Bounds between contraction coefficients

Bounds between Contraction Coefficients

Anuran Makur and Lizhong Zheng
EECS Department, Massachusetts Institute of Technology
Email: {a_makur, lizhong}@mit.edu

Abstract

In this paper, we delineate how the contraction coefficient of the strong data processing inequality for KL divergence can be used to learn likelihood models. We then present an alternative formulation to learn likelihood models that forces the input KL divergence of the data processing inequality to vanish, and achieves a contraction coefficient equivalent to the squared maximal correlation. This formulation turns out to admit a linear algebraic solution. To analyze the performance loss in using this simple but suboptimal procedure, we bound these contraction coefficients in the discrete and finite regime, and prove their equivalence in the Gaussian regime.

CONTENTS

I Introduction 1
- I-A Local Approximations of KL Divergence . . . 1
- I-B Learning Likelihood Models using Contraction Coefficients . . . 2
- I-C Bounding the Performance Loss . . . 5

II Bounds in the Discrete and Finite Setting 5
- II-A Bounds on KL Divergence . . . 6
- II-B Bounds on Global Contraction Coefficient . . . 9

III Equivalence in the Gaussian Setting 11
- III-A Maximal Correlation of AWGN Channel . . . 12
- III-B Global Contraction Coefficients of AWGN Channel . . . 12

IV Conclusion 17

References 17

I. INTRODUCTION

Strong data processing inequalities (SDPIs) for Kullback-Leibler (KL) divergence and mutual information [1]–[5], and more generally f-divergences [6], have been studied extensively in various contexts in information theory. The contraction coefficients of such strong data processing inequalities can serve as convenient variational representations of certain statistical learning problems. We introduce such an inference problem, elucidate its relation to the contraction coefficients for KL and χ^2-divergences, and derive bounds between these contraction coefficients to provide performance guarantees.

A. Local Approximations of KL Divergence

We commence our discussion with a brief prelude on local approximations of KL divergence, because such approximations will underlie our learning approach. Moreover, such approximations are geometrically appealing because they transform neighborhoods of stochastic manifolds, with KL divergence as the distance measure, into inner product spaces. To recognize this, consider a discrete and finite sample space $\Omega = \{1, \ldots, n\}$, and probability mass functions (pmfs) $P = [P(1) \cdots P(n)]^T$ and $Q = [Q(1) \cdots Q(n)]^T$ on Ω, which can be construed as vectors in $\mathbb{R}^n$. Let us arbitrarily fix P as the reference pmf which is in the relative interior of the probability simplex in $\mathbb{R}^n$ (of pmfs on Ω): $\forall x \in \Omega,\ P(x) > 0$. This allows us to consider a local neighborhood of pmfs around P, and we assume that Q is in this neighborhood in the sense that:

$$Q = P + \epsilon J \tag{1}$$

where $J = [J(1) \cdots J(n)]^T$ is an additive perturbation vector (which provides the direction of perturbation) satisfying:

$$\sum_{x \in \Omega} J(x) = 0 \tag{2}$$

and $\epsilon \neq 0$ (which controls how close P and Q are) is sufficiently small to ensure that Q is a valid pmf: $\forall x \in \Omega,\ 0 \leq Q(x) \leq 1$. In our ensuing discussion, we simply assume this condition on ϵ holds without explicitly imposing it.

Using second order Taylor expansions of the natural logarithm function, we can show that KL divergence is locally a weighted Euclidean metric:

$$D(Q||P) = \frac{1}{2}\epsilon^2 \sum_{x \in \Omega} \frac{J^2(x)}{P(x)} + o\left(\epsilon^2\right) = \frac{1}{2}\chi^2\left(Q, P\right) + o\left(\epsilon^2\right) \tag{3}$$

where $o\left(\epsilon^2\right)$ denotes a function satisfying: $\lim_{\epsilon \to 0} o\left(\epsilon^2\right)/\epsilon^2 = 0$, and $\chi^2\left(Q, P\right)$ denotes the χ^2-divergence between Q and P:

$$\chi^2\left(Q, P\right) \triangleq \sum_{x \in \Omega} \frac{\left(Q(x) - P(x)\right)^2}{P(x)} \tag{4}$$

which is closely related to χ^2-tests in statistics. Equation (3) portrays that KL divergence is locally proportional to the χ^2-divergence [7].

Compelled by (3), we may define an alternative *spherical perturbation* vector which permits us to use standard Euclidean norms:

$$K \triangleq \left[\sqrt{P}\right]^{-1} J \tag{5}$$

where $\sqrt{P}$ denotes the elementwise square root of P, and for any vector $x \in \mathbb{R}^n$, we let $[x]$ denote the $n \times n$ diagonal matrix with entries of x along its principal diagonal. K satisfies the orthogonality constraint: $K^T\sqrt{P} = 0$, which is equivalent to (2). It is called a spherical perturbation vector because it is the first order perturbation term of $\sqrt{Q}$ from $\sqrt{P}$, which are embeddings of pmfs as vectors on the unit sphere in $\mathbb{R}^n$. Using (5), we may recast (1) into:

$$Q = P + \epsilon\left[\sqrt{P}\right] K \tag{6}$$

and the expression for local KL divergence in (3) into:

$$D(Q||P) = \frac{1}{2}\epsilon^2 \left\|K\right\|_2^2 + o\left(\epsilon^2\right) = D(P||Q) \tag{7}$$

where $\left\|\cdot\right\|_2$ is the standard Euclidean ℓ^2-norm, and (7) also illustrates that KL divergence is locally symmetric. Hence, the KL divergence resembles an Euclidean metric within a neighborhood of pmfs around an arbitrary reference pmf in the relative interior of the probability simplex. Furthermore, it is easily verified that additive and spherical perturbations form isomorphic inner product spaces. In particular, the inner product space of spherical perturbations is equipped with the standard Euclidean inner product.

B. Learning Likelihood Models using Contraction Coefficients

We now introduce a seemingly pedestrian inference problem. Suppose we want to infer some hidden variable U about an individual based on the some data $(Y_1, \ldots, Y_m)$ attributed to him, where each $Y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is a discrete and finite set. For instance, U might be the individual's political affiliation, and $(Y_1, \ldots, Y_m)$ might be the list of movies he has watched over a period of time. We assume for simplicity that $U \sim$ Rademacher i.e. $U \in \mathcal{U} = \{-1, 1\}$ and $\mathbb{P}(U = 1) = 0.5$, and that $(Y_1, \ldots, Y_m)$ are conditionally independent given U. If we know the conditional distributions $P_{Y|U}$ from which $(Y_1, \ldots, Y_m)$ are generated given U, then inferring U is simply a binary hypothesis testing problem. We construct the log-likelihood ratio sufficient statistic for U:

$$Z \triangleq \sum_{i=1}^{m} \log\left(\frac{P_{Y|U}\left(Y_i|1\right)}{P_{Y|U}\left(Y_i|-1\right)}\right) \tag{8}$$

and the log-likelihood ratio test with a threshold of 0 (maximum likelihood) corresponds to the minimum probability of error estimator $\widehat{U} = \operatorname{sgn}(Z)$, where $\operatorname{sgn}(\cdot)$ is the signum function.

If $P_{Y|U}$ is unknown, we can pose a more intriguing unsupervised model selection problem which finds the "optimal" $P_{Y|U}$ from training data, $(X_1, Y_1), \ldots, (X_n, Y_n)$. Here, each $X_i \in \mathcal{X}$ and $\mathcal{X}$ is a discrete and finite set of indexes for different people, each $Y_i \in \mathcal{Y}$ as before, and a single sample of data, (X_i, Y_i), conveys that person X_i watched movie Y_i. The data is "unlabeled" as we do not observe the hidden variables $U_i \in \mathcal{U}$ corresponding to each X_i. We assume that the data is generated i.i.d. from a (marginalized) joint distribution $P_{U,X,Y} = P_U P_{X|U} P_{Y|X}$, and that $P_{X,Y} = \widehat{P}_{X_1^n, Y_1^n}$, where $\widehat{P}_{X_1^n, Y_1^n}$ is the empirical distribution of the training data. The former assumption is standard practice in statistics, and the latter is motivated by concentration of measure results like Sanov's theorem. Hence, our problem reduces to finding the "optimal" $P_{X|U}$, where $U \to X \to Y$ is a Markov chain, and P_U and $P_{X,Y}$ are fixed.

To find an appropriate optimization criterion, we introduce SDPIs. Recall the data processing inequalities (DPIs) for KL divergence and mutual information [7]. For a general channel (Markov kernel) $P_{Y|X}$:

$$D(R_Y||P_Y) \leq D(R_X||P_X) \tag{9}$$

where P_X and R_X are input distributions, and P_Y and R_Y are the corresponding output distributions. Likewise, if the (general) random variables $U \to X \to Y$ form a Markov chain, then:

$$I(U;Y) \leq I(U;X). \tag{10}$$

For fixed P_X and $P_{Y|X}$, we can tighten such DPIs to produce SDPIs by inserting pertinent contraction coefficients; these are defined next.

Definition 1 (Global Contraction Coefficient)**.** For random variables X and Y with joint distribution $P_{X,Y}$, the global contraction coefficient is given by:

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \triangleq \sup_{\substack{R_X : R_X \neq P_X \\ 0 < D(R_X||P_X) < \infty}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where R_Y is the marginal distribution of $R_{X,Y} = P_{Y|X}R_X$. If X or Y is a constant almost surely, we define $\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = 0$.

Theorem 1 (Contraction Coefficient for Mutual Information [2], [3], [6])**.** *If the random variables $U \to X \to Y$ form a Markov chain such that the joint distribution $P_{X,Y}$ is fixed, then we have:*

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \sup_{\substack{P_U, P_{X|U}: \\ U \to X \to Y \\ 0 < I(U;X) < \infty}} \frac{I(U;Y)}{I(U;X)} = \lim_{\epsilon \to 0} \sup_{\substack{P_U, P_{X|U}: \\ U \to X \to Y \\ I(U;X) = \frac{1}{2}\epsilon^2}} \frac{I(U;Y)}{I(U;X)}.$$

$\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ is called the "global" contraction coefficient to distinguish it from Definition 2 where we use local approximations. It is related to the notion of hypercontractivity in statistics [1], which is useful in studying extremal problems in probability spaces with distance measures, and has found many applications in information theory. Indeed, $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ is also the chordal slope of the lower boundary of the hypercontractivity ribbon at infinity in the discrete and finite case [4]. Theorem 1 illustrates that $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ gracefully unifies the SDPIs for KL divergence and mutual information.

The variational problem posed by $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ determines the probability model that maximizes the flow of information down the Markov chain $U \to X \to Y$, and meaningfully addresses our model selection problem (neglecting the $U \sim$ Rademacher assumption). Given $\mathcal{U} = \{0,1\}$, $X \sim$ Bernoulli(0.5), and an asymmetric erasure channel $P_{Y|X}$, the numerical example in [3] conveys that the supremum of the ratio $I(U;Y)/I(U;X)$ is achieved by a sequence of distributions, $P^k_{X|U=0}$, $P^k_{X|U=1}$, P^k_U, $k \in \mathbb{N}$, satisfying the following: $P^k_U(1) \to 0$, $D(P^k_{X|U=0}||P_X) \to 0$, $\liminf_k D(P^k_{X|U=1}||P_X) > 0$, $\limsup_k D(P^k_{Y|U=0}||P_Y)/D(P^k_{X|U=0}||P_X) < \eta_{\mathsf{glo}}(P_X, P_{Y|X})$, and $D(P^k_{Y|U=1}||P_Y)/D(P^k_{X|U=1}||P_X) \to \eta_{\mathsf{glo}}(P_X, P_{Y|X})$, as $k \to \infty$. Thus, although $I(U;Y)/I(U;X)$ is maximized when $I(U;X) \to 0$ [2], $D(R_Y||P_Y)/D(R_X||P_X)$ is typically maximized by a sequence $\{R^k_X \neq P_X : k \in \mathbb{N}\}$ that does not tend to P_X due to the non-concave nature of this extremal problem. Moreover, as the optimization problems in Definition 1 and Theorem 1 are highly non-concave, they are onerous to solve. Since (9) is tight when $R_X = P_X$, we instead maximize $D(R_Y||P_Y)/D(R_X||P_X)$ over all sequences $\{R^k_X \neq P_X : k \in \mathbb{N}\}$ satisfying $R^k_X \to P_X$ (in ℓ_p-norm) or $D(R^k_X||P_X) \to 0$, or equivalently, solve the optimization problem in Theorem 1 with the additional constraint that $U \sim$ Rademacher. This formulation turns out to admit a simple linear algebraic solution.

To formally present this alternative formulation, we assume that $\forall x \in \mathcal{X}$, $P_X(x) > 0$, and $\forall y \in \mathcal{Y}$, $P_Y(y) > 0$, and let W be the $|\mathcal{Y}| \times |\mathcal{X}|$ column stochastic transition probability matrix of conditional pmfs $P_{Y|X}$. We also define a trajectory of spherically perturbed pmfs parametrized by ϵ: $R^{(\epsilon)}_X = P_X + \epsilon \left[\sqrt{P_X}\right] K_X$, for some fixed vector $K_X \neq \vec{0}$ (zero vector) satisfying $K^T_X \sqrt{P_X} = 0$. When $R^{(\epsilon)}_X$ passes through the channel W, we get:

$$R^{(\epsilon)}_Y = W R^{(\epsilon)}_X = P_Y + \epsilon \left[\sqrt{P_Y}\right] B K_X \tag{11}$$

where $B \triangleq \left[\sqrt{P_Y}\right]^{-1} W \left[\sqrt{P_X}\right]$ is the *divergence transition matrix* (DTM) which maps input spherical perturbations to output spherical perturbations [8]. Using the trajectory $R^{(\epsilon)}_X$, we will prove that our new formulation is equivalent to the extremal problem defined next.

Definition 2 (Local Contraction Coefficient)**.** For random variables X and Y with joint pmf $P_{X,Y}$ and corresponding DTM B, the local contraction coefficient is given by:

$$\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right) \triangleq \sup_{\substack{R_X: \\ R_X \neq P_X}} \frac{\chi^2\left(R_Y, P_Y\right)}{\chi^2\left(R_X, P_X\right)} = \sup_{\substack{K_X: K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\left\|B K_X\right\|_2^2}{\left\|K_X\right\|_2^2}$$

where R_Y is the marginal pmf of the joint pmf $R_{X,Y} = P_{Y|X} R_X$. If X or Y is a constant almost surely, we define $\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right) = 0$.

Theorem 2 (Contraction Coefficient for Local Mutual Information [8], [9])**.** *If the random variables $U \to X \to Y$ form a Markov chain such that the joint pmf $P_{X,Y}$ is fixed, then we have:*

$$\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right) = \sup_{\substack{P_U, \left\{K_u \neq \vec{0}: u \in \mathcal{U}\right\}: \\ \forall u \in \mathcal{U},\, K_u^T \sqrt{P_X} = 0 \\ \sum_{u \in \mathcal{U}} P_U(u) \left[\sqrt{P_X}\right] K_u = \vec{0}}} \frac{\sum_{u \in \mathcal{U}} P_U(u) \left\|B K_u\right\|_2^2}{\sum_{u \in \mathcal{U}} P_U(u) \left\|K_u\right\|_2^2}$$

where B is the DTM corresponding to $P_{X,Y}$, the last constraint ensures that: $\sum_{u \in \mathcal{U}} P_U(u) P_{X|U=u} = P_X$, where $\forall u \in \mathcal{U}$, $P_{X|U=u} = P_X + \left[\sqrt{P_X}\right] K_u$, and $U \sim$ Rademacher without loss of generality.

$\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right)$ is called the "local" contraction coefficient because it characterizes the SDPIs for χ^2-divergence or local KL divergence (7), and local mutual information. It equals the squared second largest singular value of B by the Courant-Fischer theorem, because the largest singular value of B is 1 (as B originates from the stochastic matrix W) with right singular vector $\sqrt{P_X}$ and left singular vector $\sqrt{P_Y}$. Theorem 2 is an analog of Theorem 1, and the extremal problem in Theorem 2 is equivalent to the linear information coupling problem [8], which was developed to enable single letterization in network capacity problems by transforming them into linear algebra problems using local approximations and then applying tensor algebra arguments. Indeed, this extremal problem resembles the first extremal problem in Theorem 1 under local approximations (7). $\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right)$ is also related to a generalization of the Pearson correlation coefficient known as the Hirschfeld-Gebelein-Rényi maximal correlation [10]. We next define the maximal correlation, which was proven to be the unique measure of statistical dependence between random variables satisfying seven natural axioms (some of which are in Proposition 4) that measures of dependence should exhibit [10].

Definition 3 (Maximal Correlation)**.** For any two jointly distributed random variables X and Y, with ranges $\mathcal{X}$ and $\mathcal{Y}$ respectively, the maximal correlation between X and Y is given by:

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

where the supremum is taken over all Borel measurable functions. If X or Y is a constant almost surely, there exist no functions f and g which satisfy the constraints, and we define $\rho(X;Y) = 0$.

The next theorem demonstrates that $\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right)$, $\rho^2(X;Y)$, and the supremum of $D(R_Y||P_Y)/D(R_X||P_X)$ as $D(R_X||P_X) \to 0$, are all equivalent. Furthermore, the trajectory $R_X^{(\epsilon)}$ achieves the aforementioned supremum with a particular choice of K_X.

Theorem 3 (Characterizations of Local Contraction Coefficient)**.** *For random variables X and Y with joint pmf $P_{X,Y}$, we have:*

$$\eta_{\mathsf{loc}}\left(P_X, P_{Y|X}\right) = \rho^2(X;Y) = \lim_{\epsilon \to 0} \sup_{\substack{R_X: R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where R_Y is the marginal pmf of the joint pmf $R_{X,Y} = P_{Y|X} R_X$.

Proof. The first equality follows from interpreting Definition 3 as the Courant-Fischer characterization of a singular value [3]. To prove the second equality, observe that for sufficiently small $\epsilon \neq 0$, we have:

$$\begin{aligned}\sup_{\substack{R_X: R_X \neq P_X \\ D(R_X||P_X)=\frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} &= \sup_{\substack{K_X: K_X^T\sqrt{P_X}=0 \\ \|K_X\|_2^2+o(1)=1}} \frac{\|BK_X\|_2^2+o(1)}{\|K_X\|_2^2+o(1)} \\ &= \sup_{\substack{K_X: K_X^T\sqrt{P_X}=0 \\ \|K_X\|_2^2=1+o(1)}} \frac{\frac{\|BK_X\|_2^2}{\|K_X\|_2^2}+o(1)}{1+o(1)} \\ &= \frac{\eta_{\mathsf{loc}}(P_X, P_{Y|X})+o(1)}{1+o(1)}\end{aligned}$$

where the first equality follows from using the trajectory $R_X^{(\epsilon)}$ (as every R_X can be decomposed in this way), (7), and (11), the third equality follows from Definition 2 after applying the squeeze theorem to pull out the $o(1)$ terms, and $o(1)$ denotes a function satisfying $\lim_{\epsilon\to 0} o(1) = 0$ uniformly with respect to K_X. Letting $\epsilon \to 0$ produces the second equality in Theorem 3. ■

Since $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ can be strictly greater than $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$ [3], Theorem 3 starkly contrasts Theorem 1 because $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ is achieved when $I(U;X) \to 0$ in Theorem 1, but often not achieved when $D(R_X||P_X) \to 0$ in Definition 1. The extremal problems in Theorem 3 are readily solved using the singular value decomposition (SVD). Indeed, the optimal spherical perturbation $K_X^\star$ that achieves $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$ in Definition 2 is the normalized right singular vector of the DTM corresponding to its second largest singular value. Equivalently, the trajectory $\widehat{R}_X^{(\epsilon)} = P_X + \epsilon\left[\sqrt{P_X}\right] K_X^\star$ maximizes the ratio $D(R_Y||P_Y)/D(R_X||P_X)$ as $\epsilon \to 0$ under the alternative formulation where we require that $D(R_X||P_X) \to 0$.

Therefore, under this formulation, the optimal $P_{X|U}$ for our model selection problem is given by: $P_{X|U=u}^\star = P_X + u\epsilon\left[\sqrt{P_X}\right] K_X^\star$ for some fixed small $\epsilon \neq 0$, where $u \in \mathcal{U} = \{1, -1\}$. We close this subsection by presenting some common properties of contraction coefficients, many of which hold for general random variables.

Proposition 4 (Contraction Coefficient Properties [3]–[5], [9], [10])**.** *The global and local contraction coefficients, simultaneously denoted $\eta(P_X, P_{Y|X})$ for simplicity, satisfy the following properties:*

1) *(Normalization)* $0 \leq \eta(P_X, P_{Y|X}) \leq 1$.
2) *(Tensorization) If (X_1, Y_1) is independent of (X_2, Y_2), then:*
$\eta(P_{X_1,X_2}, P_{Y_1,Y_2|X_1,X_2}) = \max\{\eta(P_{X_1}, P_{Y_1|X_1}), \eta(P_{X_2}, P_{Y_2|X_2})\}$.
3) *(Monotonicity) For random variables (X, Y), if $W = r(X)$ and $Z = s(Y)$ for some Borel measurable functions $r: \mathcal{X} \to \mathbb{R}$ and $s: \mathcal{Y} \to \mathbb{R}$, then $\eta(P_X, P_{Y|X}) \geq \eta(P_W, P_{Z|W})$.*
4) *(Vanishing)* $X \perp\!\!\!\perp Y \Leftrightarrow \eta(P_X, P_{Y|X}) = 0$.

C. Bounding the Performance Loss

We have illustrated that although the trajectory of distributions achieving $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ globally solves our model selection problem in a data processing sense, it is significantly easier to compute the trajectory $\widehat{R}_X^{(\epsilon)}$ that achieves $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$. This alternative approach is carefully analyzed in [11], where singular vectors of the DTM are identified with informative score functions, and variants of the alternating conditional expectations algorithm to compute the SVD of the DTM are presented. To estimate the performance loss in using $\widehat{R}_X^{(\epsilon)}$, we are propelled to bound $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ above and below by $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$. Section II presents such bounds in the discrete and finite regime by bounding KL divergence using χ^2-divergence, and Section III demonstrates that $\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \eta_{\mathsf{loc}}(P_X, P_{Y|X})$ for additive white Gaussian noise (AWGN) channels.

II. Bounds in the Discrete and Finite Setting

Recall that we are considering discrete random variables X and Y with finite ranges $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint pmf $P_{X,Y}$ such that $\forall x \in \mathcal{X}$, $P_X(x) > 0$, and $\forall y \in \mathcal{Y}$, $P_Y(y) > 0$. In this discrete and finite setting, locally approximating the objective functions in Definition 1 or Theorem 1 produces the following lower bound.

Theorem 5 (Lower Bound on Global Contraction Coefficient)**.**

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \geq \eta_{\mathsf{loc}}(P_X, P_{Y|X}) .$$

Proof. This result is well-known in the literature [5], and follows trivially from Theorem 3. We now offer an alternative direct proof. Starting from Definition 1 of $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$, we have:

$$\sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} \geq \lim_{\epsilon \to 0} \sup_{\substack{K_X : \|K_X\|_2 = 1 \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2 + o(1)}{\|K_X\|_2^2 + o(1)}$$

$$\geq \sup_{\substack{K_X : \|K_X\|_2 = 1 \\ K_X^T \sqrt{P_X} = 0}} \lim_{\epsilon \to 0} \frac{\|BK_X\|_2^2 + o(1)}{\|K_X\|_2^2 + o(1)}$$

where the first inequality follows from restricting the supremum over all pmfs $R_X^{(\epsilon)}$ such that $\|K_X\|_2 = 1$ and ϵ is some small enough fixed value, and then letting $\epsilon \to 0$, and the second inequality follows from the minimax inequality (by interpreting the $\lim$ as a $\liminf$). The theorem statement then follows from Definition 2. ■

This argument, mutatis mutandis, also proves that $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$ lower bounds the contraction coefficients for any f-divergence where $f''(1) > 0$ exists for the convex function $f : (0, \infty) \to \mathbb{R}$. The inequality in Theorem 5 is tight. For instance, $\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \eta_{\mathsf{loc}}(P_X, P_{Y|X}) = (1 - 2\alpha)^2$ for a doubly symmetric binary source with parameter $\alpha \in [0, 1]$, which is a joint pmf of uniform Bernoulli random variables (X, Y) where X is passed through a binary symmetric channel with crossover probability α to produce Y [1]. As another example, $\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \eta_{\mathsf{loc}}(P_X, P_{Y|X}) = 1 - \beta$ for a binary erasure channel with erasure probability $\beta \in [0, 1]$ regardless of the input Bernoulli distribution. This can be proven using brute force computation of the two contraction coefficients.

A. Bounds on KL Divergence

To upper bound $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ using $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$, we must first upper and lower bound the KL divergence using χ^2-divergence or local KL divergence. The next lemma presents one such lower bound on KL divergence which we will eventually tighten.

Lemma 6 (KL Divergence Lower Bound)**.** *Given pmfs P_X and R_X on $\mathcal{X}$, such that $\forall x \in \mathcal{X}$, $P_X(x) > 0$, and $R_X = P_X + J_X = P_X + \left[\sqrt{P_X}\right] K_X$, where J_X and K_X are additive and spherical perturbations, we have:*

$$D(R_X||P_X) \geq \frac{\min_{x \in \mathcal{X}} P_X(x)}{2} \|K_X\|_2^2 = \frac{\min_{x \in \mathcal{X}} P_X(x)}{2} \chi^2(R_X, P_X).$$

Proof. Since we essentially want to bound KL divergence using a norm, we can use Pinsker's inequality [7], [12], which lower bounds KL divergence using the total variation distance, or the ℓ^1-norm when $\mathcal{X}$ is discrete and finite. Starting from Pinsker's inequality, we have:

$$D(R_X||P_X) \geq \frac{1}{2} \|J_X\|_1^2 \geq \frac{1}{2} \|J_X\|_2^2 \geq \frac{\min_{x \in \mathcal{X}} P_X(x)}{2} \|K_X\|_2^2$$

where the second inequality holds because the ℓ^1-norm of a finite dimensional vector is greater than or equal to its ℓ^2-norm, and the third inequality holds by the definition of the spectral norm:

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

where $1/\min_{x \in \mathcal{X}} P_X(x)$ is the squared largest eigenvalue (or equivalently, squared spectral norm) of $\left[\sqrt{P_X}\right]^{-1}$. Recognizing that $\|K_X\|_2^2 = \chi^2(R_X, P_X)$ completes the proof. ■

This proof is statistical in flavor. We provide an alternative proof of Lemma 6 which has a more convex analysis flavor.

Proof. Let $\mathcal{P} \subseteq \mathbb{R}^{|\mathcal{X}|}$ be the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$, and let relint$(\mathcal{P})$ be the relative interior of $\mathcal{P}$. Furthermore, let $H_n : \mathcal{P} \to \mathbb{R}$ be the negative Shannon entropy function which is defined as:

$$\forall P = \begin{bmatrix} p_1 & \dots & p_{|\mathcal{X}|} \end{bmatrix}^T \in \mathcal{P}, \;\; H_n(P) \triangleq \sum_{i=1}^{|\mathcal{X}|} p_i \log(p_i)$$

where $\log(\cdot)$ denotes the natural logarithm. Since the Bregman divergence corresponding to H_n is the KL divergence [13], we have $\forall R_X \in \mathcal{P},\ \forall P_X \in \text{relint}(\mathcal{P})$:

$$D(R_X||P_X) = H_n(R_X) - H_n(P_X) - \nabla H_n(P_X)^T(R_X - P_X).$$

As H_n is twice continuously differentiable on relint$(\mathcal{P})$, we have: $\forall P \in \text{relint}(\mathcal{P})$, $\nabla^2 H_n(P) = [P]^{-1} \succeq I$, where $\succeq$ denotes the Löwner partial order, which means $[P]^{-1} - I$ is positive semidefinite, and I denotes the identity matrix. So, H_n is a strongly convex function on relint$(\mathcal{P})$, which is equivalent to the existence of the following quadratic lower bound [14]: $\forall R_X \in \mathcal{P},\ \forall P_X \in \text{relint}(\mathcal{P})$,

$$H_n(R_X) \geq H_n(P_X) + \nabla H_n(P_X)^T J_X + \frac{1}{2}\left\|J_X\right\|_2^2$$

where we allow $R_X \in \mathcal{P}\backslash\text{relint}(\mathcal{P})$ due to the continuity of H_n, and J_X is the additive perturbation between R_X and P_X. This gives us:

$$\forall R_X \in \mathcal{P}, \forall P_X \in \text{relint}(\mathcal{P}),\ \ D(R_X||P_X) \geq \frac{1}{2}\left\|J_X\right\|_2^2$$

which is precisely what we had in the previous proof after loosening Pinsker's inequality. Hence, the remainder of this proof is identical to the previous proof. ■

We note that such a convexity based approach cannot be used to easily derive an upper bound on KL divergence. It is well-known that if $\exists r > 0$ such that $\forall P \in \text{relint}(\mathcal{P})$, $\nabla^2 H_n(P) \preceq rI$, or equivalently, if ∇H_n is Lipschitz continuous on relint$(\mathcal{P})$, then a quadratic upper bound on H_n can be derived [14]. Unfortunately, the natural logarithm is not Lipschitz continuous on the domain $(0, \infty)$. We next present a tighter variant of Lemma 6 which does not slacken the ℓ^1-norm in Pinsker's inequality using an ℓ^2-norm.

Lemma 7 (KL Divergence Tighter Lower Bound)**.** *Given distinct pmfs P_X and R_X on $\mathcal{X}$, such that $\forall x \in \mathcal{X}$, $P_X(x) > 0$, and $R_X = P_X + J_X = P_X + \left[\sqrt{P_X}\right] K_X$, we have:*

$$D(R_X||P_X) \geq \frac{\left\|K_X\right\|_2^2 \left\|J_X\right\|_1}{2 \max\limits_{x\in\mathcal{X}} \left|\dfrac{J_X(x)}{P_X(x)}\right|} = \frac{\chi^2\left(R_X, P_X\right) \left\|J_X\right\|_1}{2 \max\limits_{x\in\mathcal{X}} \left|\dfrac{J_X(x)}{P_X(x)}\right|}.$$

Proof. Again starting from Pinsker's inequality [7], [12], we have:

$$D(R_X||P_X) \geq \frac{1}{2}\left\|J_X\right\|_1^2 \geq \frac{\left\|K_X\right\|_2^2 \left\|J_X\right\|_1}{2 \max\limits_{x\in\mathcal{X}} \left|\dfrac{J_X(x)}{P_X(x)}\right|}$$

where the denominator of the rightmost expression is strictly positive as $R_X \neq P_X$, and the second inequality follows from the ℓ^1-norm, ℓ^∞-norm form of Hölder's inequality:

$$\left\|K_X\right\|_2^2 = \sum_{x\in\mathcal{X}} \left|J_X(x)\right| \left|\frac{J_X(x)}{P_X(x)}\right| \leq \left(\max_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right) \left\|J_X\right\|_1.$$

As before, recognizing that $\left\|K_X\right\|_2^2 = \chi^2\left(R_X, P_X\right)$ completes the proof. ■

The bounds in Lemmata 6 and 7 are tighter than their counterparts in [9]. The former is tightened using the spectral norm instead of the Cauchy-Schwarz inequality with the Frobenius norm of $\left[\sqrt{P_X}\right]^{-1}$, and the latter is tightened by lower bounding $\left\|J_X\right\|_1$ instead of $\left\|J_X\right\|_1^2$. It turns out that we can further tighten Lemma 7 using "a distribution dependent refinement of Pinsker's inequality" derived in [15] instead of using the standard Pinsker's inequality. The following lemma presents this tighter bound.

Lemma 8 (KL Divergence Refined Lower Bound)**.** *Given distinct pmfs P_X and R_X on $\mathcal{X}$, such that $\forall x \in \mathcal{X}$, $P_X(x) > 0$, and $R_X = P_X + J_X = P_X + \left[\sqrt{P_X}\right] K_X$, we have:*

$$D(R_X||P_X) \geq \frac{\phi\left(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right)}{4} \frac{\left\|J_X\right\|_1}{\max\limits_{x\in\mathcal{X}} \left|\dfrac{J_X(x)}{P_X(x)}\right|} \left\|K_X\right\|_2^2$$

where for any subset $A \subseteq \mathcal{X}$, $P_X(A) = \sum_{x \in A} P_X(x)$, *and the function* $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ *is defined as:*

$$\phi(p) \triangleq \begin{cases} \frac{1}{1-2p} \log\left(\frac{1-p}{p}\right) & , \ p \in \left[0, \frac{1}{2}\right) \\ 2 & , \ p = \frac{1}{2} \end{cases} .$$

Proof. We now start with the refinement of Pinsker's inequality in [15]:

$$D(R_X || P_X) \geq \frac{\phi\left(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right)}{4} \left\|J_X\right\|_1^2 .$$

Following the proof of Lemma 7, using the ℓ^1-norm, ℓ^∞-norm form of Hölder's inequality:

$$\left\|K_X\right\|_2^2 = \sum_{x \in \mathcal{X}} \left|J_X(x)\right| \left|\frac{J_X(x)}{P_X(x)}\right| \leq \left(\max_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right) \left\|J_X\right\|_1$$

we have the following result:

$$D(R_X || P_X) \geq \frac{\phi\left(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right)}{4} \frac{\left\|J_X\right\|_1}{\max_{x \in \mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|} \left\|K_X\right\|_2^2$$

where the denominator on the right hand side is strictly positive as $R_X \neq P_X$. This completes the proof. ■

Lemma 8 is indeed tighter than Lemma 7, because $0 \leq \max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} \leq \frac{1}{2}$ and:

$$\phi\left(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right) \geq 2 \tag{12}$$

with equality if and only if $\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} = \frac{1}{2}$ [15]. Finally, we derive an upper bound on KL divergence using χ^2-divergence [16], which trivially follows from Jensen's inequality.

Lemma 9 (KL Divergence Upper Bound)**.** *Given pmfs* P_X *and* R_X *on* $\mathcal{X}$, *such that* $\forall x \in \mathcal{X}$, $P_X(x) > 0$, *and* $R_X = P_X + \left[\sqrt{P_X}\right] K_X$, *we have:*

$$D(R_X || P_X) \leq \log\left(1 + \left\|K_X\right\|_2^2\right) \leq \left\|K_X\right\|_2^2 = \chi^2\left(R_X, P_X\right).$$

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

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

where $D(R_X || P_X) = \mathbb{E}_{R_X}\left[\log\left(\frac{R_X(X)}{P_X(X)}\right)\right]$, and:

$$\mathbb{E}_{R_X}\left[\frac{R_X(X)}{P_X(X)}\right] = \sum_{x \in \mathcal{X}} \frac{R_X(x)^2}{P_X(x)} = 1 + \left\|K_X\right\|_2^2$$

using $R_X = P_X + \left[\sqrt{P_X}\right] K_X$. Hence, we have:

$$D(R_X || P_X) \leq \log\left(1 + \left\|K_X\right\|_2^2\right) \leq \left\|K_X\right\|_2^2 = \chi^2\left(R_X, P_X\right)$$

using the fact that: $\forall x > -1, \ \log(1 + x) \leq x$. ■

B. Bounds on Global Contraction Coefficient

Using the lemmata from the previous subsection, we can upper bound the global contraction coefficient in terms of the local contraction coefficient. In particular, combining Lemmata 6 and 9 produces Theorem 10, combining the tighter Lemma 7 with Lemma 9 produces the tighter upper bound in Theorem 11, and combining the even tighter Lemma 8 with Lemma 9 produces our tightest upper bound in Theorem 13.

Theorem 10 (Upper Bound on Global Contraction Coefficient)**.**

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq \frac{2}{\min_{x\in\mathcal{X}} P_X(x)}\, \eta_{\mathsf{loc}}(P_X, P_{Y|X})\,.$$

Proof. For any pmf $R_X = P_X + \left[\sqrt{P_X}\right] K_X$ on $\mathcal{X}$ such that $K_X \neq \vec{0}$ is a spherical perturbation, we have $R_Y = P_Y + \left[\sqrt{P_Y}\right] BK_X$, where B is the DTM. Hence, using Lemmata 6 and 9, we get:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{2}{\min_{x\in\mathcal{X}} P_X(x)} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2}.$$

Taking the supremum over R_X on the left hand side, and then the supremum over K_X on the right hand side, we can conclude the theorem statement from Definitions 1 and 2. ■

We remark that this result is independently derived in [17], where the author studies SDPIs for general f-divergences in the discrete setting. The next theorem tightens this result.

Theorem 11 (Contraction Coefficient Bound)**.**

$$\eta_{\mathsf{loc}}(P_X, P_{Y|X}) \leq \eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq \frac{\eta_{\mathsf{loc}}(P_X, P_{Y|X})}{\min_{x\in\mathcal{X}} P_X(x)}.$$

Proof. The lower bound is simply a restatement of Theorem 5. To derive the upper bound, we use Lemmata 7 and 9 to get:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{2}{\|J_X\|_1} \left(\max_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|\right) \frac{\|BK_X\|_2^2}{\|K_X\|_2^2}.$$

As in the proof of Theorem 10, this produces:

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq 2\,\eta_{\mathsf{loc}}(P_X, P_{Y|X}) \sup_{\substack{J_X : J_X \neq \vec{0} \\ J_X^T \vec{1} = 0}} \frac{\max_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|}{\|J_X\|_1}$$

since the supremum of a non-negative product is less than or equal to the product of the suprema, where $\vec{1}$ is a vector with all entries equal to unity. We now observe that:

$$\frac{\max_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|}{\|J_X\|_1} \leq \frac{\max_{x\in\mathcal{X}} |J_X(x)|}{\min_{x\in\mathcal{X}} P_X(x)\, \|J_X\|_1} \leq \frac{1}{2\min_{x\in\mathcal{X}} P_X(x)}$$

where the second inequality follows from $2\max_{x\in\mathcal{X}} |J_X(x)| \leq \|J_X\|_1$, which holds because $J_X^T\vec{1} = 0$. The upper bound of $1/(2\min_{x\in\mathcal{X}} P_X(x))$ can be achieved by choosing $J_X(x_0) = \delta$ for $x_0 = \arg\min_{x\in\mathcal{X}} P_X(x)$ and some sufficiently small $\delta > 0$, $J_X(x_1) = -\delta$ for some $x_1 \neq x_0$, and $J_X(x) = 0$ for every $x \in \mathcal{X}\backslash\{x_0, x_1\}$. Hence, we have:

$$\sup_{\substack{J_X : J_X \neq \vec{0} \\ J_X^T \vec{1} = 0}} \frac{\max_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|}{\|J_X\|_1} = \frac{1}{2\min_{x\in\mathcal{X}} P_X(x)}$$

which completes the proof. ■

We intuitively expect a bound between contraction coefficients to depend on $|\mathcal{X}|$ or $|\mathcal{Y}|$. Indeed, the constant in the upper bound of Theorem 11 satisfies $1/\min_{x\in\mathcal{X}} P_X(x) \geq |\mathcal{X}|$, and can therefore be construed as "modeling" $|\mathcal{X}|$. The lower bound of Theorem 11 asserts the intuitive fact that using the trajectory $\widehat{R}_X^{(\epsilon)}$ for model selection is suboptimal in a data processing sense, while the upper bound limits how much worse we can perform by using $\widehat{R}_X^{(\epsilon)}$.

Simulations in the binary case illustrate that $\eta_{\mathsf{glo}}(P_X, P_{Y|X})/\eta_{\mathsf{loc}}(P_X, P_{Y|X})$ increases significantly if some $P_X(x)$ is close to 0. This effect is captured in the upper bound, and is unsurprising given the skewed nature of stochastic manifolds (or probability simplices) at their edges with respect to KL divergence as the distance measure. However, the upper bound can be rendered arbitrarily loose since the constant $1/\min_{x\in\mathcal{X}} P_X(x)$ does not tensorize, while both $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ and $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$ do (Proposition 4). For example, if $X \sim$ Bernoulli(0.5), then the constant $1/\min_{x\in\{0,1\}} P_X(x) = 2$. If we instead consider $X_1^n = (X_1, \ldots, X_n)$ i.i.d. Bernoulli(0.5), then the constant in the upper bound is: $1/\min_{x_1^n\in\{0,1\}^n} P_{X_1^n}(x_1^n) = 2^n$. However, when we have i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, $\eta_{\mathsf{glo}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\mathsf{glo}}(P_{X_1}, P_{Y_1|X_1})$ and $\eta_{\mathsf{loc}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\mathsf{loc}}(P_{X_1}, P_{Y_1|X_1})$ by the tensorization property in Proposition 4. The next corollary presents a partial fix for this i.i.d. slackening attack.

Corollary 12 (Tensorized Contraction Coefficient Bound)**.** *If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$, then:*

$$\eta_{\mathsf{loc}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) \leq \eta_{\mathsf{glo}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) \leq \frac{\eta_{\mathsf{loc}}(P_{X_1^n}, P_{Y_1^n|X_1^n})}{\min\limits_{x\in\mathcal{X}} P_X(x)}.$$

Corollary 12 trivially follows from Theorem 11 and Proposition 4, and permits us to use the tighter factor of $1/\min_{x\in\mathcal{X}} P_X(x)$ in this product distribution context. While this corollary partially remedies the tensorization issue that ails Theorem 11, finding an upper bound which tensorizes gracefully remains a direction for future research. We next present our tightest bound between the global and local contraction coefficients. Its proof is almost identical to that of Theorem 11.

Theorem 13 (Tighter Contraction Coefficient Bound)**.**

$$\eta_{\mathsf{loc}}(P_X, P_{Y|X}) \leq \eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq \frac{2}{\phi\left(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right) \min\limits_{x\in\mathcal{X}} P_X(x)} \eta_{\mathsf{loc}}(P_X, P_{Y|X})$$

where for any subset $A \subseteq \mathcal{X}$, $P_X(A) = \sum\limits_{x\in A} P_X(x)$, and the function $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ is defined in Lemma 8.

Proof. Since the lower bound is a restatement of Theorem 5, we only prove the upper bound. Following the proof of Theorem 11, we have:

$$\frac{D(R_Y||P_Y)}{D(R_X||P_X)} \leq \frac{4}{\phi\left(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right)} \frac{\max\limits_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|}{\|J_X\|_1} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2}$$

using Lemmata 8 and 9. This then gives us:

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq \frac{4}{\phi\left(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right)} \eta_{\mathsf{loc}}(P_X, P_{Y|X}) \sup_{\substack{J_X : J_X \neq \vec{0} \\ J_X^T \vec{1} = 0}} \frac{\max\limits_{x\in\mathcal{X}} \left|\frac{J_X(x)}{P_X(x)}\right|}{\|J_X\|_1}.$$

We can compute the rightmost supremum as in the proof of Theorem 11 to get:

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \leq \frac{2}{\phi\left(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\right) \min\limits_{x\in\mathcal{X}} P_X(x)} \eta_{\mathsf{loc}}(P_X, P_{Y|X})$$

which completes the proof. ∎

Theorems 11 and 13 are the main results of this section. The bounds in these theorems are illustrated in Figure 1 for the binary symmetric channel. Although the upper bound in Theorem 13 is tighter than that in Theorem 11, testing it with our aforementioned i.i.d. slackening attack demonstrates that it unfortunately does not tensorize. We note that Section 3.4 in [17] also presents upper bounds on the global contraction coefficient that use the function, $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$, which stems from the refined Pinsker's inequality in [15]. However, these upper bounds do not admit the local contraction coefficient in a multiplicative way, which is essential in analyzing our model selection approach. We close this section with a tensorized version of Theorem 13 analogous to Corollary 12.

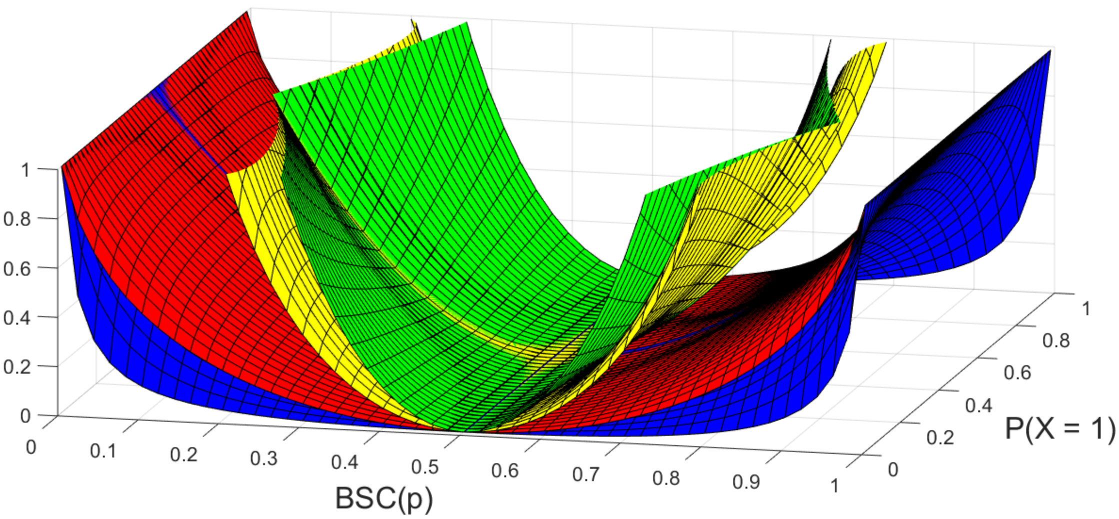

Fig. 1. Plot of the contraction coefficient bounds in Theorems 11 and 13 for a binary symmetric channel (BSC), $P_{Y|X}$, with crossover probability $p \in [0,1]$, and input random variable $X \sim \text{Bernoulli}(\mathbb{P}(X=1))$. The green mesh denotes the upper bound in Theorem 11, the yellow mesh denotes the tighter upper bound in Theorem 13, the red mesh denotes the global contraction coefficient $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$, and the blue mesh denotes the local contraction coefficient $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$.

Corollary 14 (Tighter Tensorized Contraction Coefficient Bound)**.** *If $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. with joint pmf $P_{X,Y}$, then:*

$$\eta_{\mathsf{loc}}\left(P_{X_1^n}, P_{Y_1^n|X_1^n}\right) \leq \eta_{\mathsf{glo}}\left(P_{X_1^n}, P_{Y_1^n|X_1^n}\right) \leq \frac{2}{\phi\left(\max\limits_{A \subseteq \mathcal{X}} \min\{P_X(A), 1-P_X(A)\}\right) \min\limits_{x \in \mathcal{X}} P_X(x)} \eta_{\mathsf{loc}}\left(P_{X_1^n}, P_{Y_1^n|X_1^n}\right)$$

where for any subset $A \subseteq \mathcal{X}$, $P_X(A) = \sum\limits_{x \in A} P_X(x)$, and the function $\phi : \left[0, \frac{1}{2}\right] \to \mathbb{R}$ is defined in Lemma 8.

As in the case of Corollary 12, Corollary 14 trivially follows from Theorem 13 and Proposition 4.

III. Equivalence in the Gaussian Setting

Our discussion regarding model selection using contraction coefficients can be naturally extended to include probability density functions (pdfs). For instance, the local approximations introduced in Subsection I-A were used to study AWGN channels in a network information theory context in [18]. We now consider the relationship between the local and global contraction coefficients in the Gaussian regime. To this end, we introduce the classical AWGN channel [7].

Definition 4 (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 $\sigma_X^2 > 0$.

It is well-known that the capacity of the AWGN channel is [7]:

$$C \triangleq \frac{1}{2} \log\left(1 + \frac{\sigma_X^2}{\sigma_W^2}\right) \tag{13}$$

where the capacity achieving input distribution (caid) is $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$. In the ensuing discussion, we fix $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, $\sigma_X^2 > 0$, as the input distribution to the AWGN channel. This defines a jointly Gaussian distribution $P_{X,Y}$, for which we will consider the global contraction coefficient defined earlier in Definition 1, as well as a variant of it that characterizes the SDPI for KL divergence when a power constraint is imposed on all input distributions.

Definition 5 (Global Contraction Coefficient with Power Constraint)**.** For a pair of jointly continuous random variables X and Y with joint pdf $P_{X,Y}$ and average power constraint $\mathbb{E}\left[X^2\right] \leq p$, $p > 0$, the global contraction coefficient with power constraint is defined as:

$$\eta^p_{\mathsf{glo}}\big(P_X, P_{Y|X}\big) \triangleq \sup_{\substack{R_X : R_X \neq P_X \\ 0 < D(R_X || P_X) < \infty \\ \mathbb{E}_{R_X}[X^2] \leq p}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where we take the supremum over all pdfs R_X which differ from P_X on a set with non-zero Lebesgue measure and satisfy the power constraint, and R_Y denotes the marginal pdf of $R_{X,Y} = P_{Y|X} R_X$.

We also consider the squared maximal correlation of X and Y, which equals the local contraction coefficient in general [6]. In this section, we prove that $\eta_{\mathsf{glo}}\big(P_X, P_{Y|X}\big)$, $\eta^p_{\mathsf{glo}}\big(P_X, P_{Y|X}\big)$ for any $p \geq \sigma_X^2$, and $\rho^2(X;Y)$ are all equal for the AWGN channel.

A. Maximal Correlation of AWGN Channel

The maximal correlation of the AWGN channel with caid can be computed from Rényi's seventh axiom [10] as we delineate next.

Proposition 15 (AWGN Maximal Correlation)**.** *For the AWGN channel with caid $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, the maximal correlation between X and Y is $\rho(X;Y) = \sigma_X / \sqrt{\sigma_X^2 + \sigma_W^2}$.*

Proof. According to Rényi's seventh axiom [10], the maximal correlation of a pair of jointly Gaussian random variables (X, Y) is the absolute value of the Pearson correlation coefficient. Furthermore, a pair of optimizing functions for maximal correlation (Definition 3) can be directly verified to be: $\forall x \in \mathbb{R}$, $f^*(x) = (x - \mathbb{E}[X])/\sqrt{\mathbb{VAR}(X)}$ and $\forall y \in \mathbb{R}$, $g^*(y) = \pm(y - \mathbb{E}[Y])/\sqrt{\mathbb{VAR}(Y)}$, where the sign of g^* is chosen so that $\mathbb{E}[f^*(X)g^*(Y)] \geq 0$. Intuitively, this holds because the minimum mean squared error estimator of X given Y is also the linear least squared error estimator, and optimizing functions of maximal correlation satisfy $\rho(X;Y)f^*(X) = \mathbb{E}\left[g^*(Y)|X\right]$ $a.s.$ and $\rho(X;Y)g^*(Y) = \mathbb{E}\left[f^*(X)|Y\right]$ $a.s.$ [10]. Hence, we have:

$$\rho(X;Y) = \frac{|\mathbb{COV}(X,Y)|}{\sqrt{\mathbb{VAR}(X)\mathbb{VAR}(Y)}} = \frac{\sigma_X}{\sqrt{\sigma_X^2 + \sigma_W^2}}$$

where the last equality holds because $\mathbb{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X(X+W)] = \mathbb{E}\left[X^2\right] + \mathbb{E}[X]\mathbb{E}[W] = \sigma_X^2$ as $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, $\mathbb{VAR}(X) = \sigma_X^2$, and $\mathbb{VAR}(Y) = \sigma_X^2 + \sigma_W^2$. ■

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 operator which takes spherical perturbations of the Gaussian input along right singular vector directions of Hermite polynomials to spherical perturbations of the Gaussian output [18].

B. Global Contraction Coefficients of AWGN Channel

To compute the global contraction coefficients for AWGN channels with caids, we first estimate them with the additional constraint that marginal distributions lie along exponential families. Then, we prove these estimates are precise using the entropy power inequality [7]. 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 deviations bounds. We next define one-parameter exponential families that are pdfs with respect to the Lebesgue measure, and list some of their properties [19].

Definition 6 (Regular Canonical Exponential Family)**.** The family of pdfs with natural parameter μ, $\{P_X(\cdot;\mu) : \mu \in \mathcal{M}\}$, is called a regular canonical exponential family when the support of the pdfs does not depend on μ, and each pdf in the family has the form:

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

where $t : \mathbb{R} \to \mathbb{R}$ is the sufficient statistic of the family, $P_X(x;0) = \exp(\beta(x))$ is a valid pdf known as the base distribution, and:

$$\forall \mu \in \mathcal{M}, \;\; \alpha(\mu) \triangleq \log\left(\int_{\mathbb{R}} \exp(\mu t(x) + \beta(x)) \, d\lambda(x)\right)$$

is the log-partition function with $\alpha(0) = 0$ without loss of generality, where λ is the Lebesgue measure and the integral is the Lebesgue integral. Furthermore, we define $\mathcal{M} \triangleq \{\mu \in \mathbb{R} : \alpha(\mu) < \infty\}$ as the natural parameter space, which ensures that $P_X(\cdot;\mu)$ is a valid pdf.

Proposition 16 (Properties of Canonical Exponential Family)**.** *For a regular canonical exponential family given by* $\{P_X(\cdot;\mu) : \mu \in \mathcal{M}\}$ *where* $\mathcal{M} \subseteq \mathbb{R}$ *is an open set, under regularity conditions such that the order of differentiation and integration can be exchanged using the dominated convergence theorem, we have:*

1) $\alpha(\mu) = \log(\mathbb{E}_{P_X(\cdot;0)}[e^{\mu t(X)}])$, $\alpha'(\mu) = \mathbb{E}_{P_X(\cdot;\mu)}[t(X)]$, *and* $\alpha''(\mu) = \mathbb{VAR}_{P_X(\cdot;\mu)}(t(X)) = J_X(\mu)$ *for every* $\mu \in \mathcal{M}$*, where the Fisher information,* $J_X : \mathcal{M} \to \mathbb{R}^+$*, is defined as:*

$$\forall \mu \in \mathcal{M},\ \ J_X(\mu) \triangleq \mathbb{E}_{P_X(\cdot;\mu)}\left[\left(\frac{\partial}{\partial \mu}\log\left(P_X(X;\mu)\right)\right)^2\right].$$

2) $\forall \mu \in \mathcal{M}$, $D\left(P_X(\cdot;\mu)||P_X(\cdot;0)\right) = \mu\mathbb{E}_{P_X(\cdot;\mu)}[t(X)] - \alpha(\mu)$.

The next proposition uses these properties to relate a global contraction coefficient like quantity, with the additional constraint that the marginal distributions are canonical exponential families, to the ratio of Fisher information terms of the marginal families.

Proposition 17 (Contraction 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 regular canonical exponential families,* $\forall x \in \mathbb{R},\ \ P_X(x;\mu) = \exp(\mu t(x) - \alpha(\mu) + \beta(x))$*, and* $\forall y \in \mathbb{R},\ \ P_Y(y;\mu) = \exp(\mu\tau(y) - A(\mu) + B(y))$*, with common natural parameter* $\mu \in \mathbb{R}$*, under regularity conditions, we have:*

$$\sup_{\substack{\mu \in \mathbb{R}:\ \mu \neq 0 \\ 0 < D(P_X(\cdot;\mu)||P_X(\cdot;0)) < \infty}} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{J_Y(\mu^*)}{J_X(\mu^*)}$$

where μ^* *is the value of* μ *at which the supremum is achieved.*

Proof. We assume the regularity conditions of Proposition 16, sufficient smoothness conditions, $\forall \mu \in \mathbb{R},\ \ J_X(\mu) > 0$, and $D(P_Y(\cdot;\mu)||P_Y(\cdot;0))$, $D(P_X(\cdot;\mu)||P_X(\cdot;0)) \to \infty$ as $\mu \to \pm\infty$. There are three possible cases. Firstly, if $\mu^* = 0$, then we have:

$$\lim_{\mu \to 0} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{J_Y(0)}{J_X(0)}$$

where we use $\lim_{\mu\to 0} D(P_X(\cdot;\mu)||P_X(\cdot;0))/\mu^2 = J_X(0)/2$ [7], which follows from Taylor approximation arguments. Secondly, if $\mu^* \in \mathbb{R}\backslash\{0\}$, then consider the function:

$$\begin{aligned} f(\mu) &= \log\left(\frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))}\right) \\ &= \log\left(\mu A'(\mu) - A(\mu)\right) - \log\left(\mu\alpha'(\mu) - \alpha(\mu)\right) \end{aligned}$$

where the second equality follows from Proposition 16. Since μ^* is a stationary point of $f(\mu)$ as $\log(\cdot)$ is monotonically increasing, we must have $f'(\mu^*) = 0$. Using Proposition 16, this translates to:

$$\begin{aligned} \frac{\mu^* A'(\mu^*) - A(\mu^*)}{\mu^*\alpha'(\mu^*) - \alpha(\mu^*)} &= \frac{A''(\mu^*)}{\alpha''(\mu^*)} \\ \Rightarrow \sup_{\mu \in \mathbb{R}:\ \mu \neq 0} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} &= \frac{J_Y(\mu^*)}{J_X(\mu^*)}. \end{aligned}$$

Finally, if $\mu^* = \pm\infty$, then l'Hôpital's rule and Proposition 16 give:

$$\lim_{\mu \to \pm\infty} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \lim_{\mu \to \pm\infty} \frac{\mu A''(\mu)}{\mu\alpha''(\mu)} = \frac{J_Y(\pm\infty)}{J_X(\pm\infty)}$$

where $J_Y(\pm\infty)/J_X(\pm\infty) \triangleq \lim_{\mu\to\pm\infty} J_Y(\mu)/J_X(\mu)$. ■

The elegant emergence of Fisher information in Proposition 17 is primarily due to the canonical exponential family constraints. Indeed, while KL divergence only locally approximates to the Fisher information metric in general, it is precisely the double integral of Fisher information for canonical exponential families. In the jointly Gaussian case, it turns out that the contraction coefficient under exponential family constraints actually equals $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$. Intuitively, this is because Gaussian distributions are completely characterized locally (by first and second moments). We next use Propositions 16 and 17 to derive the global contraction coefficient for the AWGN channel under marginal exponential family constraints.

Lemma 18 (AWGN Contraction 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 $\epsilon > 0$, and a family of pdfs $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$, we have:*

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

Proof. Observe that $P_X(x;\mu) = \exp(\mu t(x) - \alpha(\mu) + \beta(x))$ is a canonical exponential family with natural parameter $\mu \in \mathbb{R}$, $\exp(\beta(x)) = \mathcal{N}\left(0, \sigma_X^2\right)$, $\alpha(\mu) = \mu^2/(2\sigma_X^2)$, and $t(x) = x/\sigma_X^2$. 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, $P_Y(y;\mu) = \exp(\mu\tau(y) - A(\mu) + B(y))$ is also a canonical exponential family with common natural parameter $\mu \in \mathbb{R}$, $\exp(B(y)) = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right)$, $A(\mu) = \mu^2/(2(\sigma_X^2 + \sigma_W^2))$, and $\tau(y) = y/(\sigma_X^2 + \sigma_W^2)$. Since such Gaussian canonical exponential families with fixed variance and exponentially tilted means have constant Fisher information, using Proposition 16, we get:

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

Proposition 17 then produces:

$$\sup_{\mu\in\mathbb{R}:\, \mu\neq 0} \frac{D(P_Y(\cdot;\mu)||P_Y(\cdot;0))}{D(P_X(\cdot;\mu)||P_X(\cdot;0))} = \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$$

where we must ensure that the 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, is satisfied. To this end, notice from Proposition 16 that for every $\mu \neq 0$:

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

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:

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

As $\epsilon > 0$, $\exists \mu \neq 0$ such that $|\mu| \leq \sqrt{\epsilon}$, which satisfies the average power constraint. Hence, we have:

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

It is worth noting that Proposition 17 is not essential to this proof, and Proposition 16 suffices. ■

The proof of Lemma 18 elucidates why we use the additional slack of $\epsilon > 0$ in the average power constraint. Taking the supremum over $P_X(\cdot;\mu) = \mathcal{N}\left(\mu, \sigma_X^2\right)$ for $\mu \neq 0$ when $\epsilon = 0$ produces $-\infty$, because the average power constraint is never satisfied and the supremum is taken over an empty set. Hence, we add a slack of ϵ in Lemma 18 to ensure the supremum is well-defined. We now prove the main theorem of this section.

Theorem 19 (AWGN Channel Equivalence)**.** *For 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 $\mathbb{E}\left[X^2\right] \leq \sigma_X^2 + \epsilon$ for any $\epsilon \geq 0$, we have:*

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \eta_{\mathsf{glo}}^{\sigma_X^2+\epsilon}(P_X, P_{Y|X}) = \rho^2(X;Y) = \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}.$$

Proof. The last equality follows from Proposition 15. So, we prove the remaining equalities. Recall from Definition 1 that:

$$\eta_{\mathsf{glo}}(P_X, P_{Y|X}) = \sup_{\substack{R_X: R_X\neq P_X\\ 0<D(R_X||P_X)<\infty}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

and from Definition 5 that:

$$\eta_{\mathsf{glo}}^{\sigma_X^2+\epsilon}(P_X, P_{Y|X}) = \sup_{\substack{R_X: R_X\neq P_X\\ 0<D(R_X||P_X)<\infty\\ \mathbb{E}_{R_X}[X^2]\leq\sigma_X^2+\epsilon}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)}$$

where the pdfs R_X and P_X differ on a set with non-zero Lebesgue measure. Lemma 18 implies that $\eta_{\mathsf{glo}}(P_X, P_{Y|X}) \geq \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$ because we are constraining R_X to be $\mathcal{N}\left(\mu, \sigma_X^2\right)$ in it, and $\eta_{\mathsf{glo}}^{\sigma_X^2+\epsilon}(P_X, P_{Y|X}) \geq \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$ for any $\epsilon > 0$. Furthermore, for any pdf R_X and $P_X = \mathcal{N}\left(0, \sigma_X^2\right)$, we have:

$$D(R_X||P_X) = \frac{1}{2}\log\left(2\pi\sigma_X^2\right) + \frac{\mathbb{E}_{R_X}\left[X^2\right]}{2\sigma_X^2} - h\left(R_X\right) \tag{14}$$

where for any pdf $P : \mathbb{R} \rightarrow \mathbb{R}^+$, $h\left(P\right) \triangleq -\mathbb{E}_P\left[\log\left(P\right)\right]$ is the differential entropy of P. Letting $R_X = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 - \delta)$ and $R_Y = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 + \sigma_W^2 - \delta)$ for any $\delta > 0$, we get:

$$\lim_{\delta\rightarrow 0^+} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \lim_{\delta\rightarrow 0^+} \frac{\log\left(\frac{\sigma_X^2+\sigma_W^2}{\sigma_X^2+\sigma_W^2-\delta}\right)}{\log\left(\frac{\sigma_X^2}{\sigma_X^2-\delta}\right)} = \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}$$

where $P_X = \mathcal{N}\left(0, \sigma_X^2\right)$ and $P_Y = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right)$, and the second equality follows from l'Hôpital's rule. Hence, we have $\eta_{\mathsf{glo}}^{\sigma_X^2}\left(P_X, P_{Y|X}\right) \geq \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$. Although this trivially implies $\eta_{\mathsf{glo}}^{\sigma_X^2+\epsilon}(P_X, P_{Y|X}) \geq \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$ for every $\epsilon > 0$, we used Lemma 18 to derive it earlier to elucidate several connections to exponential families. We now observe that it suffices to prove that:

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

for every pdf $R_X \neq P_X$. Let R_X and $R_Y = R_X \star \mathcal{N}\left(0, \sigma_W^2\right)$, where $\star$ denotes the convolution operation, 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 $p > -\sigma_X^2$. Using (14), we have:

$$\begin{aligned} D(R_X||P_X) &= \frac{1}{2}\log\left(2\pi\sigma_X^2\right) + \frac{\sigma_X^2+p}{2\sigma_X^2} - h\left(R_X\right) \\ &= h\left(P_X\right) - h\left(R_X\right) + \frac{p}{2\sigma_X^2}, \\ D(R_Y||P_Y) &= h\left(P_Y\right) - h\left(R_Y\right) + \frac{p}{2\left(\sigma_X^2+\sigma_W^2\right)}. \end{aligned}$$

Hence, it suffices to prove that:

$$h(P_Y) - h(R_Y) \leq \frac{\sigma_X^2}{\sigma_X^2+\sigma_W^2}\left(h(P_X) - h(R_X)\right) \tag{15}$$

which we recast using entropy power terms as:

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

where for any pdf $P : \mathbb{R} \rightarrow \mathbb{R}^+$, $N(P) \triangleq e^{2h(P)}/(2\pi e)$ is the entropy power of 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 [7] to the AWGN channel, we have:

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

Hence, it is sufficient to prove that:

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

Let $a = \sigma_X^2 + \sigma_W^2$, $b = \sigma_X^2 - N(R_X)$, and $c = \sigma_X^2$. Then, we have $a > c > 0$ and $c > b$ (which follows from $h(R_X) > -\infty$ due to (14) and $D(R_X||P_X) < \infty$), and it is sufficient to prove that:

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

which is equivalent to proving:

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

This statement is a variant of Bernoulli's inequality proved in (r_7') and (r_7'') in [20]. This completes the proof. ■

Variants of Theorem 19 are well-known in the literature; a mutual information analog of this result is presented in [2]. However, our alternative proof and context offer some new perspective on this result, whose implications are quite profound. Recall that for model selection, $\eta_{\mathsf{glo}}^{\sigma_X^2}(P_X, P_{Y|X})$ and $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$ are achieved by globally optimal models in a data processing sense for an AWGN channel with or without a power constraint, respectively, and $\rho^2(X;Y)$ is achieved by locally optimal models. Theorem 19 portrays that models achieving $\rho^2(X;Y)$ are also globally optimal for AWGN channels. This conforms to our understanding of Gaussian distributions, where many local properties determine global ones. We next derive an interesting corollary of Theorem 19, which bounds 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 caid.

Corollary 20 (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:*

$$0 \leq C - I\left(R_X; P_{Y|X}\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right)$$

where C is the AWGN channel capacity defined in (13).

Proof. Let $P_X = \mathcal{N}\left(0, \sigma_X^2\right)$ be the caid, which is also the maximum differential entropy pdf given the second moment constraint $\mathbb{E}\left[X^2\right] \leq \sigma_X^2$, and let $P_Y = \mathcal{N}\left(0, \sigma_X^2 + \sigma_W^2\right)$ be the corresponding output pdf. From (15) in the proof of Theorem 19, we have:

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

for every pdf R_X satisfying $\mathbb{E}_{R_X}\left[X^2\right] \leq \sigma_X^2$. This trivially holds if $R_X = P_X$ *a.e.* with respect to the Lebesgue measure. Furthermore, $I\left(R_X; P_{Y|X}\right) = h(R_Y) - h\left(\mathcal{N}\left(0, \sigma_W^2\right)\right)$ for every pdf R_X for the AWGN channel, where $R_Y = R_X \star \mathcal{N}\left(0, \sigma_W^2\right)$. Hence, we have: $I\left(P_X; P_{Y|X}\right) - I\left(R_X; P_{Y|X}\right) = h(P_Y) - h(R_Y)$, which produces:

$$0 \leq C - I\left(R_X; P_{Y|X}\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right)$$

because $C = I\left(P_X; P_{Y|X}\right)$ and $h(P_X) = \frac{1}{2}\log\left(2\pi e \sigma_X^2\right)$. ■

Corollary 20 has a compelling analog in the discrete and finite setting. Consider a discrete memoryless channel with input random variable $X \in \mathcal{X}$, output random variable $Y \in \mathcal{Y}$, and conditional pmfs $\left\{P_{Y|X=x} : x \in \mathcal{X}\right\}$, where $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Let P_X be a capacity achieving input pmf, and P_Y be the unique capacity achieving output pmf. If $\forall x \in \mathcal{X}$, $P_X(x) > 0$, then for every pmf R_X on $\mathcal{X}$:

$$0 \leq I\left(P_X; P_{Y|X}\right) - I\left(R_X; P_{Y|X}\right) = D\left(R_Y || P_Y\right) \tag{16}$$

where R_Y is the marginal pmf of $R_{X,Y} = P_{Y|X}R_X$. This can be proved using the "equidistance" property of channel capacity [21], which states that P_X achieves capacity C_{DMC} if and only if $D\left(P_{Y|X=x} || P_Y\right) = C_{\mathsf{DMC}}$ for every $x \in \mathcal{X}$ such that $P_X(x) > 0$, and $D\left(P_{Y|X=x} || P_Y\right) \leq C_{\mathsf{DMC}}$ for every $x \in \mathcal{X}$ such that $P_X(x) = 0$. Using (16) and the SDPI: $D\left(R_Y || P_Y\right) \leq \eta_{\mathsf{glo}}(P_X, P_{Y|X})\, D\left(R_X || P_X\right)$, we have:

$$0 \leq C_{\mathsf{DMC}} - I\left(R_X; P_{Y|X}\right) \leq \eta_{\mathsf{glo}}(P_X, P_{Y|X})\, D\left(R_X || P_X\right) \tag{17}$$

which parallels Corollary 20, as can be seen using (14).

The inequalities in Corollary 20 and (17) are tight and equalities can be achieved. Moreover, we can recast Corollary 20 as:

$$C - I\left(R_X; P_{Y|X}\right) \leq \frac{\mathsf{snr}}{1 + \mathsf{snr}}\left(\frac{1}{2}\log\left(2\pi e \sigma_X^2\right) - h(R_X)\right) \tag{18}$$

where $\mathsf{snr} \triangleq \sigma_X^2 / \sigma_W^2$ is the signal-to-noise ratio. Hence, if $\mathsf{snr} \to 0$, then $I\left(R_X; P_{Y|X}\right) \to C$, which intuitively means that any input pdf 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. More generally, the capacity gap, $C - I(R_X; P_{Y|X})$, is sensitive to perturbations of the input distribution from the caid, and the input distribution achieves capacity if and only if it is Gaussian.

IV. Conclusion

In closing, we reiterate our main results. Our goal was to capture the performance loss in learning likelihood models using the extremal problem posed by $\eta_{\mathsf{loc}}(P_X, P_{Y|X})$, which admits a simple linear algebraic solution, instead of that posed by $\eta_{\mathsf{glo}}(P_X, P_{Y|X})$. In the discrete and finite regime, we accomplished this by appropriately bounding these contraction coefficients in Theorems 11 and 13. In the Gaussian regime, we proved in Theorem 19 that the local and global contraction coefficients are equal for AWGN channels with caids.

References

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *The Annals of Probability*, vol. 4, no. 6, pp. 925–939, December 1976.

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

[3] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," April 2013, arXiv:1304.6133 [cs.IT].

[4] ——, "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, October 2-4 2013, pp. 13–19.

[5] 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, October 1-5 2012, pp. 1057–1064.

[6] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," May 2015, arXiv:1405.3629 [cs.IT].

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New Jersey: John Wiley & Sons, Inc., 2006.

[8] S.-L. Huang and L. Zheng, "Linear information coupling problems," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Cambridge, MA, USA, July 1-6 2012, pp. 1029–1033.

[9] A. Makur, "A study of local approximations in information theory," Master's thesis, Massachusetts Institute of Technology, June 2015.

[10] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[11] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control and Computing*, Allerton House, UIUC, Illinois, USA, September 29-October 2 2015.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. New York: Cambridge University Press, 2011.

[13] 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.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[15] E. Ordentlich and M. J. Weinberger, "A distribution dependent refinement of pinsker's inequality," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1836–1840, May 2005.

[16] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating f-divergences," in *IEEE Information Theory Workshop (ITW)*, Jerusalem, Israel, April 26 - May 1 2015.

[17] M. Raginsky, "Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels," November 2014, arXiv:1411.3575 [cs.IT].

[18] E. Abbe and L. Zheng, "A coordinate system for gaussian networks," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 721–733, February 2012.

[19] R. W. Keener, *Theoretical Statistics: Topics for a Core Course*, ser. Springer Texts in Statistics. New York: Springer, 2010.

[20] Y.-C. Li and C.-C. Yeh, "Some equivalent forms of bernoulli's inequality: A survey," *Applied Mathematics*, vol. 4, no. 7, pp. 1070–1093, 2013.

[21] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, Inc., 1968.