_0 = 1$, where the right singular vectors are normalized generalized Laguerre polynomials weighted by the square root of the gamma pdf, and the left singular vectors are normalized Meixner polynomials weighted by the square root of the negative binomial pmf.

Proof.
From the discussion in subsection 5.3.1, it suffices to show that the moments of the Poisson distribution are strictly degree preserving polynomials of the rate parameter, and the moments of the gamma distribution are strictly degree preserving polynomials of the parameter α (see equation 5.50). This will imply that the conditional expectation operators are closed over polynomials and strictly degree preserving, and then Theorem 5.2.8 will beget the SVD presented in the theorem statement.

We first prove that the moments of the Poisson random variable Z with rate parameter $x > 0$ are strictly degree preserving polynomials of x. The pmf of Z is precisely the conditional pmf of Y given $X = x \in (0, \infty)$:

$$\forall z \in \mathbb{N}, \;\; P_Z(z) = \frac{x^z e^{-x}}{z!} = P_{Y|X}(z|x),$$

and $\forall n \in \mathbb{N}, \;\; \mathbb{E}[Z^n] = \mathbb{E}[Y^n | X = x]$. The moment generating function (MGF) of Z is:

$$\forall s \in \mathbb{R}, \;\; M_Z(s) \triangleq \mathbb{E}\left[e^{sZ}\right] = e^{x(e^s - 1)} \tag{5.81}$$

and as the MGF is finite on an open interval around $s = 0$, the moments of Z are given by:

$$\forall n \in \mathbb{N},\ \mathbb{E}\left[Z^n\right] = \left.\frac{d^n}{ds^n} M_Z(s)\right|_{s=0} = \left.\frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right)\right|_{s=0}. \tag{5.82}$$

To show these moments are polynomials of x, we prove by induction that:

$$\forall n \in \mathbb{N},\ \frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(\sum_{k=0}^{n} a_k^n(s)x^k\right) \tag{5.83}$$

for some infinitely differentiable functions $\{a_k^n : \mathbb{R} \to \mathbb{R} \mid n \in \mathbb{N},\ 0 \leq k \leq n\}$. For the base case of $n = 0$:

$$e^{x(e^s-1)} = \frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(a_0^0(s)x^0\right),$$

where we take $a_0^0(s) \triangleq 1$, which is infinitely differentiable. We then assume that for some fixed $n \in \mathbb{N}$:

$$\frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(\sum_{k=0}^{n} a_k^n(s)x^k\right)$$

for some infinitely differentiable functions $a_k^n,\ 0 \leq k \leq n$. This is the inductive hypothesis. The inductive step then requires us to prove that:

$$\frac{d^{n+1}}{ds^{n+1}}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(\sum_{k=0}^{n+1} a_k^{n+1}(s)x^k\right)$$

for some infinitely differentiable functions $a_k^{n+1},\ 0 \leq k \leq n+1$. By the inductive hypothesis, we have:

$$\begin{aligned}
\frac{d^{n+1}}{ds^{n+1}}\left(e^{x(e^s-1)}\right) &= \frac{d}{ds}\left(e^{x(e^s-1)}\left(\sum_{k=0}^{n} a_k^n(s)x^k\right)\right) \\
&= e^{x(e^s-1)}\left(\sum_{k=0}^{n} x^k \frac{d}{ds}a_k^n(s)\right) + xe^s e^{x(e^s-1)}\left(\sum_{k=0}^{n} a_k^n(s)x^k\right) \\
&= e^{x(e^s-1)}\left(\frac{d}{ds}a_0^n(s) + \sum_{k=1}^{n} x^k\left(\frac{d}{ds}a_k^n(s) + a_{k-1}^n(s)e^s\right) + a_n^n(s)e^s x^{n+1}\right)
\end{aligned}$$

where we define $a_0^{n+1}(s) \triangleq a_0^n(s)$, $a_k^{n+1}(s) \triangleq \frac{d}{ds}a_k^n(s) + a_{k-1}^n(s)e^s$ for $1 \leq k \leq n$, and $a_{n+1}^{n+1}(s) \triangleq a_n^n(s)e^s$. The functions $a_k^{n+1},\ 0 \leq k \leq n+1$, are infinitely differentiable because e^s and the functions $a_k^n,\ 0 \leq k \leq n$, are infinitely differentiable. Hence, we have:

$$\frac{d^{n+1}}{ds^{n+1}}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(\sum_{k=0}^{n+1} a_k^{n+1}(s)x^k\right)$$

as required. By induction, we get:

$$\forall n \in \mathbb{N},\ \frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right) = e^{x(e^s-1)}\left(\sum_{k=0}^{n} a_k^n(s)x^k\right)$$

for some infinitely differentiable functions $\{a_k^n : \mathbb{R} \to \mathbb{R} \mid n \in \mathbb{N},\ 0 \leq k \leq n\}$. This implies that:

$$\forall n \in \mathbb{N},\ \mathbb{E}\left[Z^n\right] = \frac{d^n}{ds^n}\left(e^{x(e^s-1)}\right)\Big|_{s=0} = \sum_{k=0}^{n} a_k^n(0)x^k \tag{5.84}$$

from equation 5.82. Note that from our recursive definition, $a_{n+1}^{n+1}(s) = a_n^n(s)e^s$ and $a_0^0(s) = 1$, we can conclude that $\forall n \in \mathbb{N},\ a_n^n(0) = a_0^0(0) = 1$. Therefore, $\forall n \in \mathbb{N},\ \mathbb{E}\left[Z^n\right] = \mathbb{E}\left[Y^n|X = x\right]$ is a polynomial in x with degree n.

Next, we prove that the moments of a gamma distribution are strictly degree preserving polynomials of α. Let X be a gamma distributed random variable with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$:

$$\forall x > 0,\ P_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}.$$

The MGF of X is:

$$M_X(s) \triangleq \mathbb{E}\left[e^{sX}\right] = \begin{cases} \left(\frac{\beta}{\beta-s}\right)^\alpha & ,\ s < \beta \\ \infty & ,\ s \geq \beta \end{cases} \tag{5.85}$$

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

$$\forall n \in \mathbb{N},\ \mathbb{E}\left[X^n\right] = \frac{d^n}{ds^n}M_X(s)\Big|_{s=0} = \frac{d^n}{ds^n}\left(\frac{\beta}{\beta-s}\right)^\alpha\Big|_{s=0} = \left(\frac{\beta}{\beta-s}\right)^\alpha \frac{(\alpha)_n}{(\beta-s)^n}\Big|_{s=0}$$

where we have used the Pochhammer symbol defined in equation 5.68. This simplifies to:

$$\forall n \in \mathbb{N},\ \mathbb{E}\left[X^n\right] = \frac{(\alpha)_n}{\beta^n} \tag{5.86}$$

from which it is evident that for every $n \in \mathbb{N}$, $\mathbb{E}\left[X^n\right]$ is a polynomial in α with degree n. This completes the proof. □

Firstly, we note that the SVD of the DTM of the Poisson channel with exponential input distribution and geometric output distribution was derived in [8]. This result is a special case of Theorem 5.3.3. Indeed, letting $\alpha = 1$ causes X to have exponential distribution and Y to have geometric distribution. Moreover, the generalized Laguerre polynomials become Laguerre polynomials, which are orthogonal with respect to the exponential distribution. This is precisely the SVD result in [8]. Furthermore, [8] provides an explicit formula for the singular values in the $\alpha = 1$ case as well. Secondly, we remark that there are more insightful avenues to derive that the moments of a Poisson pmf are strictly degree preserving polynomials of the rate parameter. However, such methods require a study of combinatorial tools like Bell polynomials and Stirling numbers, and we avoid developing these concepts for the sake of brevity. Finally, we note that in Theorem 5.3.3, the right and left singular vectors form complete orthonormal bases of $\mathcal{L}^2\left((0, \infty), \lambda\right)$ and $\mathcal{L}^2\left(\mathbb{N}, \mu\right)$, respectively. Moreover, the normalized generalized Laguerre polynomials form a complete orthonormal basis of $\mathcal{L}^2\left((0, \infty), \mathbb{P}_X\right)$, and the normalized Meixner polynomials form a complete orthonormal basis of $\mathcal{L}^2\left(\mathbb{N}, \mathbb{P}_Y\right)$. Like the SVD of the DTM of the AWGN channel derived earlier, the result in Theorem 5.3.3 also illustrates the utility of Theorem 5.2.8 in deriving SVDs of DTMs of infinite alphabet channels.

5.3.5 Beta Input, Binomial Channel, and Jacobi and Hahn Polynomials

In the final subsection of this chapter, we derive the SVD of the DTM of the binomial channel. Recall from subsection 5.3.1 that the binomial channel has input random variable X with range $\mathcal{X} = (0,1)$, and output random variable Y with range $\mathcal{Y} = \{0,\ldots,n\}$ for some fixed $n \in \mathbb{Z}^+$. The channel conditional distributions of Y given X are binomial distributions:

$$\forall x \in (0,1), \forall y \in \{0,\ldots,n\}, \;\; P_{Y|X}(y|x) = \binom{n}{y} x^y (1-x)^{n-y}.$$

We assume that X is beta distributed with parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$:

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

This implies that Y is beta-binomial distributed with the same parameters α and β:

$$\forall y \in \{0,\ldots,n\}, \;\; P_Y(y) = \binom{n}{y} \frac{\mathrm{B}(\alpha+y, \beta+n-y)}{\mathrm{B}(\alpha,\beta)}.$$

The next theorem presents the SVD of the DTM of the binomial channel using Theorem 5.2.9.

Theorem 5.3.4 (Binomial Channel SVD)**.** *Given the binomial channel with input random variable X that is beta distributed with parameters $\alpha \in (0,\infty)$ and $\beta \in (0,\infty)$, the SVD of the DTM, $B : \mathcal{L}^2((0,1),\lambda) \to \mathcal{L}^2(\{0,\ldots,n\},\mu)$, is given by:*

$$\begin{aligned} \forall k \in \{0,\ldots,n\}, \;\; B\left(\sqrt{P_X} J_k^{(\alpha,\beta)}\right) &= \sigma_k \sqrt{P_Y} Q_k^{(\alpha,\beta)} \\ \forall k \in \mathbb{N} \backslash \{0,\ldots,n\}, \;\; B\left(\sqrt{P_X} J_k^{(\alpha,\beta)}\right) &= 0 \end{aligned}$$

for some singular values σ_k such that $\forall k \in \{0,\ldots,n\}, \; 0 < \sigma_k \leq 1$ and $\sigma_0 = 1$, where the right singular vectors are normalized Jacobi polynomials weighted by the square root of the beta pdf, and the left singular vectors are normalized Hahn polynomials weighted by the square root of the beta-binomial pmf.

Proof.
It suffices to show that the conditional expectation operators are closed over polynomials and strictly degree preserving (in the sense given in the proof of Theorem 5.2.9). Using Theorem 5.2.9, this will imply the SVD presented in the theorem statement.

We first prove that the conditional moments of Y given X are strictly degree preserving polynomials. From the discussion in subsection 5.3.1, it suffices to show that the zeroth to nth moments of the binomial distribution are strictly degree preserving polynomials of the success probability parameter, and the remaining moments are only degree preserving polynomials. Let Z be a binomial random variable on $\{0,\ldots,n\}$ with success probability parameter $x \in (0,1)$. The the pmf of Z is precisely the conditional pmf of Y given $X = x \in (0,1)$:

$$\forall z \in \mathbb{N}, \;\; P_Z(z) = \binom{n}{z} x^z (1-x)^{n-z} = P_{Y|X}(z|x),$$

and $\forall m \in \mathbb{N},\ \mathbb{E}\left[Z^m\right] = \mathbb{E}\left[Y^m | X = x\right]$. It is well-known that a binomial random variable is a sum of i.i.d. Bernoulli random variables. So, let $Z_1, \ldots, Z_n$ be i.i.d. Bernoulli random variables with success probability $x \in (0,1)$ i.e. $\mathbb{P}(Z_i = 1) = x$ and $\mathbb{P}(Z_i = 0) = 1 - x$. Then, we have:

$$Z = \sum_{i=1}^{n} Z_i$$

and the moments of Z are given by:

$$\begin{aligned}
\forall m \in \mathbb{N},\ \mathbb{E}\left[Z^m\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i\right)^m\right] \\
&= \sum_{1 \leq i_1, \ldots, i_m \leq n} \mathbb{E}\left[Z_{i_1} \cdots Z_{i_m}\right] \\
&= \sum_{\substack{0 \leq k_1, \ldots, k_n \leq m \\ k_1 + \cdots + k_n = m}} \binom{m}{k_1, \ldots, k_n} \mathbb{E}\left[Z_1^{k_1} \cdots Z_n^{k_n}\right] \\
&= \sum_{\substack{0 \leq k_1, \ldots, k_n \leq m \\ k_1 + \cdots + k_n = m}} \binom{m}{k_1, \ldots, k_n} \prod_{i=1}^{n} \mathbb{E}\left[Z_i^{k_i}\right] \qquad (5.87)
\end{aligned}$$

where the second equality follows from the linearity of expectation, the third equality follows from the multinomial theorem, and the fourth equality follows from the independence of $Z_1, \ldots, Z_n$. Note that the multinomial coefficient is defined for any $m \in \mathbb{N}$ and $k_1, \ldots, k_n \in \mathbb{N}$, such that $k_1 + \cdots + k_n = m$, as:

$$\binom{m}{k_1, \ldots, k_n} \triangleq \frac{m!}{k_1! \ldots k_n!}.$$

Since the moments of the Bernoulli random variables are $\mathbb{E}\left[Z_i^0\right] = 1$ and $\forall m \in \mathbb{N}, m \geq 1,\ \mathbb{E}\left[Z_i^m\right] = x$, each term in equation 5.87 is given by:

$$\prod_{i=1}^{n} \mathbb{E}\left[Z_i^{k_i}\right] = x^{N(k_1, \ldots, k_n)}$$

where $N(k_1, \ldots, k_n)$ represents the number of k_i that are strictly greater than 0. Evidently, $N(k_1, \ldots, k_n) \leq \min(m, n)$ and $N(k_1, \ldots, k_n) = \min(m, n)$ for at least one of the terms. Hence, for every $m \leq n$, $\mathbb{E}\left[Z^m\right] = \mathbb{E}\left[Y^m | X = x\right]$ is a polynomial in x with degree m, and for every $m > n$, $\mathbb{E}\left[Z^m\right] = \mathbb{E}\left[Y^m | X = x\right]$ is a polynomial in x with degree n.

We now prove that the conditional moments of X given Y are strictly degree preserving polynomials. To this end, we compute the moments of X, which is beta distributed with parameters $\alpha \in (0, \infty)$ and $\beta \in (0, \infty)$:

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

in order to infer how the moments of a beta distribution depend on α and β.

$$\begin{aligned}\forall m \in \mathbb{N},\ \ \mathbb{E}\left[X^m\right] &= \int_0^1 x^m \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\ dx \\ &= \frac{\mathrm{B}(\alpha+m,\beta)}{\mathrm{B}(\alpha,\beta)} \\ &= \frac{\Gamma(\alpha+m)\Gamma(\alpha+\beta)}{\Gamma(\alpha+m+\beta)\Gamma(\alpha)} \\ &= \frac{(\alpha)_m}{(\alpha+\beta)_m} \end{aligned} \tag{5.88}$$

where the second and third equalities follow from the definition of the beta function in equation 5.55, and the fourth equality follows from the definition of the Pochhammer symbol in equation 5.68. Recall from equation 5.56 in subsection 5.3.1 that:

$$\forall y \in \{0,\ldots,n\}, \forall x \in (0,1),\ \ P_{X|Y}(x|y) = \frac{x^{\alpha+y-1}(1-x)^{\beta+n-y-1}}{\mathrm{B}(\alpha+y,\beta+n-y)}$$

from which we have for any $y \in \{0,\ldots,n\}$ that:

$$\forall m \in \mathbb{N},\ \ \mathbb{E}\left[X^m|Y=y\right] = \frac{(\alpha+y)_m}{(\alpha+y+\beta+n-y)_m} = \frac{(\alpha+y)_m}{(\alpha+\beta+n)_m}. \tag{5.89}$$

Notice how y fortuitously cancels in the denominator. Hence, for every $m \in \mathbb{N}$, $\mathbb{E}\left[X^m|Y=y\right]$ is a polynomial in y with degree m. This completes the proof. □

In Theorem 5.3.4, the right and left singular vectors form complete orthonormal bases of $\mathcal{L}^2\left((0,1),\lambda\right)$ and $\mathcal{L}^2\left(\{0,\ldots,n\},\mu\right)$, respectively. Furthermore, the normalized Jacobi polynomials form a complete orthonormal basis of $\mathcal{L}^2\left((0,1),\mathbb{P}_X\right)$, and the normalized Hahn polynomials form a complete orthonormal basis of $\mathcal{L}^2\left(\{0,\ldots,n\},\mathbb{P}_Y\right)$. Since Jacobi and Hahn polynomials generalize various other orthogonal polynomials, we can specialize Theorem 5.3.4 to obtain SVDs with other orthogonal polynomial singular vectors. For example, when $\alpha=\beta=1$, the normalized Jacobi polynomials are reduced to normalized Legendre polynomials. In this case, the input beta distribution becomes a uniform distribution, and the Legendre polynomials are orthogonal with respect to this uniform distribution. The binomial channel SVD in Theorem 5.3.4 illustrates the applicability of Theorem 5.2.9 in deriving SVDs of DTMs of channels which have finite alphabet size for the input or output. This concludes the final example of this chapter.

In closing, we briefly recapitulate the principal result of this chapter. We have determined a set of intuitively sound equivalent conditions on the conditional moments that ensure that the SVD of the DTM of a channel contains weighted orthogonal polynomials as singular vectors. Although our primary focus was on infinite alphabet channels, the result generalizes to channels with discrete and finite alphabets as well. We portrayed some pertinent examples of applying the conditional moment conditions by computing SVDs of DTMs of AWGN, Poisson, and binomial channels, and more generally indicated how such examples may be generated from natural exponential families with quadratic variance functions (NEFQVF) and their conjugate priors. We note that similar

work in deriving spectral decompositions of Markov chains was carried out in [36], where the authors were interested in bounding the convergence rate of Gibbs sampling. In fact, [36] presents a comprehensive list of bivariate distributions which admit polynomial spectral decompositions. However, we emphasize that our investigation was conducted independently of this work. The spirit of our investigation was in alleviating the latent technicalities of computing explicit SVDs of certain channels. We hope that this will engender linear information coupling type of analysis for such channels.

Chapter 6

Conclusion

We have explored several topics in this thesis under the common umbrella of local approximations in information theory. To bring our discourse to an end, the next section summarizes our main contributions in the preceding chapters. We then propose a few promising directions for future research in section 6.2.

6.1 Main Contributions

In chapter 1, we introduced the local approximation of KL divergence using Taylor expansions, defined the local perturbation vector spaces, and delineated the linear information coupling style of analysis developed in [4] which transforms information theory problems into linear algebra problems. We then considered two notable classes of statistical divergences in chapter 2: the f-divergences and the Bregman divergences, both of which generalize the KL divergence. In Theorems 2.1.1 and 2.2.2, we used Taylor expansions to locally approximate the f-divergence and the Bregman divergence, respectively. The local f-divergence was identical to the local KL divergence up to a scale factor. However, the local Bregman divergence had a different form which only reduced to that of the local KL divergence when the original Bregman divergence was itself a KL divergence.

In chapter 3, we identified that the locally optimal performance of data processing algorithms employing the linear information coupling approach is given by the Rényi correlation, and the globally optimal performance is given by the hypercontractive constant. In the discrete and finite setting of section 3.3, we provided an alternative proof that the Rényi correlation lower bounds the hypercontractive constant in Theorem 3.3.1, and determined upper bounds on the hypercontractive constant in terms of Rényi correlation in Theorems 3.3.7 and 3.3.8. These results were used to derive the overall performance bound in Theorem 3.3.9. In section 3.4, we considered the AWGN channel setting. We proved that the Rényi correlation actually equals the hypercontractive constant (with power constraint) for AWGN channels in Corollary 3.4.2, Theorem 3.4.6, and Theorem 3.4.8. Hence, we validated the intuitive belief that for the Gaussian case, locally optimal performance using a linear information coupling approach is globally optimal in the sense of preserving information down a Markov chain.

Chapter 4 admitted a more communications oriented perspective of local approximations and considered the scenario of sending codewords (usually drawn i.i.d. from a source distribution) down

a discrete memoryless channel. We first interpreted the results of [4] as characterizing the most probable source perturbation under local approximations in a large deviations sense, when the output empirical distribution is given and the channel is fixed. We then established two complementary counterparts of this result. In Theorems 4.2.1 and 4.2.2, we characterized the most probable source and channel perturbations in a large deviations sense when the output empirical distribution is given. Furthermore, in Theorems 4.3.1 and 4.3.4, we characterized the most probable channel perturbations in a large deviations sense when the output empirical distribution is given and the source composition is fixed. We finally used Theorem 4.3.4 to prove Theorem 4.4.1, which specifies an additive Gaussian noise MIMO channel model for normalized perturbations under local and exponential approximations. In particular, Theorem 4.4.1 stated that the normalized output perturbation is a sum of the normalized input perturbation multiplied by the DTM and some independent additive Gaussian noise, where the noise represented the normalized channel perturbations. Hence, Theorem 4.4.1 concretely expressed a source-channel decomposition of the output perturbation under local and exponential approximations.

Finally, chapter 5 considered the problem of computing SVDs of general channels like infinite alphabet channels (which are not discrete and finite), because this would enable us to perform linear information coupling style of analysis for such channels in the future. Theorem 5.2.1 presented necessary and sufficient conditions for a compact self-adjoint linear operator on a separable Hilbert space to have a complete eigenbasis of orthonormal polynomials. This result was used to derive Theorem 5.2.7, which specified sufficient conditions on the conditional expectation operators (associated with the input and output random variables of a channel) to ensure that the Gramian operator of the DTM of a channel has a complete eigenbasis of orthonormal polynomials. Theorem 5.2.7 was then used to prove the pivotal results of the chapter: Theorems 5.2.8 and 5.2.9. For infinite alphabet channels, Theorem 5.2.8 gave necessary and sufficient conditions on the conditional expectation operators so that the singular vectors of the SVDs of a DTM and its adjoint form complete orthonormal bases of weighted orthonormal polynomials. Theorem 5.2.9 presented the analogous result for channels which have finite alphabets to emphasize the generality of the result. Theorems 5.3.2, 5.3.3, and 5.3.4 utilized Theorems 5.2.8 and 5.2.9 to compute weighted orthonormal polynomial SVDs of the DTMs of the AWGN, Poisson, and binomial channels. In a nutshell, we established that the right and left singular vectors of the AWGN channel with Gaussian input are normalized Hermite polynomials, the right and left singular vectors of the Poisson channel with gamma input are normalized generalized Laguerre and normalized Meixner polynomials, and the right and left singular vectors of the binomial channel with beta input are normalized Jacobi and normalized Hahn polynomials. These SVD examples were all generated from our discussion in subsection 5.3.1. Subsection 5.3.1 presented the framework of natural exponential families with quadratic variance functions and their conjugate priors as a means of constructing channels with orthonormal polynomial SVDs. This framework unveiled the beautiful relationship between certain natural exponential families, their conjugate priors, and the corresponding orthogonal polynomials.

We hope that the richness and elegance of many of our results will encourage further research on the general topic of local approximations in information theory. On this note, the next section proposes some specific areas for future research that we consider to be interesting and potentially fruitful.

6.2 Future Directions

We suggest future directions in a chapter by chapter basis starting with chapter 3. In section 3.3 of chapter 3, we derived (performance) bounds on the hypercontractive constant in Theorem 3.3.9. In particular, for discrete and finite random variables X and Y with joint pmf $P_{X,Y}$, we showed that:

$$\rho^2(X;Y) \leq s^*(X;Y) \leq \left(\frac{2}{\min\limits_{x \in \mathcal{X}} P_X(x)} \right) \rho^2(X;Y).$$

Since both the hypercontractive constant and the Rényi correlation tensorize, the upper bound becomes loose if we consider X_1^n and Y_1^n, where i.i.d. X_1^n are sent through a discrete memoryless channel to get Y_1^n. While Corollary 3.3.10 partially remedied this tensorization issue, we ideally seek a tighter upper bound on the hypercontractive constant using Rényi correlation, such that the constant in the bound (which is currently a scaled reciprocal of the minimum probability mass of P_X) naturally tensorizes. Proving such a tighter bound is a meaningful future endeavor. Furthermore, in section 3.4 of chapter 3, we showed in Theorem 3.4.8 that the Rényi correlation is equal to the hypercontractive constant (with power constraint) for an AWGN channel. Hence, using weak convergence (or convergence in distribution) arguments like the CLT, we may be able to prove that the local linear information coupling approach is asymptotically globally optimal for any channel in some appropriate sense. Exploring this is another topic of future research.

The results of chapter 4 may be extended further as well. In chapter 4, we primarily analyzed the large deviations behavior of i.i.d. sources and discrete memoryless channels under the local lens. A natural extension is to add memory to the source or the channel. [26] presents several theorems related to large deviations theory of Markov chains. Hence, as a starting point, we may consider analyzing a discrete memoryless channel with a Markov chain input (which is a hidden Markov model). It is nontrivial to compute the large deviations exponent for such hidden Markov models using techniques in traditional large deviations theory [26]. However, under local approximations, we can derive such exponents; indeed, local approximations were introduced to make analytically intractable information theoretic problems tractable. Investigating the large deviations behavior of hidden Markov models could be a fulfilling future endeavor. We note that [8] already approaches this problem using basic tensor algebra. However, it would be intriguing to compare this tensor based approach with the locally approximated large deviations approach.

Since this thesis presents an entirely theoretical set of ideas, we now discuss some potential applications of our work (which were hinted at throughout the thesis, starting from section 1.4 in chapter 1). From a channel coding perspective, [8] explicates the communication by varying distributions coding strategy for channels which randomly permute codewords passing through them. In such channels, we can only send information by varying the empirical distributions of codewords. However, we cannot vary the empirical distributions too much because only a small set of distributions will reliably transmit information at high rates (due to the channel noise). We derived an additive Gaussian noise MIMO channel model for normalized perturbations pertinent to this scenario in chapter 4. There are ample opportunities to construct and analyze concrete coding strategies for such channels. Moreover, [8] also presents a message passing algorithm for computing informative score functions for inference on Ising models (or hidden Markov models) based on linear information coupling style of SVD analysis. Thus, there are also opportunities for developing such inference

algorithms and analyzing their performance through bounds like those in chapter 3.

From a data processing standpoint, the elegant MMSE characterization of Rényi correlation in Theorem 3.2.6 insinuates an alternating projections algorithm for finding the optimizing functions of Rényi correlation. Moreover, maximizing Rényi correlation is essentially equivalent to solving a regularized least squares problem to find the "best" linear model between two random variables by applying arbitrary measurable functions to them. Using such intuition, the authors of [31] derive the alternating conditional expectations (ACE) algorithm to perform regression (or equivalently, maximize Rényi correlation) and unearth such models from bivariate (or more general) data. Our local approximation perspective can be used to understand and further develop this algorithm. Since the Rényi correlation is the second largest singular value of the DTM when we view the bivariate data as realizations of the input and output of a channel, its optimizing functions are weighted versions of the corresponding left and right singular vectors. The proof of Theorem 3.2.4 illustrates this for the discrete and finite case. Moreover, the singular vectors of the DTM associated to larger singular values carry more information about data (that is generated i.i.d. from a bivariate distribution). Devising algorithms with strong convergence guarantees that estimate the informative singular vectors of the DTM empirically from data is a practical direction of future research. We can perceive such algorithms as performing lossy source coding, because they effectively find compact models for data which is tantamount to data compression. Therefore, the local approximation method can be employed in both source and channel coding applications.

In light of our discussion on source and channel coding algorithms which estimate or compute singular vectors of the DTM, we recall that Theorems 5.2.8 and 5.2.9 in chapter 5 characterize the channels whose DTMs have SVDs with orthogonal polynomial singular vectors. For such channels, calculating singular vectors is computationally efficient (for inference algorithms like that in [8]) because we are computing values of polynomials. Furthermore, estimating singular vectors for regression or data processing purposes becomes a much simpler task if the singular vectors are polynomials. Research can be done to develop alternative algorithms which estimate polynomial singular vectors from data derived from the channels characterized in chapter 5.

On a related note, even when the DTM of a channel does not admit weighted orthogonal polynomial singular vectors, we may still want to approximate the singular vectors using representative polynomials. It is usually rather difficult to explicitly compute spectral decompositions or SVDs of such DTMs because they are bounded linear operators on infinite-dimensional Hilbert spaces. This leads us to the topic of polynomial approximations of eigenfunctions of the Gramian operators of DTMs (which is the first step in computing SVDs of DTMs). A promising starting point in the literature of polynomial approximations of functions is the Weierstrass approximation theorem, which is presented next. A version of this theorem and its generalization, the Stone-Weierstrass theorem, can be found in the classic real analysis text [30].

Theorem 6.2.1 (Weierstrass Approximation Theorem)**.** *Given a continuous function $f : [a, b] \to \mathbb{R}$ on a compact interval, there exists a sequence of polynomial functions $p_n : [a, b] \to \mathbb{R}$, $n \in \mathbb{N}$, that converge uniformly to f on the interval $[a, b]$. Formally, we have:*

$$\lim_{n\to\infty} \sup_{x\in[a,b]} |p_n(x) - f(x)| = 0.$$

There are several proofs of the Weierstrass approximation theorem in the literature, and fortunately, some of these proofs are constructive. For instance, Bernstein polynomials can be used to uniformly approximate continuous functions on a compact (closed and bounded) interval in $\mathbb{R}$. This means that although the Weierstrass approximation theorem is an existence theorem, we can easily find polynomials to approximate continuous functions. However, our problem is more difficult than this because we do not have the explicit form of the eigenfunctions we wish to approximate. They are defined implicitly using the Gramian operator of the DTM. To make matters worse, we do not even know the Gramian operator, because we do not know any theoretical distribution. We simply have a large volume of data from which we must estimate the eigenfunctions of the Gramian operator. For simplicity, we consider the problem of estimating the eigenfunctions when the Gramian operator is known. So, we are essentially given the kernel of a (possibly compact) self-adjoint linear integral operator, and we seek to approximate the eigenfunctions of the operator with polynomials. Appealing to the theory of integral equations [37], we see that we must approximate solutions to a homogeneous Fredholm equation of the second kind. [37] states that computing eigenfunctions of integral operators is straightforward when we have degenerate kernels. A kernel, $\Lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, of an integral operator is called degenerate if it has the form:

$$\forall x, t \in \mathbb{R}, \;\; \Lambda(x,t) = \sum_{i=1}^{n} \phi_i(x)\psi_i(t)$$

for some $n \in \mathbb{Z}^+$, where without loss of generality, $\phi_i : \mathbb{R} \to \mathbb{R}$ are linearly independent and $\psi_i : \mathbb{R} \to \mathbb{R}$ are linearly independent. For a degenerate kernel, [37] proves that the eigenfunctions are linear combinations of ϕ_i, where the coefficients of the linear combinations can be found by solving systems of n linear equations (which is easily done using matrix methods). Moreover, it can be shown that the Weierstrass approximation theorem can be used to approximate any kernel using degenerate kernels where ϕ_i and ψ_i are polynomials. Chapter 11 of [38] provides a broad introduction to approximating the eigenfunctions of integral operators with arbitrary kernels by approximating these kernels with degenerate kernels. Exploring such polynomial approximations of eigenfunctions of the Gramian operators of DTMs is another viable direction of future research.

Until now, our discussion has entirely been about furthering the work in different chapters of this thesis. A much broader goal of future research is attempting to expand the horizons of local information theory. We hope that this can be achieved by relating it to quantum information theory. Quantum information theory studies communication channels which have inherently quantum attributes. For example, in a classical-quantum channel, the decoder is a positive operator valued measurement (POVM) which is a generalization of the von Neumann measurement in quantum mechanics. [39] provides a comprehensive introduction to quantum information theory and [40] provides a terse summary of its results and open problems.

Recent advances in our understanding of classical bounds on zero error capacity have come from quantum information theory. When Shannon introduced the zero error capacity problem in [41], it became apparent to information theorists that the problem had a highly combinatorial flavor. The best known general upper bound for zero error capacity is the Lovász theta function. Unsurprisingly, its proof is combinatorial. [42] correctly observes that there has been a divergence of techniques in information theory where traditional capacity problems are solved probabilistically while zero error information theory is attacked using combinatorics. The work in [42] suggests a

resolution to this divergence by using quantum probability. It provides a long-awaited interpretation of Lovász's bound and proceeds to prove the quantum sphere packing bound.

Curiously, quantum probability derives its strength from representing and interpreting probability distributions as unit norm complex vectors. So, while classical probability preserves the ℓ_1-norm, quantum probability preserves the ℓ_2-norm and effectively takes us to the domain of linear algebra. Fortunately, linear algebra is very well understood by the applied mathematics community. [42] portrays how this approach helps unify many aspects of information theory. On a seemingly different front, we are performing local approximations to solve information theory problems. Recall that one reason for performing these approximations is to simplify information theoretic problems meaningfully so that they are solvable using linear algebra techniques. Thus, at a high level, the quantum probability approach parallels the local approach. Exploring some concrete relations between the quantum and local information theory frameworks might be a truly fulfilling future endeavor.

Appendix A

Proof of MMSE Characterization of Rényi Correlation

In this section of the appendix, we provide a proof of Theorem 3.2.6. This theorem provides an MMSE characterization of Rényi correlation. We also restate the theorem below for convenience.

Theorem A.0.2 (MMSE Characterization of Rényi Correlation)**.** *Suppose we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with joint distribution $P_{X,Y}$ on this space, and Rényi correlation $\rho(X;Y)$. If the optimizing functions of the Rényi correlation are $f^* : \mathcal{X} \to \mathbb{R}$ and $g^* : \mathcal{Y} \to \mathbb{R}$, then f^* and g^* satisfy:*

$$\begin{aligned} \rho(X;Y) f^*(X) &= \mathbb{E}\left[g^*(Y)|X\right] \quad a.s. \\ \rho(X;Y) g^*(Y) &= \mathbb{E}\left[f^*(X)|Y\right] \quad a.s. \end{aligned}$$

where the equalities hold almost surely (with probability 1), and $\rho^2(X;Y) = \mathbb{E}\left[\mathbb{E}\left[f^*(X)|Y\right]^2\right] = \mathbb{E}\left[\mathbb{E}\left[g^*(Y)|X\right]^2\right]$.

Proof.
We use variational calculus assuming all appropriate regularity conditions. We also assume that pmfs or pdfs exist. A more general proof can be found in [19]. The Lagrangian which represents the optimization in Definition 3.2.1 of $\rho(X;Y)$ is given by:

$$L(f(x), g(y), \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \mathbb{E}\left[f(X)g(Y)\right] + \lambda_1 \mathbb{E}\left[f(X)\right] + \lambda_2 \mathbb{E}\left[g(Y)\right] + \lambda_3 \mathbb{E}\left[f^2(X)\right] + \lambda_4 \mathbb{E}\left[g^2(Y)\right]$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}$ are the Lagrange multipliers. We differentiate $L(f(x), g(y))$ with respect to $f(a)$ and $g(b)$ for any $a \in \mathcal{X}$ and $b \in \mathcal{Y}$:

$$\frac{\partial L}{\partial f(a)} = \mathbb{E}_{P_{X,Y}(a,\cdot)}\left[g(Y)\right] + \lambda_1 P_X(a) + 2\lambda_3 f(a) P_X(a)$$

$$\frac{\partial L}{\partial g(b)} = \mathbb{E}_{P_{X,Y}(\cdot,b)}\left[f(X)\right] + \lambda_2 P_Y(b) + 2\lambda_4 g(b) P_Y(b)$$

where $\mathbb{E}_{P_{X,Y}(a,\cdot)}\left[g(Y)\right] = \int_{\mathcal{Y}} P_{X,Y}(a,y) g(y) d\lambda(y)$ in the continuous case (λ denotes the Lebesgue measure and the integral is the Lebesgue integral), and $\mathbb{E}_{P_{X,Y}(a,\cdot)}\left[g(Y)\right] = \sum_{y \in \mathcal{Y}} P_{X,Y}(a,y) g(y)$ in

the discrete case. Replacing a and b with x and y, respectively, and setting these partial derivatives equal to 0 to find stationary points, we get:

$$\mathbb{E}\left[g(Y)|X=x\right]+\lambda_1+2\lambda_3 f(x)=0,$$
$$\mathbb{E}\left[f(X)|Y=y\right]+\lambda_2+2\lambda_4 g(y)=0.$$

Since the Lagrange multipliers are real constants, we rearrange these equations and re-label constants to get:

$$f(x)=a_1+a_2\mathbb{E}[g(Y)|X=x]$$
$$g(y)=b_1+b_2\mathbb{E}[f(X)|Y=y]$$

where the re-labeled Lagrange multipliers $a_1,a_2,b_1,b_2\in\mathbb{R}$ have values which satisfy the expectation and variance constraints in Definition 1. Imposing the constraints $\mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=0$, we have:

$$\mathbb{E}[f(X)]=a_1+a_2\mathbb{E}[\mathbb{E}[g(Y)|X]]=a_1+a_2\mathbb{E}[g(Y)]=a_1=0$$
$$\mathbb{E}[g(Y)]=b_1+b_2\mathbb{E}[\mathbb{E}[f(X)|Y]]=b_1+b_2\mathbb{E}[f(X)]=b_1=0$$

which gives:

$$f(x)=a_2\mathbb{E}[g(Y)|X=x],$$
$$g(y)=b_2\mathbb{E}[f(X)|Y=y].$$

Furthermore, imposing the constraints $\mathbb{E}\left[f^2(X)\right]=\mathbb{E}\left[g^2(Y)\right]=1$ gives:

$$\mathbb{E}\left[f^2(X)\right]=a_2^2\mathbb{E}\left[\mathbb{E}[g(Y)|X]^2\right]=1$$
$$\mathbb{E}\left[g^2(Y)\right]=b_2^2\mathbb{E}\left[\mathbb{E}[f(X)|Y]^2\right]=1$$

which implies:

$$a_2=\frac{1}{\sqrt{\mathbb{E}\left[\mathbb{E}[g(Y)|X]^2\right]}},$$
$$b_2=\frac{1}{\sqrt{\mathbb{E}\left[\mathbb{E}[f(X)|Y]^2\right]}}.$$

Plugging in the values of a_2 and b_2 into the expressions for $f(x)$ and $g(y)$, and re-labeling the optimizing functions as f^* and g^*, respectively, gives the equations:

$$\forall x\in\mathcal{X},\ \ \sqrt{\mathbb{E}\left[\mathbb{E}[g^*(Y)|X]^2\right]}f^*(x)=\mathbb{E}[g^*(Y)|X=x],$$
$$\forall y\in\mathcal{Y},\ \ \sqrt{\mathbb{E}\left[\mathbb{E}[f^*(X)|Y]^2\right]}g^*(y)=\mathbb{E}[f^*(X)|Y=y].$$

Since the uniqueness of the conditional expectation holds when equality between random variables is defined as equality with probability 1, the above equalities rigorously hold with probability 1. All that remains is to show that $\rho^2(X;Y)=\mathbb{E}\left[\mathbb{E}\left[f^*(X)|Y\right]^2\right]=\mathbb{E}\left[\mathbb{E}\left[g^*(Y)|X\right]^2\right]$. To this end, consider:

$$\rho^2(X;Y)=\mathbb{E}\left[f^*(X)g^*(Y)\right]^2=\mathbb{E}\left[\mathbb{E}\left[f^*(X)g^*(Y)|Y\right]\right]^2=\mathbb{E}\left[g^*(Y)\mathbb{E}\left[f^*(X)|Y\right]\right]^2$$

which, using $\sqrt{\mathbb{E}\left[\mathbb{E}[f^*(X)|Y]^2\right]}g^*(y)=\mathbb{E}[f^*(X)|Y=y]$, becomes:

$$\rho^2(X;Y)=\mathbb{E}\left[g^*(Y)^2\sqrt{\mathbb{E}\left[\mathbb{E}[f^*(X)|Y]^2\right]}\right]^2=\mathbb{E}\left[\mathbb{E}[f^*(X)|Y]^2\right]\mathbb{E}\left[g^*(Y)^2\right]^2=\mathbb{E}\left[\mathbb{E}[f^*(X)|Y]^2\right]$$

where the last equality holds because $\mathbb{E}\left[g^*(Y)^2\right]=1$. Likewise, $\rho^2(X;Y)=\mathbb{E}\left[\mathbb{E}\left[g^*(Y)|X\right]^2\right]$. This completes the proof. □

Bibliography

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Cambridge University Press, second ed., 2011.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New Jersey: John Wiley & Sons, Inc., second ed., 2006.

[3] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, Inc., 1968.

[4] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, (Cambridge, MA, USA), pp. 1029–1033, July 1-6 2012.

[5] S. Borade and L. Zheng, "Euclidean information theory," in *IEEE International Zurich Seminar on Communications*, pp. 14–17, March 12-14 2008.

[6] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July-October 1948.

[7] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, (Allerton House, UIUC, Illinois, USA), pp. 368–377, September 1999.

[8] S.-L. Huang, A. Makur, F. Kozynski, and L. Zheng, "Efficient statistics: Extracting information from iid observations," in *Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing*, (Allerton House, UIUC, Illinois, USA), October 1-3 2014.

[9] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences." Submitted to 2015 IEEE Information Theory Workshop, Jerusalem, Israel, arXiv:1502.06428 [cs.IT], March 2015.

[10] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, vol. 1 of *Foundations and Trends in Communications and Information Theory*. Hanover: now Publishers Inc., 2004.

[11] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, October 2005.

[12] J. Stewart, *Calculus: Early Transcendentals*. Belmont: Cengage Learning, sixth ed., 2008.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[14] V. Anantharam, A. Gohari, S. Kamath, and C. Nair., "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover." arXiv:1304.6133 [cs.IT], April 2013.

[15] R. O'Donnell, *Analysis of Boolean Functions*. New York: Cambridge University Press, 2014.

[16] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between boolean functions," in *Proceedings of the 51st Annual Allerton Conference on Communication, Control and Computing*, (Allerton House, UIUC, Illinois, USA), pp. 13–19, October 2-4 2013.

[17] S. Kamath and V. Anantharam, "Non-interactive simulation of joint distributions: The hirschfeld-gebelein-rényi maximal correlation and the hypercontractivity ribbon," in *Proceedings of the 50th Annual Allerton Conference on Communication, Control and Computing*, (Allerton House, UIUC, Illinois, USA), pp. 1057–1064, October 1-5 2012.

[18] G. R. Kumar and T. A. Courtade, "Which boolean functions are most informative?," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, (Istanbul, Turkey), pp. 226–230, July 7-12 2013.

[19] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[20] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM Journal on Applied Mathematics*, vol. 28, pp. 100–113, January 1975.

[21] F. P. Calmon, M. Varia, M. Médard, M. M. Christiansen, K. R. Duffy, and S. Tessaro, "Bounds on inference," in *Proceedings of the 51st Annual Allerton Conference on Communication, Control and Computing*, (Allerton House, UIUC, Illinois, USA), pp. 567–574, October 2-4 2013.

[22] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Transactions on Information Theory*, vol. 44, pp. 1026–1040, May 1998.

[23] E. Abbe and L. Zheng, "A coordinate system for gaussian networks," *IEEE Transactions on Information Theory*, vol. 58, pp. 721–733, February 2012.

[24] Y.-C. Li and C.-C. Yeh, "Some equivalent forms of bernoullis inequality: A survey," *Applied Mathematics*, vol. 4, no. 7, pp. 1070–1093, 2013.

[25] R. W. Keener, *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics, New York: Springer, 2010.

[26] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38 of *Stochastic Modelling and Applied Probability*. New York: Springer, second ed., 1998.

[27] W. Forst and D. Hoffmann, *Optimization - Theory and Practice*. Springer Undergraduate Texts in Mathematics and Technology, New York: Springer, 2010.

[28] E. M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, vol. 3 of *Princeton Lectures in Analysis*. New Jersey: Princeton University Press, 2005.

[29] R. E. Megginson, *An Introduction to Banach Space Theory*. Graduate Texts in Mathematics, New York: Springer, October 1998.

[30] W. Rudin, *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics, New York: McGraw-Hill, Inc., third ed., 1976.

[31] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, pp. 580–598, September 1985.

[32] C. N. Morris, "Natural exponential families with quadratic variance functions," *The Annals of Statistics*, vol. 10, no. 1, pp. 65–80, 1982.

[33] D. Fink, "A compendium of conjugate priors." Environmental Statistics Group, Department of Biology, Montana State Univeristy, May 1997.

[34] G. E. Andrews and R. Askey, *Polynômes Orthogonaux et Applications*, vol. 1171 of *Lecture Notes in Mathematics*, ch. Classical Orthogonal Polynomials, pp. 36–62. Springer, 1985.

[35] T. S. Chihara, *An Introduction to Orthogonal Polynomials*. New York: Dover Publications, dover reprint ed., 2011.

[36] P. Diaconis, K. Khare, and L. Saloff-Coste, "Gibbs sampling, exponential families and orthogonal polynomials," *Statisical Science*, vol. 23, no. 2, pp. 151–178, 2008.

[37] F. Porter, "Integral equations." Caltech Physics 129a Online Notes, November 2007.

[38] R. Kress, *Linear Integral Equations*, vol. 82 of *Applied Mathematical Sciences*. New York: Springer, third ed., 2014.

[39] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. New York: Cambridge University Press, tenth anniversary ed., 2010.

[40] P. Shor, *Visions in Mathematics, Geometric and Functional Analysis 2000 Special Volume, Part II*, ch. Quantum Information Theory: Results and Open Problems, pp. 816–838. Modern Birkhäuser Classics, Birkhäuser Basel, 2010.

[41] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, pp. 706–715, September 1956.

[42] M. Dalai, "Lower bounds on the probability of error for classical and classical-quantum channels," *IEEE Transactions on Information Theory*, vol. 59, pp. 8027–8056, December 2013.