Bounds between Contraction Coefficients

Anuran Makur and Lizhong Zheng

EECS Department, Massachusetts Institute of Technology

Allerton Conference 2015

Outline

1. Motivation
   - Inference Problem
   - Unsupervised Model Selection
2. Contraction Coefficients of Strong Data Processing Inequalities
3. Bounds between Contraction Coefficients

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.

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

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 $Y_1, \ldots, Y_m \in \mathcal{Y}$ 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: $U \in \{\text{conservative, liberal}\}$ and $\mathcal{Y} =$ 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.

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

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 $(X_1, Y_1), \ldots, (X_n, Y_n)$:

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

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

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 $\widehat{P}^n_{X,Y}$:

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

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 $\widehat{P}^n_{X,Y}$:

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

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 the $P_{X|U}$

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

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 the $P_{X|U}$ that maximizes the proportion of information that passes through the Markov chain,
i.e. find $P_{X|U}$ that maximizes $\frac{I(U;Y)}{I(U;X)}$.

Outline

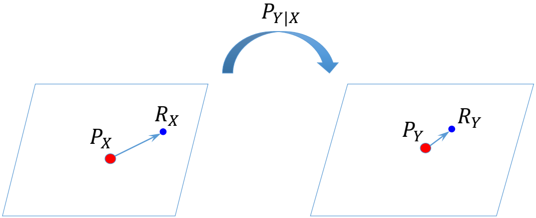
1. Motivation
2. Contraction Coefficients of Strong Data Processing Inequalities
   - Data Processing Inequalities
   - Contraction Coefficient for KL Divergence
   - Local Approximation of KL Divergence
   - Contraction Coefficient for χ^2-Divergence
3. Bounds between Contraction Coefficients

Data Processing Inequalities

Data Processing Inequality for KL Divergence:
Given a source P_X and a channel $P_{Y|X}$:

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

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



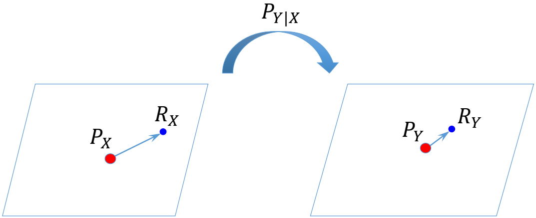
Data Processing Inequalities

Data Processing Inequality for KL Divergence:

Given a source P_X and a channel $P_{Y|X}$:

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

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:

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

Data Processing Inequalities

Data Processing Inequality for KL Divergence:
Given a source P_X and a channel $P_{Y|X}$:

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

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:

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

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

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

Data Processing Inequalities

Data Processing Inequality for KL Divergence:
Given a source P_X and a channel $P_{Y|X}$:

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

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:

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

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 $P_{Y|X}$:

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

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

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

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

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2015].

Contraction Coefficient for KL Divergence

Definition (Contraction Coefficient for KL Divergence)

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

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

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2015].

- 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 can define the contraction coefficient for KL divergence:

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

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2015].

- 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 can define the contraction coefficient for KL divergence:

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

where the second equality is proven in [Anantharam et al., 2013] and [Polyanskiy and Wu, 2015].

- 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: $\forall x \in \mathcal{X}$, $R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$, 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: $\forall x \in \mathcal{X},\ R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$, where we can think of K_X and $\sqrt{P_X}$ as vectors, and $K_X^T\sqrt{P_X} = 0$.
Then, using Taylor's theorem, we have:

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

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

where $R_Y^{(\epsilon)} = P_{Y|X} \cdot R_X^{(\epsilon)}$, and the matrix B is defined element-wise as $B(x,y) \triangleq P_{X,Y}(x,y)/\sqrt{P_X(x)P_Y(y)} = \sqrt{P_{X|Y}(x|y)P_{Y|X}(y|x)}$, or equivalently, $B \triangleq \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$, and it captures the effect of the channel on K_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: $\forall x \in \mathcal{X},\ R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X(x)$,
where we can think of K_X and $\sqrt{P_X}$ as vectors, and $K_X^T\sqrt{P_X} = 0$.
Then, using Taylor's theorem, we have:

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

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

where $R_Y^{(\epsilon)} = P_{Y|X} \cdot R_X^{(\epsilon)}$, and the matrix B is defined element-wise as
$B(x,y) \triangleq P_{X,Y}(x,y)/\sqrt{P_X(x)P_Y(y)} = \sqrt{P_{X|Y}(x|y)P_{Y|X}(y|x)}$, or
equivalently, $B \triangleq \operatorname{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \operatorname{diag}\left(\sqrt{P_X}\right)$, and it captures the effect of the channel on K_X.

Contraction Coefficient for χ^2-Divergence

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

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

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

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.

Contraction Coefficient for χ^2-Divergence

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

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

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

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 $\forall x \in \mathcal{X}$, $R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$ achieves the supremum in the LHS as $\epsilon \to 0$.

Contraction Coefficient for χ^2-Divergence

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

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

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

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 $\forall x \in \mathcal{X}$, $R_X^{(\epsilon)}(x) = P_X(x) + \epsilon \sqrt{P_X(x)} K_X^*(x)$ achieves the supremum in the LHS as $\epsilon \to 0$.
- This formulation admits an easy solution using the SVD.

Contraction Coefficient for χ^2-Divergence

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

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

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

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 $\forall x \in \mathcal{X}$, $R_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$ achieves the supremum in the LHS as $\epsilon \to 0$.
- This formulation admits an easy solution using the SVD.
- **Model Selection:** $\forall x \in \mathcal{X}$, $P_{X|U}(x|1) = P_X(x) + \epsilon\sqrt{P_X(x)}K_X^*(x)$ & $\forall x \in \mathcal{X}$, $P_{X|U}(x|-1) = P_X(x) - \epsilon\sqrt{P_X(x)}K_X^*(x)$, for fixed small ϵ.

Contraction Coefficient for χ^2-Divergence

Definition (Contraction Coefficient for χ^2-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we can define the contraction coefficient for χ^2-divergence:

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \triangleq \max_{\substack{K:K\neq 0 \\ K^T\sqrt{P_X}=0}} \frac{\|BK\|_2^2}{\|K\|_2^2}$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$. It is also known as the squared Hirschfeld-Gebelein-Rényi maximal correlation.

Recall that $\chi^2(Q,P) = \|K\|_2^2$, where $Q(x) = P(x) + \sqrt{P(x)}K(x)$ and $K^T\sqrt{P} = 0$.

Contraction Coefficient for χ^2-Divergence

Definition (Contraction Coefficient for χ^2-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we can define the contraction coefficient for χ^2-divergence:

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \triangleq \max_{\substack{K: K \neq 0 \\ K^T \sqrt{P_X} = 0}} \frac{\|BK\|_2^2}{\|K\|_2^2}$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$. It is also known as the squared Hirschfeld-Gebelein-Rényi maximal correlation.

Recall that $\chi^2(Q, P) = \|K\|_2^2$, where $Q(x) = P(x) + \sqrt{P(x)}K(x)$ and $K^T\sqrt{P} = 0$.

Learning models using maximal correlation was covered in Lizhong's talk [Makur et al., 2015].

Contraction Coefficient for χ^2-Divergence

Definition (Contraction Coefficient for χ^2-Divergence)

For a fixed source distribution P_X and channel $P_{Y|X}$, we can define the contraction coefficient for χ^2-divergence:

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \triangleq \max_{\substack{K:K\neq 0 \\ K^T\sqrt{P_X}=0}} \frac{\|BK\|_2^2}{\|K\|_2^2}$$

where $B = \text{diag}\left(\sqrt{P_Y}\right)^{-1} \cdot P_{Y|X} \cdot \text{diag}\left(\sqrt{P_X}\right)$. It is also known as the squared Hirschfeld-Gebelein-Rényi maximal correlation.

Recall that $\chi^2(Q,P) = \|K\|_2^2$, where $Q(x) = P(x) + \sqrt{P(x)}K(x)$ and $K^T\sqrt{P} = 0$.

Learning models using maximal correlation was covered in Lizhong's talk [Makur et al., 2015].

Compare $\eta_{\text{loc}}\left(P_X, P_{Y|X}\right)$ and $\eta_{\text{glo}}\left(P_X, P_{Y|X}\right)$

Outline

1. Motivation
2. Contraction Coefficients of Strong Data Processing Inequalities
3. Bounds between Contraction Coefficients
    - Contraction Coefficient Bound
    - Upper Bound on Contraction Coefficient of KL Divergence
    - Bounding KL Divergence with χ^2-Divergence
    - Binary Symmetric Channel Example

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:

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

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:

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

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:

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

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

Lower Bound:

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

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:

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \leq \eta_{\text{glo}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\text{loc}}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

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

Lower Bound:

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

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:

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \leq \eta_{\text{glo}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\text{loc}}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

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:

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

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:

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

Upper Bound Proof Sketch:

Suppose we have:

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

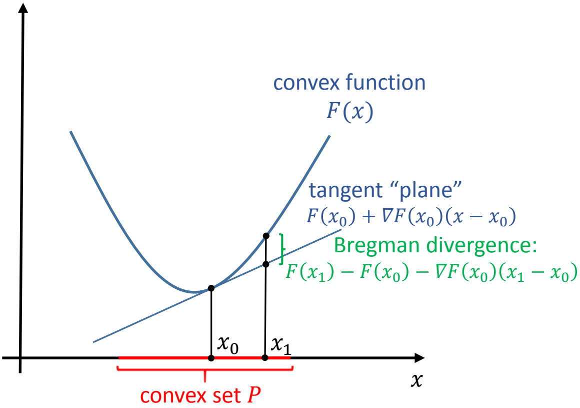
where $\forall x \in \mathcal{X}$, $R_X(x) = P_X(x) + \sqrt{P_X(x)}\, K_X(x)$.

Then, we can prove an upper bound because:

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

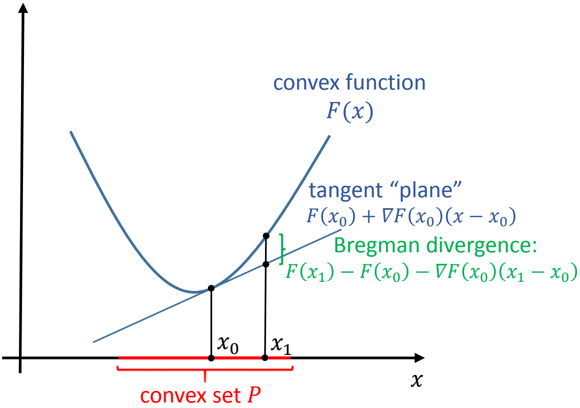
Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:



Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:



Let $\mathcal{P}_\mathcal{X}$ be the probability simplex of pmfs on $\mathcal{X}$, and $H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ be the negative Shannon entropy function: $H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log(Q(x))$.

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

The KL divergence is the Bregman divergence corresponding to H_n [Banerjee et al., 2005]:

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

where $H_n : \mathcal{P}_\mathcal{X} \to \mathbb{R}$ is the negative Shannon entropy function: $H_n(Q) \triangleq \sum_{x\in\mathcal{X}} Q(x) \log(Q(x))$.

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:
The KL divergence is the Bregman divergence corresponding to H_n [Banerjee et al., 2005]:

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

where $H_n : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ is the negative Shannon entropy function:
$H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log(Q(x))$.
$H_n : \mathcal{P}_{\mathcal{X}} \rightarrow \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}(Q)^{-1} \succeq I$, where I denotes the identity matrix.

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

The KL divergence is the Bregman divergence corresponding to H_n [Banerjee et al., 2005]:

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

where $H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ is the negative Shannon entropy function: $H_n(Q) \triangleq \sum_{x \in \mathcal{X}} Q(x) \log(Q(x))$.

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

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

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:
The KL divergence is the Bregman divergence corresponding to H_n [Banerjee et al., 2005]:

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

where $H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ is the negative Shannon entropy function: $H_n(Q) \triangleq \sum_{x\in\mathcal{X}} Q(x) \log (Q(x))$.
$H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ is strongly convex because $\nabla^2 H_n(Q) = \text{diag}(Q)^{-1} \succeq I$, where I denotes the identity matrix. Hence, we have:

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

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

Bounding KL Divergence with χ^2-Divergence

KL Divergence Lower Bound:

The KL divergence is the Bregman divergence corresponding to H_n [Banerjee et al., 2005]:

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

where $H_n : \mathcal{P}_\mathcal{X} \rightarrow \mathbb{R}$ is the negative Shannon entropy function: $H_n(Q) \triangleq \sum_{x\in\mathcal{X}} Q(x) \log(Q(x))$.

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

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

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

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

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

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

Lemma (KL Divergence Upper Bound)

Furthermore, for a fixed channel $P_{Y|X}$ we have:

$$D(R_Y||P_Y) \leq \|BK_X\|_2^2$$

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

Contraction Coefficient Bound

Using a tighter lower bound on KL divergence, we can show that:

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

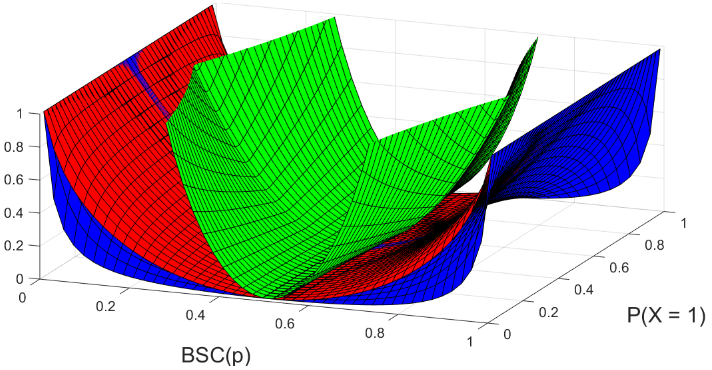
For a fixed source distribution P_X and channel $P_{Y|X}$, we have:

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

Example of Contraction Coefficient Bound

Binary Symmetric Channel Bounds:

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



Conclusion

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

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

$$\eta_{\text{loc}}\left(P_X, P_{Y|X}\right) \leq \eta_{\text{glo}}\left(P_X, P_{Y|X}\right) \leq \frac{\eta_{\text{loc}}\left(P_X, P_{Y|X}\right)}{\min\limits_{x \in \mathcal{X}} P_X(x)}.$$

Summary:

- Global contraction coefficient can perform model selection, but no simple algorithm to solve it.
- Local contraction coefficient performs (sub-optimal) model selection using the SVD.
- Bounds exist between these contraction coefficients.


That's all Folks!