Comparison of Local and Global Contraction Coefficients for KL Divergence

Anuran Makur and Lizhong Zheng

EECS Department, Massachusetts Institute of Technology

5 November 2015

Outline

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \to \pi \mathbf{1}^T$ (rank 1 matrix)

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \to \pi \mathbf{1}^T$ (rank 1 matrix)

Rate of convergence?

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \to \pi \mathbf{1}^T$ (rank 1 matrix)

Rate of convergence?

Perron-Frobenius:

$$1 = \lambda_1(W) > |\lambda_2(W)| \geq \cdots \geq |\lambda_n(W)|$$

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \to \pi \mathbf{1}^T$ (rank 1 matrix)

Rate of convergence?

Perron-Frobenius:

$$1 = \lambda_1(W) > |\lambda_2(W)| \geq \cdots \geq |\lambda_n(W)|$$

Rate of convergence determined by $|\lambda_2(W)| \longleftarrow$ coefficient of ergodicity

Measuring Ergodicity

Consider an ergodic Markov chain with $n \times n$ column stochastic transition matrix W.

- irreducible $\Rightarrow$ *unique stationary distribution* π: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \to \pi \mathbf{1}^T$ (rank 1 matrix)

Rate of convergence?
Perron-Frobenius:

$$1 = \lambda_1(W) > |\lambda_2(W)| \geq \cdots \geq |\lambda_n(W)|$$

Rate of convergence determined by $|\lambda_2(W)| \longleftarrow$ coefficient of ergodicity

Want: A guarantee on the relative improvement
i.e. for any distribution p, $W^{k+1}p$ is "closer" to π than $W^k p$.

Measuring Ergodicity

Let $d : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a divergence measure on the simplex $\mathcal{P}$.

$$\textbf{Want:} \quad \forall p \in \mathcal{P}, \quad d(Wp, \underbrace{W\pi}_{=\pi}) \leq \eta_d(\pi, W) d(p, \pi)$$

for some contraction coefficient $\eta_d(\pi, W) \in [0, 1]$.

Measuring Ergodicity

Let $d : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a divergence measure on the simplex $\mathcal{P}$.

$$\textbf{Want:} \quad \forall p \in \mathcal{P}, \quad d(Wp, \underbrace{W\pi}_{=\pi}) \leq \eta_d(\pi, W) d(p, \pi)$$

for some contraction coefficient $\eta_d(\pi, W) \in [0, 1]$. This would mean that:

$$\forall p \in \mathcal{P}, \quad d(W^k p, \pi) \leq \eta_d(\pi, W)^k d(p, \pi).$$

Measuring Ergodicity

Let $d : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a divergence measure on the simplex $\mathcal{P}$.

$$\textbf{Want:} \quad \forall p \in \mathcal{P}, \quad d(Wp, \underbrace{W\pi}_{=\pi}) \leq \eta_d(\pi, W) d(p, \pi)$$

for some contraction coefficient $\eta_d(\pi, W) \in [0, 1]$. This would mean that:

$$\forall p \in \mathcal{P}, \quad d(W^k p, \pi) \leq \eta_d(\pi, W)^k d(p, \pi).$$

$\eta_d(\pi, W) < 1 \;\; \Rightarrow \;\; W^k p \xrightarrow{d} \pi$ geometrically fast with rate $\eta_d(\pi, W)$.

Measuring Ergodicity

Let $d : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a divergence measure on the simplex $\mathcal{P}$.

$$\textbf{Want:} \quad \forall p \in \mathcal{P}, \quad d(Wp, \underbrace{W\pi}_{=\pi}) \leq \eta_d(\pi, W) d(p, \pi)$$

for some contraction coefficient $\eta_d(\pi, W) \in [0, 1]$. This would mean that:

$$\forall p \in \mathcal{P}, \quad d(W^k p, \pi) \leq \eta_d(\pi, W)^k d(p, \pi).$$

$\eta_d(\pi, W) < 1 \;\Rightarrow\; W^k p \xrightarrow{d} \pi$ geometrically fast with rate $\eta_d(\pi, W)$.

So, $\eta_d(\pi, W)$ is a coefficient of ergodicity, and we define it as:

$$\eta_d(\pi, W) \triangleq \sup_{p : p \neq \pi} \frac{d(Wp, W\pi)}{d(p, \pi)}.$$

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Does the ℓ^2-norm work?

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Does the ℓ^2-norm work?

$$\|W\pi - Wp\|_2 = \|W(\pi - p)\|_2 \leq \|W\|_2 \|\pi - p\|_2$$

where the spectral norm $\|W\|_2$ is the largest singular value of W.

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Does the ℓ^2-norm work?

$$\|W\pi - Wp\|_2 = \|W(\pi - p)\|_2 \leq \|W\|_2 \|\pi - p\|_2$$

where the spectral norm $\|W\|_2$ is the largest singular value of W.

$\|W\|_2 > 1$ *is possible...* ☹

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Does the ℓ^2-norm work?

$$\|W\pi - Wp\|_2 = \|W(\pi - p)\|_2 \leq \|W\|_2 \|\pi - p\|_2$$

where the spectral norm $\|W\|_2$ is the largest singular value of W.

$\|W\|_2 > 1$ *is possible...* ☹

Dobrushin-Doeblin Coefficient of Ergodicity:

The ℓ^1-norm (total variation distance) works! ☺

Measuring Ergodicity

Can we define notions of distance between distributions which make W a contraction?

Does the ℓ^2-norm work?

$$\|W\pi - Wp\|_2 = \|W(\pi - p)\|_2 \leq \|W\|_2 \|\pi - p\|_2$$

where the spectral norm $\|W\|_2$ is the largest singular value of W.

$\|W\|_2 > 1$ *is possible...* ☹

Dobrushin-Doeblin Coefficient of Ergodicity:

The ℓ^1-norm (total variation distance) works! ☺

$$\|W\pi - Wp\|_1 = \|W(\pi - p)\|_1 \leq \eta_{\mathsf{TV}}(\pi, W) \|\pi - p\|_1$$

where $\eta_{\mathsf{TV}}(\pi, W) \triangleq \sup_{p: p \neq \pi} \frac{\|W\pi - Wp\|_1}{\|\pi - p\|_1} \in [0, 1]$ is the Dobrushin-Doeblin contraction coefficient.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

- **Non-negativity:** $D_f(R_X||P_X) \geq 0$ with equality iff $R_X = P_X$.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ is convex and $f(1) = 0$.

- **Non-negativity:** $D_f(R_X||P_X) \geq 0$ with equality iff $R_X = P_X$.
- **Data Processing Inequality:** For a fixed channel $P_{Y|X}$:

$$\forall R_X, P_X, \quad D_f(R_Y||P_Y) \leq D_f(R_X||P_X)$$

where R_Y and P_Y are output pmfs corresponding to R_X and P_X.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Theorem [Amari and Cichocki, 2010]:
A *decomposable* divergence measure satisfies data processing if and only if it is an f-divergence.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Theorem [Amari and Cichocki, 2010]:
A *decomposable* divergence measure satisfies data processing if and only if it is an f-divergence.

Definition: A divergence d is *decomposable* if it can be written as:

$$d(R_X, P_X) = \sum_{x \in \mathcal{X}} g\left(R_X(x), P_X(x)\right)$$

for some function $g : [0,1]^2 \to \mathbb{R}$.

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Some Examples:

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Some Examples:

- **Total Variation Distance:** $f(t) = |t - 1|$ produces $D_f(R_X||P_X) = \|R_X - P_X\|_1$.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Some Examples:

- **Total Variation Distance:** $f(t) = |t - 1|$ produces $D_f(R_X||P_X) = \|R_X - P_X\|_1$.
- **KL Divergence:** $f(t) = t\log(t)$ produces $D_f(R_X||P_X) = D(R_X||P_X) = \sum_{x \in \mathcal{X}} R_X(x) \log\left(\frac{R_X(x)}{P_X(x)}\right)$.

Csiszár f-Divergence

Definition (Csiszár f-Divergence)

Given distributions R_X and P_X on $\mathcal{X}$, we define their f-divergence as:

$$D_f(R_X||P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is convex and $f(1) = 0$.

Some Examples:

- **Total Variation Distance:** $f(t) = |t - 1|$ produces $D_f(R_X||P_X) = \|R_X - P_X\|_1$.
- **KL Divergence:** $f(t) = t\log(t)$ produces $D_f(R_X||P_X) = D(R_X||P_X) = \sum_{x\in\mathcal{X}} R_X(x) \log\left(\frac{R_X(x)}{P_X(x)}\right)$.
- χ^2**-Divergence:** $f(t) = (t-1)^2$ produces $D_f(R_X||P_X) = \chi^2(R_X, P_X) = \sum_{x\in\mathcal{X}} \frac{(R_X(x)-P_X(x))^2}{P_X(x)}$.

Contraction Coefficients

Definition (Contraction Coefficient for f-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for f-divergence as:

$$\eta_f\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D_f(R_Y||P_Y)}{D_f(R_X||P_X)}$$

where R_Y is the output distribution when R_X passes through $P_{Y|X}$.

Contraction Coefficients

Definition (Contraction Coefficient for f-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for f-divergence as:

$$\eta_f\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D_f(R_Y||P_Y)}{D_f(R_X||P_X)}$$

where R_Y is the output distribution when R_X passes through $P_{Y|X}$.

Strong Data Processing Inequality

For fixed P_X and $P_{Y|X}$, we have:

$$\forall R_X, \quad D_f(R_Y||P_Y) \leq \eta_f\left(P_X, P_{Y|X}\right) D_f(R_X||P_X).$$

Contraction Coefficients

Definition (Contraction Coefficient for f-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for f-divergence as:

$$\eta_f\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D_f(R_Y||P_Y)}{D_f(R_X||P_X)}$$

where R_Y is the output distribution when R_X passes through $P_{Y|X}$.

Strong Data Processing Inequality

For fixed P_X and $P_{Y|X}$, we have:

$$\forall R_X, \quad D_f(R_Y||P_Y) \leq \eta_f\left(P_X, P_{Y|X}\right) D_f(R_X||P_X).$$

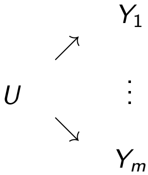
We will use the following instances of contraction coefficients:

1. $f(t) = t\log(t)$: $\eta_f\left(P_X, P_{Y|X}\right) = \eta_{\text{KL}}\left(P_X, P_{Y|X}\right)$
2. $f(t) = (t-1)^2$: $\eta_f\left(P_X, P_{Y|X}\right) = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$

Outline

Motivation: Inference Problem

Problem: Infer a hidden variable U about a "person X" given some data $Y_1, \ldots, Y_m \in \mathcal{Y}$ about the person that is conditionally independent given U.



Motivation: Inference Problem

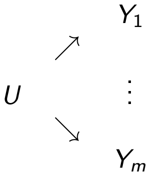
Problem: Infer a hidden variable U about a "person X" given some data $Y_1, \ldots, Y_m \in \mathcal{Y}$ about the person that is conditionally independent given U.

$$\begin{array}{ccc} & & Y_1 \\ & \nearrow & \\ U & & \vdots \\ & \searrow & \\ & & Y_m \end{array}$$

Assume U is binary with $\mathbb{P}(U = -1) = \mathbb{P}(U = 1) = \frac{1}{2}$.

Motivation: Inference Problem

Problem: Infer a hidden variable U about a "person X" given some data $Y_1, \dots, Y_m \in \mathcal{Y}$ about the person that is conditionally independent given U.



Assume U is binary with $\mathbb{P}(U = -1) = \mathbb{P}(U = 1) = \frac{1}{2}$.
Example: $U \in \{\text{conservative, liberal}\}$ and $\mathcal{Y}$ = movies watched on Netflix

Problem: Infer a hidden variable U about a "person X" given some data $Y_1, \ldots, Y_m \in \mathcal{Y}$ about the person that is conditionally independent given U.

$$\begin{array}{ccc} & & Y_1 \\ & \nearrow & \\ U & & \vdots \\ & \searrow & \\ & & Y_m \end{array}$$

Assume U is binary with $\mathbb{P}(U = -1) = \mathbb{P}(U = 1) = \frac{1}{2}$.
Example: $U \in$ {conservative, liberal} and $\mathcal{Y}$ = movies watched on Netflix
Log-likelihood Ratio Test: Construct sufficient statistic Z

$$U \longrightarrow (Y_1, \ldots, Y_m) \longrightarrow Z \triangleq \sum_{i=1}^{m} \log\left(\frac{P_{Y|U}(Y_i|1)}{P_{Y|U}(Y_i|-1)}\right)$$

Maximum Likelikood Estimate: $\widehat{U} = sign(Z)$

Motivation: Unsupervised Model Selection

How do we learn $P_{Y|U}$?

Motivation: Unsupervised Model Selection

How do we learn $P_{Y|U}$?
Given i.i.d. training data $(X_1, Y_1), \dots, (X_n, Y_n)$:

$$\begin{array}{ccccc} U_1 & \longrightarrow & X_1 & \longrightarrow & Y_1 \\ U_2 & \longrightarrow & X_2 & \longrightarrow & Y_2 \\ \vdots & & \vdots & & \vdots \\ U_n & \longrightarrow & X_n & \longrightarrow & Y_n \end{array}$$

where each $X_i \in \mathcal{X} = \{1, 2, \dots, |\mathcal{X}|\}$ and $\mathcal{X}$ indexes different people.

Motivation: Unsupervised Model Selection

How do we learn $P_{Y|U}$?
Given i.i.d. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

$$\begin{array}{ccccc} U_1 & \longrightarrow & X_1 & \longrightarrow & Y_1 \\ U_2 & \longrightarrow & X_2 & \longrightarrow & Y_2 \\ \vdots & & \vdots & & \vdots \\ U_n & \longrightarrow & X_n & \longrightarrow & Y_n \end{array}$$

where each $X_i \in \mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ and $\mathcal{X}$ indexes different people.
Training data gives us empirical distribution $\widehat{P}^n_{X,Y}$:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad \widehat{P}^n_{X,Y}(x, y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(X_i = x, Y_i = y)$$

Motivation: Unsupervised Model Selection

How do we learn $P_{Y|U}$?
Given i.i.d. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

$$\begin{array}{ccccc} U_1 & \longrightarrow & X_1 & \longrightarrow & Y_1 \\ U_2 & \longrightarrow & X_2 & \longrightarrow & Y_2 \\ \vdots & & \vdots & & \vdots \\ U_n & \longrightarrow & X_n & \longrightarrow & Y_n \end{array}$$

where each $X_i \in \mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ and $\mathcal{X}$ indexes different people.
Training data gives us empirical distribution $\widehat{P}^n_{X,Y}$:

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad \widehat{P}^n_{X,Y}(x, y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(X_i = x, Y_i = y)$$

We assume that the true distribution $P_{X,Y} = \widehat{P}^n_{X,Y}$
(motivated by concentration of measure results).

Motivation: Unsupervised Model Selection

Model Selection Problem:
Given $U \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and the joint pmf $P_{X,Y}$ for the Markov chain:

$$\begin{array}{ccccc} P_U & P_{X|U} & P_X & P_{Y|X} & P_Y \\ U & \longrightarrow & X & \longrightarrow & Y \end{array}$$

Find $P_{X|U}$

Motivation: Unsupervised Model Selection

Model Selection Problem:
Given $U \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and the joint pmf $P_{X,Y}$ for the Markov chain:

$$\begin{array}{ccccc} P_U & P_{X|U} & P_X & P_{Y|X} & P_Y \\ U & \longrightarrow & X & \longrightarrow & Y \end{array}$$

Find $P_{X|U}$ that maximizes the proportion of information that passes through the Markov chain:

$$\max \frac{I(U;Y)}{I(U;X)}.$$

Motivation: Unsupervised Model Selection

Model Selection Problem:
Given $U \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and the joint pmf $P_{X,Y}$ for the Markov chain:

$$\begin{array}{ccccc} P_U & P_{X|U} & P_X & P_{Y|X} & P_Y \\ U & \longrightarrow & X & \longrightarrow & Y \end{array}$$

Find $P_{X|U}$ that maximizes the proportion of information that passes through the Markov chain:

$$\max \frac{I(U;Y)}{I(U;X)}.$$

Remark: $\frac{I(U;Y)}{I(U;X)} = 1 \Rightarrow I(U;Y) = I(U;X)$ which means Y is a sufficient statistic for U.

Outline

1. Introduction to Contraction Coefficients
2. Motivation from Inference
3. Contraction Coefficients for KL and χ^2-Divergences
   - Data Processing Inequalities
   - Contraction Coefficient for KL Divergence
   - Local Approximation of KL Divergence
   - Local Contraction Coefficient
4. Bounds between Contraction Coefficients

Data Processing Inequalities

Data Processing Inequality for KL Divergence:
Fix P_X and $P_{Y|X}$. Then, for any R_X:

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

where R_Y is the output when R_X passes through $P_{Y|X}$.

Strong Data Processing Inequality for KL Divergence:
Fix P_X and $P_{Y|X}$. Then, for any R_X:

$$D(R_Y||P_Y) \leq \eta_{\mathsf{KL}}(P_X, P_{Y|X}) D(R_X||P_X)$$

Data Processing Inequalities

Data Processing Inequality for KL Divergence:
Fix P_X and $P_{Y|X}$. Then, for any R_X:

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

where R_Y is the output when R_X passes through $P_{Y|X}$.

Strong Data Processing Inequality for KL Divergence:
Fix P_X and $P_{Y|X}$. Then, for any R_X:

$$D(R_Y||P_Y) \leq \eta_{\mathsf{KL}}(P_X, P_{Y|X}) D(R_X||P_X)$$

Data Processing Inequality for Mutual Information:
Given a Markov chain $U \to X \to Y$:

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

Strong Data Processing Inequality for Mutual Information:
For fixed P_X and $P_{Y|X}$:

$$I(U;Y) \leq \eta_{\mathsf{KL}}(P_X, P_{Y|X}) I(U;X)$$

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for KL divergence and mutual information as:

$$\eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \sup_{\substack{P_U, P_{X|U}: \\ U \to X \to Y}} \frac{I(U;Y)}{I(U;X)}$$

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2016].

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for KL divergence and mutual information as:

$$\eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D(R_Y || P_Y)}{D(R_X || P_X)} = \sup_{\substack{P_U, P_{X|U}: \\ U \rightarrow X \rightarrow Y}} \frac{I(U;Y)}{I(U;X)}$$

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2016].

- This provides an optimization criterion which finds both P_U and $P_{X|U}$ for our model selection problem.

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for KL divergence and mutual information as:

$$\eta_{\mathrm{KL}}\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D(R_Y || P_Y)}{D(R_X || P_X)} = \sup_{\substack{P_U, P_{X|U}: \\ U \to X \to Y}} \frac{I(U;Y)}{I(U;X)}$$

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2016].

- This provides an optimization criterion which finds both P_U and $P_{X|U}$ for our model selection problem.
- The problem is not concave. So, it is difficult to solve.

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we define the contraction coefficient for KL divergence and mutual information as:

$$\eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \triangleq \sup_{R_X : R_X \neq P_X} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \sup_{\substack{P_U, P_{X|U}: \\ U \to X \to Y}} \frac{I(U;Y)}{I(U;X)}$$

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2016].

- This provides an optimization criterion which finds both P_U and $P_{X|U}$ for our model selection problem.
- The problem is not concave. So, it is difficult to solve.
- **Observation:** $D(R_Y||P_Y) \leq D(R_X||P_X)$ is tight when $R_X = P_X$, but the sequence of pmfs R_X achieving the supremum do not tend to P_X.

Local Approximation of KL Divergence

Idea: Find sequence of pmfs $R_X \to P_X$ that maximizes $\frac{D(R_Y||P_Y)}{D(R_X||P_X)}$.

Local Approximation of KL Divergence

Idea: Find sequence of pmfs $R_X \to P_X$ that maximizes $\frac{D(R_Y||P_Y)}{D(R_X||P_X)}$.

Consider the trajectory:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$$

where we can think of K_X and $\sqrt{P_X}$ as vectors, and $K_X^T\sqrt{P_X} = 0$.

Local Approximation of KL Divergence

Idea: Find sequence of pmfs $R_X \to P_X$ that maximizes $\frac{D(R_Y||P_Y)}{D(R_X||P_X)}$.

Consider the trajectory:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$$

where we can think of K_X and $\sqrt{P_X}$ as vectors, and $K_X^T\sqrt{P_X} = 0$.

Taylor's theorem:

$$D(R_X^{(\epsilon)}||P_X) = \frac{1}{2}\ \epsilon^2 \left\|K_X\right\|_2^2 \ + o\left(\epsilon^2\right)$$

$$D(R_Y^{(\epsilon)}||P_Y) = \frac{1}{2}\epsilon^2 \left\|BK_X\right\|_2^2 \ + o\left(\epsilon^2\right)$$

where $R_Y^{(\epsilon)} = P_{Y|X} \cdot R_X^{(\epsilon)}$, and B captures the effect of the channel on K_X:

$$B \triangleq \operatorname{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \operatorname{diag}\left(\sqrt{P_X}\right).$$

Local Approximation of KL Divergence

Idea: Find sequence of pmfs $R_X \to P_X$ that maximizes $\frac{D(R_Y||P_Y)}{D(R_X||P_X)}$.

Consider the trajectory:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$$

where we can think of K_X and $\sqrt{P_X}$ as vectors, and $K_X^T\sqrt{P_X} = 0$.

Taylor's theorem:

$$D(R_X^{(\epsilon)}||P_X) = \frac{1}{2}\underbrace{\epsilon^2\left\|K_X\right\|_2^2}_{=\chi^2(R_X^{(\epsilon)},P_X)} + o\left(\epsilon^2\right)$$

$$D(R_Y^{(\epsilon)}||P_Y) = \frac{1}{2}\underbrace{\epsilon^2\left\|BK_X\right\|_2^2}_{=\chi^2(R_Y^{(\epsilon)},P_Y)} + o\left(\epsilon^2\right)$$

where $R_Y^{(\epsilon)} = P_{Y|X} \cdot R_X^{(\epsilon)}$, and B captures the effect of the channel on K_X:

$$B \triangleq \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right).$$

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

- The trajectory:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$$

achieves the supremum in the LHS as $\epsilon \to 0$.

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X || P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y || P_Y)}{D(R_X || P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

- The trajectory:

$$\forall x \in \mathcal{X}, \quad R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$$

  achieves the supremum in the LHS as $\epsilon \to 0$.
- This formulation admits an easy solution using the SVD.

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

Model Selection Solution:

$$\forall x \in \mathcal{X}, \quad P_{X|U}(x|1) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$$
$$\forall x \in \mathcal{X}, \quad P_{X|U}(x|-1) = P_X(x) - \epsilon\sqrt{P_X(x)}K_X^*(x)$$

for fixed small ϵ.

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

- $\eta_{\chi^2}\left(P_X, P_{Y|X}\right)$ is also equal to the squared **Hirschfeld-Gebelein-Rényi maximal correlation**.

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

- $\eta_{\chi^2}\left(P_X, P_{Y|X}\right)$ is also equal to the squared **Hirschfeld-Gebelein-Rényi maximal correlation**.
- Other singular vectors of B can be used to decompose information into "mutually orthogonal" parts [Makur et al., 2015].

Local Contraction Coefficient

Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

For random variables X and Y with joint pmf $P_{X,Y}$, we have:

$$\lim_{\epsilon \to 0} \sup_{\substack{R_X : R_X \neq P_X \\ D(R_X||P_X) = \frac{1}{2}\epsilon^2}} \frac{D(R_Y||P_Y)}{D(R_X||P_X)} = \max_{\substack{K_X : K_X \neq \vec{0} \\ K_X^T \sqrt{P_X} = 0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2}\left(P_X, P_{Y|X}\right)$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and the RHS is maximized by K_X^*, which is the right singular vector of B corresponding to its "largest" singular value.

Compare $\eta_{\chi^2}\left(P_X, P_{Y|X}\right)$ **and** $\eta_{\text{KL}}\left(P_X, P_{Y|X}\right)$

Outline

Contraction Coefficient Bound

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min_{x \in \mathcal{X}} P_X(x)}.$$

Contraction Coefficient Bound

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Remark: Our local model selection method cannot perform "too poorly."

Contraction Coefficient Bound

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Remark: Our local model selection method cannot perform "too poorly."

Lower Bound:

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

Contraction Coefficient Bound

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Remark: Our local model selection method cannot perform “too poorly.”

Lower Bound:

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

Result is known in the literature, and inequality can be strict, as demonstrated in [Anantharam et al., 2013].

Upper Bound on Contraction Coefficient of KL Divergence

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Upper Bound Proof Sketch:

Upper Bound on Contraction Coefficient of KL Divergence

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathrm{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Upper Bound Proof Sketch:

Suppose we have:

- $D(R_Y || P_Y) \leq \alpha \left\| B K_X \right\|_2^2$, for some α
- $D(R_X || P_X) \geq \beta \left\| K_X \right\|_2^2$, for some β

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Upper Bound on Contraction Coefficient of KL Divergence

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Upper Bound Proof Sketch:

Suppose we have:

- $D(R_Y || P_Y) \leq \alpha \, \|BK_X\|_2^2$, for some α
- $D(R_X || P_X) \geq \beta \, \|K_X\|_2^2$, for some β

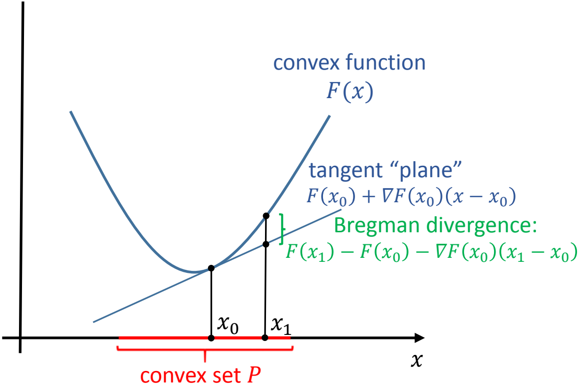
where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Then, we can prove an upper bound because:

$$\frac{D(R_Y || P_Y)}{D(R_X || P_X)} \leq \frac{\alpha}{\beta} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2}.$$

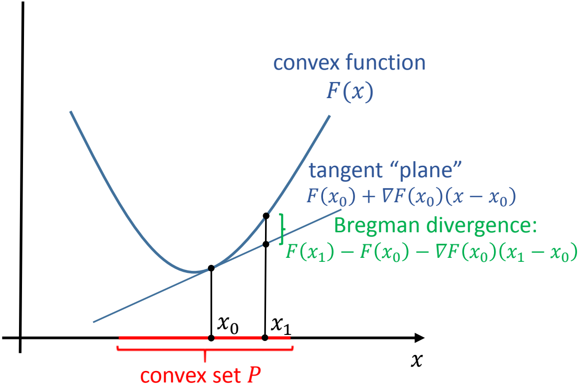
Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:



Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:



Bregman Divergence: Given $F : P \to \mathbb{R}$ convex:

$$\forall x_1, x_0 \in P, \quad B_F(x_1, x_0) \triangleq F(x_1) - F(x_0) - \nabla F(x_0)^T (x_1 - x_0)$$

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

Let $H_n : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ be the negative Shannon entropy function:

$$\forall Q \in \mathcal{P}_{\mathcal{X}}, \quad H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log\left(Q(x)\right).$$

KL divergence is a Bregman divergence [Banerjee et al., 2005]:

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

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

Let $H_n : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ be the negative Shannon entropy function:

$$\forall Q \in \mathcal{P}_{\mathcal{X}}, \quad H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log\left(Q(x)\right).$$

KL divergence is a Bregman divergence [Banerjee et al., 2005]:

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

$H_n : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}\left(Q\right)^{-1} \succeq I$, where I denotes the identity matrix.

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

Let $H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ be the negative Shannon entropy function:

$$\forall Q \in \mathcal{P}_\mathcal{X}, \quad H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log\left(Q(x)\right).$$

KL divergence is a Bregman divergence [Banerjee et al., 2005]:

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

$H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}\left(Q\right)^{-1} \succeq I$, where I denotes the identity matrix.

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

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

Let $H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ be the negative Shannon entropy function:

$$\forall Q \in \mathcal{P}_\mathcal{X}, \quad H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log\left(Q(x)\right).$$

KL divergence is a Bregman divergence [Banerjee et al., 2005]:

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

$H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}\left(Q\right)^{-1} \succeq I$, where I denotes the identity matrix.

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

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

Let $H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ be the negative Shannon entropy function:

$$\forall Q \in \mathcal{P}_\mathcal{X}, \quad H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log\left(Q(x)\right).$$

KL divergence is a Bregman divergence [Banerjee et al., 2005]:

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

$H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}\left(Q\right)^{-1} \succeq I$, where I denotes the identity matrix.

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

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

Using $\forall x \in \mathcal{X},\ R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$, we see that:

$$D(R_X||P_X) \geq \frac{1}{2} \left\|R_X - P_X\right\|_2^2 \geq \frac{\min_{x \in \mathcal{X}} P_X(x)}{2} \left\|K_X\right\|_2^2.$$

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Lower Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Lower Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

which can be improved to:

Lemma (KL Divergence Lower Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Upper Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\,K_X(x)$.

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Upper Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X},\ R_X(x) = P_X(x) + \sqrt{P_X(x)}\,K_X(x)$.

Proof:

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

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Upper Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Proof:

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

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

Bounding KL Divergence with χ^2-Divergence

Lemma (KL Divergence Upper Bound)

Given pmfs P_X and R_X, we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Proof:

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

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

Hence, we have: $D(R_X||P_X) \leq \log\left(1 + \|K_X\|_2^2\right) \leq \|K_X\|_2^2$,

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

Contraction Coefficient Bound

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

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

where R_Y is the output when R_X passes through $P_{Y|X}$, and $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$.

Contraction Coefficient Bound

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

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

where R_Y is the output when R_X passes through $P_{Y|X}$, and
$B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$.

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

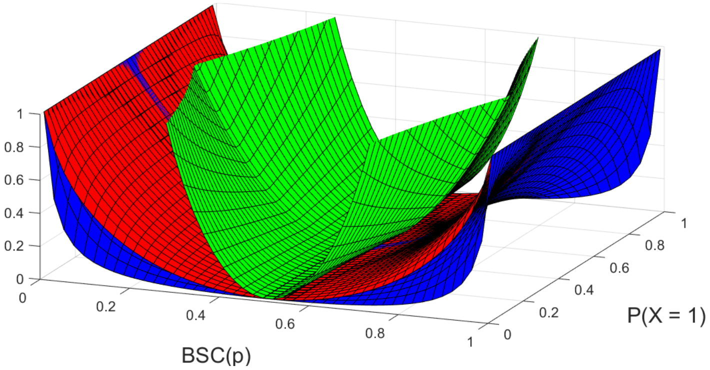
For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\text{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min_{x \in \mathcal{X}} P_X(x)}.$$

Example of Contraction Coefficient Bound

Binary Symmetric Channel Bounds:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\text{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min_{x\in\mathcal{X}} P_X(x)}$$



Conclusion

Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

$$\eta_{\chi^2}\left(P_X, P_{Y|X}\right) \leq \eta_{\mathsf{KL}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\chi^2}\left(P_X, P_{Y|X}\right)}{\min_{x \in \mathcal{X}} P_X(x)}.$$

Summary:

- Contraction coefficient for KL divergence can perform model selection, but no simple algorithm to solve it.
- Contraction coefficient for χ^2-divergence performs (suboptimal) model selection using the SVD.
- Bounds exist between these contraction coefficients.


That's all Folks!

Amari, S. and Cichocki, A. (2010).
Information geometry of divergence functions.
Bulletin of the Polish Academy of Sciences, Technical Sciences, 58(1):183–195.

Anantharam, V., Gohari, A., Kamath, S., and Nair, C. (2013).
On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover.
arXiv:1304.6133 [cs.IT].

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005).
Clustering with Bregman divergences.
Journal of Machine Learning Research, 6:1705–1749.

Makur, A., Kozynski, F., Huang, S.-L., and Zheng, L. (2015).
An efficient algorithm for information decomposition and extraction.
In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 972–979, Allerton House, UIUC, Illinois, USA.

Makur, A. and Zheng, L. (2015).

Bounds between contraction coefficients.
In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 1422–1429, Allerton House, UIUC, Illinois, USA.

Polyanskiy, Y. and Wu, Y. (2016).
Dissipation of information in channels with input constraints.
IEEE Transactions on Information Theory, 62(1):35–55.