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

1. Introduction to Contraction Coefficients
   - Measuring Ergodicity
   - Contraction Coefficients of Strong Data Processing Inequalities

2. Motivation from Inference

3. Contraction Coefficients for KL and $\chi^2$-Divergences

4. Bounds between Contraction Coefficients
Measuring Ergodicity

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

- irreducible $\Rightarrow$ unique stationary distribution $\pi$: $W\pi = \pi$
Measuring Ergodicity

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

- irreducible $\Rightarrow$ unique stationary distribution $\pi$: $W\pi = \pi$
- aperiodic $\Rightarrow W^k \rightarrow \pi 1^T$ (rank 1 matrix)
Measuring Ergodicity

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

- irreducible $\Rightarrow$ unique stationary distribution $\pi$: $W\pi = \pi$
- aperiodic $\Rightarrow$ $W^k \rightarrow \pi 1^T$ (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 $\pi$: $W \pi = \pi$
- aperiodic $\Rightarrow W^k \to \pi 1^T$ (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 \( \pi: W\pi = \pi \)
- aperiodic \( \Rightarrow \) \( W^k \rightarrow \pi 1^T \) (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 \( |\lambda_2(W)| \) \( \leftarrow \) coefficient of ergodicity
Measuring Ergodicity

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

- irreducible $\Rightarrow$ unique stationary distribution $\pi$: $W\pi = \pi$
- aperiodic $\Rightarrow W^k \to \pi 1^T$ (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 $|\lambda_2(W)| \leftarrow$ coefficient of ergodicity

Want: A guarantee on the relative improvement

i.e. for any distribution $p$, $W^{k+1}p$ is “closer” to $\pi$ than $W^k p$. 
Let $d : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a divergence measure on the simplex $\mathcal{P}$.

**Want:** \[ \forall p \in \mathcal{P}, \quad d(Wp, W\pi) \leq \eta_d(\pi, W)d(p, \pi) \]

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

A. Makur & L. Zheng (MIT) 
Local and Global Contraction Coefficients 
5 November 2015 4 / 32
Let \( d : \mathcal{P} \times \mathcal{P} \to [0, \infty] \) be a divergence measure on the simplex \( \mathcal{P} \).

\[
\textbf{Want:} \quad \forall p \in \mathcal{P}, \quad d(Wp, W\pi) \leq \eta_d(\pi, W)d(p, \pi)
\]

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 \( \mathcal{P} \).

**Want:** \( \forall p \in \mathcal{P}, \quad d(Wp, W\pi) \leq \eta_d(\pi, W)d(p, \pi) \)

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 \( \eta_d(\pi, W) \).
Let \(d : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]\) be a divergence measure on the simplex \(\mathcal{P}\).

**Want:** \(\forall p \in \mathcal{P}, \quad d(Wp_\pi, Wp) \leq \eta_d(\pi, W)d(p, \pi)\)

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:

\[
\eta_d(\pi, W) \triangleq \sup_{p : p \neq \pi} \frac{d(Wp, W\pi)}{d(p, \pi)}.
\]
Can we define notions of distance between distributions which make $W$ a contraction?

\[ \|W \|_2 > 1 \] is possible...

Dobrushin-Doeblin Coefficient of Ergodicity:

The $\ell_1$-norm (total variation distance) works!

\[ \|W \pi - Wp\|_1 \leq \eta_{TV}(\pi, W) \|\pi - p\|_1 \]

where $\eta_{TV}(\pi, W) \triangleq \sup_{p: p \neq \pi} \|W \pi - Wp\|_1 / \|\pi - p\|_1 \in [0, 1]$ is the Dobrushin-Doeblin contraction coefficient.
Can we define notions of distance between distributions which make $W$ a contraction?

Does the $\ell^2$-norm work?
Can we define notions of distance between distributions which make \( W \) a contraction?

Does the \( \ell^2 \)-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 \( W \).
Can we define notions of distance between distributions which make $W$ a contraction?

Does the $\ell^2$-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... 😞
Can we define notions of distance between distributions which make $W$ a contraction?

Does the $\ell^2$-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 $\ell^1$-norm (total variation distance) works! 😊
Measuring Ergodicity

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

Does the \( \ell^2 \)-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 \( \ell^1 \)-norm (total variation distance) works!

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

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

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

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:

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

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

**Definition (Csiszár $f$-Divergence)**

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

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

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

**Some Examples:**

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

Definition (Csiszár $f$-Divergence)

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

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

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

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

Definition (Csizár $f$-Divergence)

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

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

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

Some Examples:

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

**Definition (Csiszár $f$-Divergence)**

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

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

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

**Some Examples:**

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

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

- **$\chi^2$-Divergence:** $f(t) = (t - 1)^2$ produces
  $$D_f(R_X \Vert P_X) = \chi^2(R_X, P_X) = \sum_{x \in \mathcal{X}} \left( \frac{R_X(x) - P_X(x)}{P_X(x)} \right)^2.$$
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:

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

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. $f(t) = t \log(t): \eta_f \left( P_X, P_{Y|X} \right) = \eta_{KL} \left( P_X, P_{Y|X} \right)$
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

1. Introduction to Contraction Coefficients
2. Motivation from Inference
   - Inference Problem
   - Unsupervised Model Selection
3. Contraction Coefficients for KL and $\chi^2$-Divergences
4. Bounds between Contraction Coefficients
**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 $U$.

Diagram:

- $Y_1$ \arrow{U} \cdots \arrow{U} \arrow{Y_m}$

Assume $U$ is binary with $P(U = -1) = P(U = 1) = 1/2$.

**Example:** $U \in \{\text{conservative, liberal}\}$ and $Y = \text{movies watched on Netflix}$.

**Log-likelihood Ratio Test:**

Construct sufficient statistic $Z_U \rightarrow (Y_1, \ldots, Y_m) \rightarrow Z \equiv m \sum_{i=1}^{\cdot} \log \left( \frac{P(Y_i | U = 1)}{P(Y_i | U = -1)} \right)$.

**Maximum Likelihood Estimate:**

$\hat{U} = \text{sign}(Z)$.
**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 $U$.

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

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

Example: $U \in \{\text{conservative}, \text{liberal}\}$ and $\mathcal{Y} = \text{movies watched on Netflix}$

**Log-likelihood Ratio Test:** Construct sufficient statistic $Z$

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

Maximum Likelihood Estimate: $\hat{U} = \text{sign}(Z)$
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 $(X_1, Y_1), \ldots, (X_n, Y_n)$:

\[
\begin{align*}
U_1 & \rightarrow X_1 & \rightarrow & Y_1 \\
U_2 & \rightarrow X_2 & \rightarrow & Y_2 \\
& \vdots & \vdots & \vdots \\
U_n & \rightarrow X_n & \rightarrow & Y_n
\end{align*}
\]

where each $X_i \in \mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ and $\mathcal{X}$ indexes different people.
Motivation: Unsupervised Model Selection

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

Given i.i.d. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

\[
\begin{align*}
U_1 & \rightarrow X_1 \rightarrow Y_1 \\
U_2 & \rightarrow X_2 \rightarrow Y_2 \\
\vdots & \quad \vdots \\
U_n & \rightarrow X_n \rightarrow Y_n
\end{align*}
\]

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

Training data gives us empirical distribution $\hat{P}_{X,Y}^n$:

\[
\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad \hat{P}_{X,Y}^n(x, y) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(X_i = x, Y_i = y)
\]
Motivation: Unsupervised Model Selection

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

Given i.i.d. training data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

$$
\begin{align*}
U_1 & \rightarrow X_1 \rightarrow Y_1 \\
U_2 & \rightarrow X_2 \rightarrow Y_2 \\
\vdots & \quad \vdots \quad \vdots \\
U_n & \rightarrow X_n \rightarrow Y_n
\end{align*}
$$

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

Training data gives us empirical distribution $\hat{P}_X^n, Y^n$:

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

We assume that the true distribution $P_X, Y = \hat{P}_X^n, Y^n$ (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:

$P_U$ $P_{X|U}$ $P_X$ $P_{Y|X}$ $P_Y$

$U$ $\rightarrow$ $X$ $\rightarrow$ $Y$

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

\[
\begin{array}{cccccc}
  P_U & P_{X|U} & P_X & P_{Y|X} & P_Y \\
  U & \rightarrow & X & \rightarrow & Y
\end{array}
\]

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

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

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 $U$. 

A. Makur & L. Zheng (MIT)  Local and Global Contraction Coefficients  5 November 2015 13 / 32
1 Introduction to Contraction Coefficients

2 Motivation from Inference

3 Contraction Coefficients for KL and $\chi^2$-Divergences
   - Data Processing Inequalities
   - Contraction Coefficient for KL Divergence
   - Local Approximation of KL Divergence
   - Local Contraction Coefficient

4 Bounds between Contraction Coefficients
Data Processing 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 $R_X$:

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

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

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

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

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

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

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

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

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

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.
**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:

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

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.
- The problem is **not concave**. So, it is difficult to solve.
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:

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

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 $R_X \rightarrow P_X$ that maximizes $\frac{D(R_Y \| P_Y)}{D(R_X \| P_X)}$. 

Consider the trajectory:

$$\forall \ x \in X, R(\epsilon) X(x) = P_X(x) + \epsilon \sqrt{P_X(x) K_X(x)}$$

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

Taylor's theorem:

$$D(R(\epsilon) X || P_X) = \frac{1}{2} \epsilon^2 \| K_X \|^2 = \chi_2(R(\epsilon) X, P_X) + o(\epsilon^2)$$

$$D(R(\epsilon) Y || P_Y) = \frac{1}{2} \epsilon^2 \| B K_X \|^2 = \chi_2(R(\epsilon) Y, P_Y) + o(\epsilon^2)$$

where $R(\epsilon) Y = P_Y |_X \cdot R(\epsilon) X$, and $B$ captures the effect of the channel on $K_X$:
Local Approximation of KL Divergence

**Idea:** Find sequence of pmfs $R_X \rightarrow P_X$ 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 $R_X \rightarrow P_X$ 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:**

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

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

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 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:**

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

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

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

$$B \triangleq \text{diag} \left( \sqrt{P_Y} \right)^{-1} \cdot P_Y|_X \cdot \text{diag} \left( \sqrt{P_X} \right).$$
Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

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

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

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

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

$$\lim_{\epsilon \to 0} \sup_{R_X: R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} = \frac{1}{2} \epsilon^2$$

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 $\epsilon \to 0$. 

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

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

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

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

$$\lim_{\epsilon \to 0} \sup_{R_x: R_x \neq P_x} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} = \max_{K_X: K_X \neq \tilde{0}} \frac{\| BK_X \|_2^2}{\| K_X \|_2^2} = \eta \chi^2 (P_X, P_{Y|X})$$

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:

\begin{align*}
\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)
\end{align*}

for fixed small $\epsilon$. 
Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

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

$$\lim_{\epsilon \to 0} \sup_{R_X: R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} = \max_{K_X: K_X \neq \vec{0}} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \frac{1}{2} \epsilon^2$$

$$\frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} = \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{\chi^2} \left(P_X, P_{Y|X}\right)$$

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:

\[
\lim_{\epsilon \to 0} \sup_{R_X: R_X \neq P_X, D(R_X \| P_X) = \frac{1}{2} \epsilon^2} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} = \max_{K_X: K_X \neq \tilde{0}, K_X^T \sqrt{P_X} = 0} \frac{\|BK_X\|_2^2}{\|K_X\|_2^2} = \eta_{X}^2 \left( P_X, P_Y | X \right)
\]

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.

- $\eta_{X}^2 \left( P_X, P_Y | X \right)$ 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].
Theorem (Local Contraction Coefficient) [Makur and Zheng, 2015]

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

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

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 $\eta_{\chi^2} (P_X, P_{Y|X})$ and $\eta_{\text{KL}} (P_X, P_{Y|X})$
1. Introduction to Contraction Coefficients

2. Motivation from Inference

3. Contraction Coefficients for KL and $\chi^2$-Divergences

4. Bounds between Contraction Coefficients
   - Contraction Coefficient Bound
   - Upper Bound on Contraction Coefficient of KL Divergence
   - Bounding KL Divergence with $\chi^2$-Divergence
   - Binary Symmetric Channel Example
Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

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

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

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

Lower Bound:
$$
\lim_{\epsilon \to 0} \sup_{R_X: R_X \neq P_X} D(R_Y || P_Y) \leq \frac{1}{2} \epsilon^2 D(R_X || P_X) \leq \eta_{\chi^2} (P_X, P_{Y|X}).
$$

Result is known in the literature, and inequality can be strict, as demonstrated in [Anantharam et al., 2013].
**Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]**

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

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

**Remark:** Our local model selection method cannot perform “too poorly.”
Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

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

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

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

Lower Bound:

$$
\lim_{\epsilon \to 0} \sup_{R_X : R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} < \sup_{R_X : R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} \eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_{KL}(P_X, P_{Y|X})
$$
Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

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

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

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

Lower Bound:

$$\lim_{\epsilon \to 0} \sup_{R_X: R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} \leq \eta_{\chi^2}(P_X, P_{Y|X}) \leq \sup_{R_X: R_X \neq P_X} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)} .$$

Result is known in the literature, and inequality can be strict, as demonstrated in [Anantharam et al., 2013].
Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

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

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

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

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

**Upper Bound Proof Sketch:**

Suppose we have:

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

where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)} K_X(x)$.
Theorem (Contraction Coefficient Bound) [Makur and Zheng, 2015]

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

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

Upper Bound Proof Sketch:

Suppose we have:

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

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 $\chi^2$-Divergence

KL Divergence Lower Bound:

![Diagram showing KL Divergence Lower Bound]

- Convex function $F(x)$
- Tangent "plane" $F(x_0) + \nabla F(x_0)(x - x_0)$
- Bregman divergence: $F(x_1) - F(x_0) - \nabla F(x_0)(x_1 - x_0)$
- Convex set $P$
Bounding KL Divergence with $\chi^2$-Divergence

**KL Divergence Lower Bound:**

$\mathcal{P}$ convex set

$F(\mathbf{x})$ convex function

$\mathbf{x}$ tangent "plane"

$F(\mathbf{x}_0) + \nabla F(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0)$

**Bregman Divergence:**

Given $F : \mathcal{P} \rightarrow \mathbb{R}$ convex:

$$\forall \mathbf{x}_1, \mathbf{x}_0 \in \mathcal{P}, \quad B_F (\mathbf{x}_1, \mathbf{x}_0) \triangleq F(\mathbf{x}_1) - F(\mathbf{x}_0) - \nabla F(\mathbf{x}_0)^T (\mathbf{x}_1 - \mathbf{x}_0)$$
KL Divergence Lower Bound:

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

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

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

$$
D(R_X \| P_X) = H_n(R_X) - H_n(P_X) - \nabla H_n(P_X)^T (R_X - P_X).
$$
Bounding KL Divergence with $\chi^2$-Divergence

**KL Divergence Lower Bound:**
Let $H_n : \mathcal{P}_X \to \mathbb{R}$ be the negative Shannon entropy function:
\[
\forall Q \in \mathcal{P}_X, \quad H_n(Q) \triangleq \sum_{x \in X} Q(x) \log (Q(x)).
\]

KL divergence is a Bregman divergence [Banerjee et al., 2005]:
\[
D(R_X \| P_X) = H_n(R_X) - H_n(P_X) - \nabla H_n(P_X)^T (R_X - P_X).
\]

$H_n : \mathcal{P}_X \to \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}(Q)^{-1} \succeq I$, where $I$ denotes the identity matrix.
KL Divergence Lower Bound:
Let $H_n : \mathcal{P}_X \to \mathbb{R}$ be the negative Shannon entropy function:
\[
\forall Q \in \mathcal{P}_X, \quad H_n(Q) \triangleq \sum_{x \in X} Q(x) \log (Q(x)).
\]

KL divergence is a Bregman divergence [Banerjee et al., 2005]:
\[
D(R_X \| P_X) = H_n(R_X) - H_n(P_X) - \nabla H_n(P_X)^T (R_X - P_X).
\]

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

\[
H_n(R_X) \geq H_n(P_X) + \nabla H_n(P_X)^T (R_X - P_X) + \frac{1}{2} \| R_X - P_X \|^2_2
\]
Bounding KL Divergence with $\chi^2$-Divergence

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

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

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 $\nabla^2 H_n(Q) = \text{diag } (Q)^{-1} \succeq I$, where $I$ denotes the identity matrix.

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

$$D(R_X \| P_X) \geq \frac{1}{2} \| R_X - P_X \|_2^2$$
Bounding KL Divergence with $\chi^2$-Divergence

KL Divergence Lower Bound:
Let $H_n : \mathcal{P}_X \to \mathbb{R}$ be the negative Shannon entropy function:

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

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

$$D(\mathbb{R}_X \| \mathbb{P}_X) = H_n(\mathbb{R}_X) - H_n(\mathbb{P}_X) - \nabla H_n(\mathbb{P}_X)^T (\mathbb{R}_X - \mathbb{P}_X).$$

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

$$H_n(\mathbb{R}_X) \geq H_n(\mathbb{P}_X) + \nabla H_n(\mathbb{P}_X)^T (\mathbb{R}_X - \mathbb{P}_X) + \frac{1}{2} \| \mathbb{R}_X - \mathbb{P}_X \|^2_{\mathbb{P}_X}$$

$$D(\mathbb{R}_X \| \mathbb{P}_X) \geq \frac{1}{2} \| \mathbb{R}_X - \mathbb{P}_X \|^2_{\mathbb{P}_X}$$

Using $\forall x \in X$, $\mathbb{R}_X(x) = \mathbb{P}_X(x) + \sqrt{\mathbb{P}_X(x)} \mathbb{K}_X(x)$, we see that:

$$D(\mathbb{R}_X \| \mathbb{P}_X) \geq \frac{1}{2} \| \mathbb{R}_X - \mathbb{P}_X \|^2_{\mathbb{P}_X} \geq \frac{\min_{x \in X} \mathbb{P}_X(x)}{2} \| \mathbb{K}_X \|^2_{\mathbb{P}_X}.$$
Lemma (KL Divergence Lower Bound)

Given pmfs $P_X$ and $R_X$, we have:

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

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

A. Makur & L. Zheng (MIT)
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 $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}K_X(x)$. 
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:

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

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

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 $\chi^2$-Divergence

Lemma (KL Divergence Upper Bound)

Given pmfs $P_X$ and $R_X$, we have:

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

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

Proof:

$$D(R_X \parallel P_X) = \mathbb{E}_{R_X} \left[ \log \left( \frac{R_X(X)}{P_X(X)} \right) \right] \leq \log \left( \mathbb{E}_{R_X} \left[ \frac{R_X(X)}{P_X(X)} \right] \right) \leq \log \left( 1 + \|K_X\|^2 \right)$$

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

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

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

A. Makur & L. Zheng (MIT)
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

\[
B = \text{diag} \left( \sqrt{P_Y} \right)^{-1} \cdot P_{Y|X} \cdot \text{diag} \left( \sqrt{P_X} \right).
\]
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:

\[
\eta_{\chi^2} (P_X, P_{Y|X}) \leq \eta_{KL} (P_X, P_{Y|X}) \leq \frac{\eta_{\chi^2} (P_X, P_{Y|X})}{\min_{x \in \mathcal{X}} P_X(x)}.
\]
Example of Contraction Coefficient Bound

Binary Symmetric Channel Bounds:

\[ \eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_{KL}(P_X, P_{Y|X}) \leq \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{\min_{x \in \mathcal{X}} P_X(x)} \]
Conclusion

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

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

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

**Summary:**

- Contraction coefficient for KL divergence can perform model selection, but no simple algorithm to solve it.
- Contraction coefficient for $\chi^2$-divergence performs (suboptimal) model selection using the SVD.
- Bounds exist between these contraction coefficients.
That's all Folks!


Bounds between contraction coefficients.
