Information Contraction and Decomposition

Anuran Makur

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Doctoral Thesis Defense
15 May 2019

Thesis Committee

Supervisors:  Lizhong Zheng and Yury Polyanskiy
Reader:      Elchanan Mossel
Outline

1. Introduction
   - f-Divergence
   - Data Processing Inequalities
   - Motivation for Strong Data Processing Inequalities

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Preliminaries

- finite alphabets $\mathcal{X}$ and $\mathcal{Y}$
Preliminaries

- finite alphabets $\mathcal{X}$ and $\mathcal{Y}$
- random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
Preliminaries

- finite alphabets $\mathcal{X}$ and $\mathcal{Y}$
- random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- probability distributions are row vectors
  e.g. $P_X$ is pmf on $\mathcal{X}$, and $P_Y$ is pmf on $\mathcal{Y}$
Preliminaries

- finite alphabets $\mathcal{X}$ and $\mathcal{Y}$
- random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
- probability distributions are row vectors
  e.g. $P_X$ is pmf on $\mathcal{X}$, and $P_Y$ is pmf on $\mathcal{Y}$
- channels (conditional distributions) are row stochastic matrices
  e.g. $W = P_{Y|X}$ such that $P_Y = P_X W$

\[ \begin{array}{c}
\text{probability simplex of pmfs of } X \\
P_X \\
\text{channel } W = P_{Y|X} \\
\text{probability simplex of pmfs of } Y \\
P_Y
\end{array} \]
**Definition (f-Divergence [Csi63, Mor63, AS66, ZZ73, Aka73])**

For any convex function $f : (0, \infty) \rightarrow \mathbb{R}$ such that $f(1) = 0$, we define the $f$-divergence between any two pmfs $R_X$ and $P_X$ on $\mathcal{X}$ 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 $f(0) = \lim_{t \to 0} f(t)$, $0 f \left( \frac{0}{0} \right) = 0$, and $0 f \left( \frac{r}{0} \right) = \lim_{p \to 0} pf \left( \frac{r}{p} \right)$ for all $r > 0$. 

Intuition:
- "Distance" between distributions
- Non-negativity: $D_f(R_X \parallel P_X) \geq 0$ with equality iff $R_X = P_X$ (where we assume that $f$ is strictly convex at 1)
**f-Divergence**

**Definition (f-Divergence [Csi63, Mor63, AS66, ZZ73, Aka73])**

For any convex function $f : (0, \infty) \rightarrow \mathbb{R}$ such that $f(1) = 0$, we define the $f$-divergence between any two pmfs $R_X$ and $P_X$ on $\mathcal{X}$ as:

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

- **Intuition:**
  “Distance” between distributions

**Probability simplex of pmfs of $X$**
**f-Divergence**

**Definition (f-Divergence [Csi63, Mor63, AS66, ZZ73, Aka73])**

For any convex function \( f : (0, \infty) \to \mathbb{R} \) such that \( f(1) = 0 \), we define the \( f \)-divergence between any two pmfs \( R_X \) and \( P_X \) on \( \mathcal{X} \) as:

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

- **Intuition:**
  “Distance” between distributions

- **Non-negativity:**
  \[
  D_f(R_X \| P_X) \geq 0
  \]

  with equality iff \( R_X = P_X \) (where we assume that \( f \) is strictly convex at 1)
Examples of $f$-Divergences

Kullback-Leibler (KL) Divergence:

$$f(t) = t \log(t)$$

$$\chi^2 (R_X || P_X) = \sum_{x \in X} R_X(x) \log \frac{R_X(x)}{P_X(x)}$$

(also known as relative entropy)

$\chi^2$-Divergence:

$$f(t) = (t - 1)^2$$

$$\chi^2 (R_X || P_X) = \sum_{x \in X} (R_X(x) - P_X(x))^2 P_X(x)$$

Total Variation (TV) Distance:

$$f(t) = \frac{1}{2} |t - 1|$$

$$\| R_X - P_X \|_{TV} = \frac{1}{2} \sum_{x \in X} |R_X(x) - P_X(x)|$$
Examples of $f$-Divergences

- **Kullback-Leibler (KL) Divergence:** $f(t) = t \log(t)$

  $$D(R_X \| P_X) = \sum_{x \in X} R_X(x) \log \left( \frac{R_X(x)}{P_X(x)} \right)$$

  (also known as *relative entropy*)
Examples of $f$-Divergences

- **Kullback-Leibler (KL) Divergence**: \( f(t) = t \log(t) \)

\[
D(R_X \| P_X) = \sum_{x \in \mathcal{X}} R_X(x) \log \left( \frac{R_X(x)}{P_X(x)} \right)
\]

(also known as *relative entropy*)

- **\( \chi^2 \)-Divergence**: \( f(t) = (t - 1)^2 \) or \( f(t) = t^2 - 1 \)

\[
\chi^2(R_X \| P_X) = \sum_{x \in \mathcal{X}} \frac{(R_X(x) - P_X(x))^2}{P_X(x)}
\]
Examples of $f$-Divergences

- **Kullback-Leibler (KL) Divergence**: $f(t) = t \log(t)$

  \[ D(R_X \| P_X) = \sum_{x \in \mathcal{X}} R_X(x) \log \left( \frac{R_X(x)}{P_X(x)} \right) \]

  (also known as *relative entropy*)

- **$\chi^2$-Divergence**: $f(t) = (t - 1)^2$ or $f(t) = t^2 - 1$

  \[ \chi^2(R_X \| P_X) = \sum_{x \in \mathcal{X}} \frac{(R_X(x) - P_X(x))^2}{P_X(x)} \]

- **Total Variation (TV) Distance**: $f(t) = \frac{1}{2} |t - 1|$

  \[ \| R_X - P_X \|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{X}} |R_X(x) - P_X(x)| \]
Prop (DPI for $f$-Divergences [Csi63, Mor63, AS66, ZZ73])

Given channel $W = P_{Y|X}$, for any two pmfs $R_X$ and $P_X$ on $\mathcal{X}$:

$$D_f(R_X W \| P_X W) \leq D_f(R_X \| P_X).$$
Data Processing Inequality (DPI)

Prop (DPI for $f$-Divergences [Csi63, Mor63, AS66, ZZ73])

Given channel $W = P_Y|X$, for any two pmfs $R_X$ and $P_X$ on $\mathcal{X}$:

$$D_f(R_X W || P_X W) \leq D_f(R_X || P_X).$$

Intuition: $R_X$ and $P_X$ are “less distinguishable” from noisy observation $Y$ compared to true data $X$. 
Consider ergodic Markov chain on state space $\mathcal{X}$:

- row stochastic transition kernel $W$

Rate of convergence?

DPI states that for any initial distribution $R$:

$$\|R W_n \| \leq \|R\| \|P\| \text{ for some coefficient } \eta \in (0, 1).$$
Consider \textit{ergodic Markov chain} on state space $\mathcal{X}$:

- row stochastic transition kernel $W$
- irreducible $\Rightarrow$ unique invariant distribution $P_{\mathcal{X}}$: $P_{\mathcal{X}}W = P_{\mathcal{X}}$
Consider **ergodic Markov chain** on state space $\mathcal{X}$:

- row stochastic transition kernel $W$
- irreducible $\Rightarrow$ **unique** invariant distribution $P_X$: $P_X W = P_X$
- irreducible & aperiodic $\Rightarrow \lim_{{n \to \infty}} R_X W^n = P_X$ for all initial pmfs $R_X$
Motivation for Stronger DPIs: Measuring Ergodicity

Consider **ergodic Markov chain** on state space $\mathcal{X}$:

- row stochastic transition kernel $W$
- irreducible $\Rightarrow$ *unique* invariant distribution $P_{\mathcal{X}}$
- irreducible & aperiodic $\Rightarrow \lim_{n \to \infty} R_{\mathcal{X}} W^n = P_{\mathcal{X}}$ for all initial pmfs $R_{\mathcal{X}}$

Rate of convergence?

Anuran Makur (MIT)
Motivation for Stronger DPs: Measuring Ergodicity

Consider ergodic Markov chain on state space $\mathcal{X}$:
- row stochastic transition kernel $W$
- irreducible $\Rightarrow$ unique invariant distribution $P_\mathcal{X}$
- irreducible & aperiodic $\Rightarrow \lim_{n \to \infty} R_\mathcal{X} W^n = P_\mathcal{X}$ for all initial pmfs $R_\mathcal{X}$

Rate of convergence?

DPI states that for any initial distribution $R_\mathcal{X}$:

$$D_f(R_\mathcal{X} W^n || P_\mathcal{X}) \leq D_f(R_\mathcal{X} || P_\mathcal{X}).$$
Motivation for Stronger DPIs: Measuring Ergodicity

Consider ergodic Markov chain on state space $\mathcal{X}$:
- row stochastic transition kernel $W$
- irreducible $\Rightarrow$ unique invariant distribution $P_X$
- irreducible & aperiodic $\Rightarrow \lim_{n \to \infty} R_X W^n = P_X$ for all initial pmfs $R_X$

Rate of convergence?

DPI states that for any initial distribution $R_X$:

$$D_f(R_X W^n \| P_X) \leq D_f(R_X \| P_X).$$

Want stronger version of DPI:

$$D_f(R_X W^n \| P_X) \leq \eta^n D_f(R_X \| P_X)$$

for some coefficient $\eta \in (0, 1)$. 

Outline

1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities
   - Properties of Contraction Coefficients
   - Linear Bounds between Contraction Coefficients
   - Illustration of Binary Case

3. Extension using Comparison of Channels

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Def (Contraction Coefficient I [Dob56, AG76, Sen81, CIRRSZ93])

For a fixed channel \( W = P_Y|X \), the contraction coefficient for an \( f \)-divergence is:

\[
\eta_f(P_Y|X) \triangleq \sup_{R_X, P_X : \begin{array}{c} 0 < D_f(R_X||P_X) < +\infty \\ D_f(R_X W||P_X W) \end{array}} \frac{D_f(R_X W||P_X W)}{D_f(R_X||P_X)}.
\]
Contraction Coefficients for $f$-Divergences

Def (Contraction Coefficient I [Dob56, AG76, Sen81, CIRRSZ93])

For a fixed channel $W = P_{Y|X}$, the contraction coefficient for an $f$-divergence is:

\[
\eta_f(P_{Y|X}) \triangleq \sup_{R_X, P_X: 0 < D_f(R_X \| P_X) < +\infty} \frac{D_f(R_X W \| P_X W)}{D_f(R_X \| P_X)}.
\]

Def (Contraction Coefficient II [Sar58, AG76, MZ15, PW16, Rag16])

For a fixed source distribution $P_X$ and channel $W = P_{Y|X}$, the contraction coefficient for an $f$-divergence is:

\[
\eta_f(P_X, P_{Y|X}) \triangleq \sup_{R_X: 0 < D_f(R_X \| P_X) < +\infty} \frac{D_f(R_X W \| P_X W)}{D_f(R_X \| P_X)}.
\]
**Contraction Coefficients for $f$-Divergences**

**Def (Contraction Coefficient I [Dob56, AG76, Sen81, CIRRSZ93])**

For a fixed channel $W = P_{Y|X}$, the contraction coefficient for an $f$-divergence is:

$$
\eta_f(P_{Y|X}) \triangleq \sup_{R_X, P_X: 0 < D_f(R_X||P_X) < +\infty} \frac{D_f(R_XW||P_XW)}{D_f(R_X||P_X)} = \sup_{P_X} \eta_f(P_X, P_{Y|X}).
$$

**Def (Contraction Coefficient II [Sar58, AG76, MZ15, PW16, Rag16])**

For a fixed source distribution $P_X$ and channel $W = P_{Y|X}$, the contraction coefficient for an $f$-divergence is:

$$
\eta_f(P_X, P_{Y|X}) \triangleq \sup_{R_X: 0 < D_f(R_X||P_X) < +\infty} \frac{D_f(R_XW||P_XW)}{D_f(R_X||P_X)}.
$$
For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:

$$D_f(R_X W || P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X || P_X).$$

For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:

$$D_f(R_X W || P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X || P_X).$$
Strong Data Processing Inequality (SDPI)

- For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:
  \[
  D_f(R_X W \| P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X \| P_X).
  \]

- For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:
  \[
  D_f(R_X W \| P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X \| P_X).
  \]

Special Cases:

- KL divergence: $\eta_{KL}(P_{Y|X}), \eta_{KL}(P_X, P_{Y|X})$
Strong Data Processing Inequality (SDPI)

For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:

$$D_f(R_X W \| P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X \| P_X).$$

For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:

$$D_f(R_X W \| P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X \| P_X).$$

Special Cases:

- KL divergence: $\eta_{KL}(P_{Y|X}), \eta_{KL}(P_X, P_{Y|X})$
- $\chi^2$-divergence: $\eta_{\chi^2}(P_X, P_{Y|X})$ (squared maximal correlation)
For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:

$$D_f(R_X W || P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X || P_X).$$

For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:

$$D_f(R_X W || P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X || P_X).$$

**Special Cases:**

- KL divergence: $\eta_{KL}(P_{Y|X}), \eta_{KL}(P_X, P_{Y|X})$
- $\chi^2$-divergence: $\eta_{\chi^2}(P_X, P_{Y|X})$ (squared maximal correlation)
- TV distance: $\eta_{TV}(P_{Y|X})$ (Dobrushin contraction coefficient)
Strong Data Processing Inequality (SDPI)

- For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:
  \[
  D_f(R_X W \| P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X \| P_X).
  \]

- For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:
  \[
  D_f(R_X W \| P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X \| P_X).
  \]

- Properties of contraction coefficients I well-studied [CIRRSZ93].

Special Cases:

- KL divergence: $\eta_{KL}(P_{Y|X}), \eta_{KL}(P_X, P_{Y|X})$
- $\chi^2$-divergence: $\eta_{\chi^2}(P_X, P_{Y|X})$ (squared maximal correlation)
- TV distance: $\eta_{TV}(P_{Y|X})$ (Dobrushin contraction coefficient)
Strong Data Processing Inequality (SDPI)

- For fixed channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X, P_X$:
  \[
  D_f(R_X W \parallel P_X W) \leq \eta_f(P_{Y|X}) D_f(R_X \parallel P_X).
  \]

- For fixed source pmf $P_X$ and channel $W = P_{Y|X}$, the SDPI states that for all pmfs $R_X$:
  \[
  D_f(R_X W \parallel P_X W) \leq \eta_f(P_X, P_{Y|X}) D_f(R_X \parallel P_X).
  \]

- Properties of contraction coefficients II?

Special Cases:

- KL divergence: $\eta_{KL}(P_{Y|X}), \eta_{KL}(P_X, P_{Y|X})$
- $\chi^2$-divergence: $\eta_{\chi^2}(P_X, P_{Y|X})$ (squared maximal correlation)
- TV distance: $\eta_{TV}(P_{Y|X})$ (Dobrushin contraction coefficient)
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** \( 0 \leq \eta_f(P_X, P_{Y|X}) \leq 1 \).
Theorem (Properties of Contraction Coefficients II)

- **Normalization**: $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.
- **Independence**: $\eta_f(P_X, P_{Y|X}) = 0$ if and only if $X$ and $Y$ are independent.
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.
- **Independence:** $\eta_f(P_X, P_{Y|X}) = 0$ if and only if $X$ and $Y$ are independent.
- **Decomposability:** If $f$ is strictly convex, twice differentiable at unity with $f''(1) > 0$, and $f(0) < \infty$, then $\eta_f(P_X, P_{Y|X}) = 1$ if and only if $P_{X,Y}$ is decomposable.
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.
- **Independence:** $\eta_f(P_X, P_{Y|X}) = 0$ if and only if $X$ and $Y$ are independent.
- **Decomposability:** If $f$ is strictly convex, twice differentiable at unity with $f''(1) > 0$, and $f(0) < \infty$, then $\eta_f(P_X, P_{Y|X}) = 1$ if and only if $P_{X,Y}$ is decomposable (i.e. there exist $h : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ such that $h(X) = g(Y)$ a.s. and $\operatorname{VAR}(h(X)) > 0$ [AG76]).
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.
- **Independence:** $\eta_f(P_X, P_{Y|X}) = 0$ if and only if $X$ and $Y$ are independent.
- **Decomposability:** If $f$ is strictly convex, twice differentiable at unity with $f''(1) > 0$, and $f(0) < \infty$, then $\eta_f(P_X, P_{Y|X}) = 1$ if and only if $P_{X,Y}$ is decomposable.
- **$\eta_{\chi^2}$ Lower Bound [MZ15, Rag16, PW17]:**
  
  If $f$ is twice differentiable at unity and $f''(1) > 0$:
  
  $$\eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_f(P_X, P_{Y|X}).$$
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** $0 \leq \eta_f(P_X, P_{Y|X}) \leq 1$.
- **Independence:** $\eta_f(P_X, P_{Y|X}) = 0$ if and only if $X$ and $Y$ are independent.
- **Decomposability:** If $f$ is strictly convex, twice differentiable at unity with $f''(1) > 0$, and $f(0) < \infty$, then $\eta_f(P_X, P_{Y|X}) = 1$ if and only if $P_{X,Y}$ is decomposable.
- **$\chi^2$ Lower Bound:** For any pmf $P_X$ and channel $W = P_{Y|X}$, if $f$ is twice differentiable at unity and $f''(1) > 0$:

\[
\eta_{\chi^2}(P_X, P_{Y|X}) = \lim_{\delta \to 0^+} \sup_{R_X: 0 < D_f(R_X || P_X) \leq \delta} \frac{D_f(R_X W || P_X W)}{D_f(R_X || P_X)}.
\]
Theorem (Properties of Contraction Coefficients II)

- **Normalization:** \( 0 \leq \eta_f(P_X, P_{Y|X}) \leq 1. \)

- **Independence:** \( \eta_f(P_X, P_{Y|X}) = 0 \) if and only if \( X \) and \( Y \) are independent.

- **Decomposability:** If \( f \) is strictly convex, twice differentiable at unity with \( f''(1) > 0 \), and \( f(0) < \infty \), then \( \eta_f(P_X, P_{Y|X}) = 1 \) if and only if \( P_{X,Y} \) is decomposable.

- \( \eta_{\chi^2} \) **Lower Bound** [MZ15, Rag16, PW17]:
  If \( f \) is twice differentiable at unity and \( f''(1) > 0 \):
  \[
  \eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_f(P_X, P_{Y|X}).
  \]

- Is there an upper bound on \( \eta_f \) in terms of \( \eta_{\chi^2} \)?
Fix any pmf $P_X$ with $p_* \triangleq \min_{x \in X} P_X(x) > 0$, and any channel $P_{Y|X}$. 

Theorem (Contraction Coefficient Bound)

If $f$ satisfies certain "regularity conditions," then:

$$\eta_f(P_X, P_{Y|X}) \leq f'(1) + f(0) f''(1) p_* \eta_{\chi_2}(P_X, P_{Y|X})$$

(KL Contraction Coefficient Bound)

$$\eta_{KL}(P_X, P_{Y|X}) \leq 2 \eta_{\chi_2}(P_X, P_{Y|X}) \phi\left(\max_{A \subseteq X} \min\{P_X(A), P_X(A^c)\}\right) p_*$$

where $\phi(p) = \frac{1}{1 - 2p \log \left(\frac{1 - p}{p}\right)}$. 

Proof Idea:

Use bounds between $f$-divergences and $\chi_2$-divergence based on [Su95, OW05, Gil10, Rag16].
Fix any pmf $P_X$ with $p_\star \triangleq \min_{x \in \mathcal{X}} P_X(x) > 0$, and any channel $P_{Y|X}$.

**Theorem (Contraction Coefficient Bound)**

If $f$ satisfies certain “regularity conditions,” then:

$$
\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) p_\star} \eta^2_{\chi^2}(P_X, P_{Y|X}).
$$

Proof Idea: Use bounds between $f$-divergences and $\chi^2$-divergence based on [Su95, OW05, Gil10, Rag16].
Fix any pmf $P_X$ with $p_\star \triangleq \min_{x \in X} P_X(x) > 0$, and any channel $P_{Y|X}$.

**Theorem (Contraction Coefficient Bound)**

If $f$ satisfies certain “regularity conditions,” then:

$$\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) p_\star} \eta_{\chi^2}(P_X, P_{Y|X})$$

**Example:** This holds for *Hellinger divergences* of order $\alpha \in (0, 2]\backslash\{1\}$, i.e. $f(t) = \frac{t^{\alpha-1}}{\alpha-1}$. 
Fix any pmf $P_X$ with $p_* \triangleq \min_{x \in X} P_X(x) > 0$, and any channel $P_{Y|X}$.

**Theorem (Contraction Coefficient Bound)**

If $f$ satisfies certain “regularity conditions,” then:

$$
\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f'''(1) p_*} \eta_{\chi^2}(P_X, P_{Y|X}).
$$

**Example:** This holds for *Hellinger divergences* of order $\alpha \in (0, 2]\backslash\{1\}$, i.e. $f(t) = \frac{t^{\alpha-1}}{\alpha-1}$. What about $\alpha = 1$?
Fix any pmf \( P_X \) with \( p_* \triangleq \min_{x \in X} P_X(x) > 0 \), and any channel \( P_{Y|X} \).

**Theorem (Contraction Coefficient Bound)**

If \( f \) satisfies certain “regularity conditions,” then:

\[
\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) p_*} \eta_{\chi^2}(P_X, P_{Y|X}).
\]

**Corollary (KL Contraction Coefficient Bound)**

\[
\eta_{KL}(P_X, P_{Y|X}) \leq \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{p_*}
\]
Fix any pmf $P_X$ with $p_* \triangleq \min_{x \in X} P_X(x) > 0$, and any channel $P_{Y|X}$.

**Theorem (Contraction Coefficient Bound)**

If $f$ satisfies certain “regularity conditions,” then:

$$
\eta_f(P_X, P_{Y|X}) \leq \frac{f'(1) + f(0)}{f''(1) p_*} \eta_{\chi^2}(P_X, P_{Y|X}).
$$

**Theorem (Refined KL Contraction Coefficient Bound)**

$$
\eta_{KL}(P_X, P_{Y|X}) \leq \frac{2 \eta_{\chi^2}(P_X, P_{Y|X})}{\phi\left(\max_{A \subseteq X} \min\{P_X(A), P_X(A^c)\}\right) p_*} \leq \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{p_*}
$$

where $\phi(p) = \frac{1}{1-2p} \log\left(\frac{1-p}{p}\right)$.

**Proof Idea:** Use bounds between $f$-divergences and $\chi^2$-divergence based on [Su95, OW05, Gil10, Rag16].
Suppose $X, Y \in \{0, 1\}$ such that $X \sim \text{Ber}(\mathbb{P}(X = 1))$ and $P_{Y|X}$ is a binary symmetric channel (BSC) with crossover probability $p \in [0, 1]$. 
Illustration of KL Contraction Coefficient Bounds

Suppose $X, Y \in \{0, 1\}$ such that $X \sim \text{Ber}(\mathbb{P}(X = 1))$ and $P_{Y|X}$ is binary symmetric channel (BSC) with crossover probability $p \in [0, 1]$.

\[ \eta \chi^2(P_X, P_{Y|X}) \]

\[ \eta \leq \chi^2(P_X, P_{Y|X}) \leq 2 \eta \chi^2(P_X, P_{Y|X}) \]

\[ \phi(p) \leq \eta \chi^2(P_X, P_{Y|X}) \]

\[ \mathbb{P}(X = 1) \]

\[ \text{BSC}(p) \]

\[ P(X = 1) \]
Suppose $X, Y \in \{0, 1\}$ such that $X \sim \text{Ber}(\mathbb{P}(X = 1))$ and $P_{Y|X}$ is a binary symmetric channel (BSC) with crossover probability $p \in [0, 1]$.

\[
\eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_{\text{KL}}(P_X, P_{Y|X}) \leq 2 \eta_{\chi^2}(P_X, P_{Y|X}) \phi(p^*)
\]
Illustration of KL Contraction Coefficient Bounds

Suppose $X, Y \in \{0, 1\}$ such that $X \sim \text{Ber}(P(X = 1))$ and $P_{Y|X}$ is a binary symmetric channel (BSC) with crossover probability $p \in [0, 1]$.

\[
\eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_{KL}(P_X, P_{Y|X}) \leq \frac{2 \eta_{\chi^2}(P_X, P_{Y|X})}{\phi(p_\star) p_\star}
\]
Suppose $X, Y \in \{0, 1\}$ such that $X \sim \text{Ber}(\mathbb{P}(X = 1))$ and $P_{Y|X}$ is a binary symmetric channel (BSC) with crossover probability $p \in [0, 1]$.

\[
\eta_{\chi^2}(P_X, P_{Y|X}) \leq \eta_{\text{KL}}(P_X, P_{Y|X}) \leq \frac{2\eta_{\chi^2}(P_X, P_{Y|X})}{\phi(p^\star)p^\star} \leq \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{p^\star}
\]
Outline

1 Introduction

2 Contraction Coefficients and Strong Data Processing Inequalities

3 Extension using Comparison of Channels
   - Motivation and Main Results
   - Equivalent Characterizations of Less Noisy Preorder
   - Conditions for Less Noisy Domination by Symmetric Channels
   - Less Noisy Domination and Logarithmic Sobolev Inequalities

4 Modal Decomposition of Mutual $\chi^2$-Information

5 Information Contraction in Networks: Broadcasting on DAGs

6 Conclusion
Definition (Less Noisy Preorder [KM77])

\[ P_{Y|X} = W \text{ is less noisy than } P_{Z|X} = V, \text{ denoted } W \succeq_{\ln} V, \text{ if and only if:} \]

\[
D(P_X W \parallel Q_X W) \geq D(P_X V \parallel Q_X V)
\]

for every pair of input distributions \( P_X \) and \( Q_X \).
Main Results

- Test $\geq_{\ln}$ using different divergence measure?
Main Results

- Test $\succeq_{ln}$ using different divergence measure?
  - Yes, any non-linear operator convex $f$-divergence, e.g. $\chi^2$-divergence
Test $\geq_{\ln}$ using different divergence measure?

Yes, any non-linear operator convex $f$-divergence, e.g. $\chi^2$-divergence

Sufficient conditions for $\geq_{\ln}$ domination by symmetric channels?
Main Results

- Test $\geq_{\ln}$ using different divergence measure? 
  Yes, any non-linear operator convex $f$-divergence, e.g. $\chi^2$-divergence

- Sufficient conditions for $\geq_{\ln}$ domination by symmetric channels? 
  Yes
  - degradation criterion for general channels
  - stronger criterion for additive noise channels
Main Results

- Test $\geq_{\ln}$ using different divergence measure?
  **Yes**, any non-linear operator convex $f$-divergence, e.g. $\chi^2$-divergence

- Sufficient conditions for $\geq_{\ln}$ domination by symmetric channels?
  **Yes**
  - degradation criterion for general channels
  - stronger criterion for additive noise channels

- Why $\geq_{\ln}$ domination by symmetric channels?
Main Results

- Test $\geq_{\ln}$ using different divergence measure?
  Yes, any non-linear operator convex $f$-divergence, e.g. $\chi^2$-divergence

- Sufficient conditions for $\geq_{\ln}$ domination by symmetric channels?
  Yes
  - degradation criterion for general channels
  - stronger criterion for additive noise channels

- Why $\geq_{\ln}$ domination by symmetric channels?
  - extend SDPIs because we love information theory
  - $\geq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequality
Motivation: Extend SDPI

SDPI for KL divergence [AG76]:
For any channel $V$, for all pairs of pmfs $P_X, Q_X$:

$$\eta_{KL}(V) D(P_X \| Q_X) \geq D(P_X V \| Q_X V)$$

where $\eta_{KL}(V) \in [0, 1]$ is the contraction coefficient.
Motivation: Extend SDPI

SDPI for KL divergence [AG76]:
For any channel \( V \), for all pairs of pmfs \( P_X, Q_X \):

\[
\eta_{KL}(V) D(P_X \| Q_X) \geq D(P_X V \| Q_X V)
\]

where \( \eta_{KL}(V) \in [0, 1] \) is the contraction coefficient.

Relation to Erasure Channels [PW17]:

- **Definition:** \( q \)-ary erasure channel \( q\text{-EC}(1 - \eta) \)
  
erases input w.p. \( 1 - \eta \), and reproduces input w.p. \( \eta \).
Motivation: Extend SDPI

SDPI for KL divergence [AG76]:
For any channel $V$, for all pairs of pmfs $P_X, Q_X$:

$$\eta_{KL}(V) D(P_X \parallel Q_X) \geq D(P_X V \parallel Q_X V)$$

where $\eta_{KL}(V) \in [0, 1]$ is the contraction coefficient.

Relation to Erasure Channels [PW17]:

- **Definition**: $q$-ary erasure channel $q$-$EC(1 - \eta)$
  erases input w.p. $1 - \eta$, and reproduces input w.p. $\eta$.

- **Prop [PW17]**:
  $$q$-EC(1 - \eta) \succeq_{V} V \iff \forall P_X, Q_X, \eta D(P_X \parallel Q_X) \geq D(P_X V \parallel Q_X V).$$
Motivation: Extend SDPI

SDPI for KL divergence [AG76]:
For any channel $V$, for all pairs of pmfs $P_X, Q_X$:

$$\eta_{KL}(V) D(P_X \parallel Q_X) \geq D(P_X V \parallel Q_X V)$$

where $\eta_{KL}(V) \in [0,1]$ is the contraction coefficient.

Relation to Erasure Channels [PW17]:

- **Definition:** $q$-ary erasure channel $q$-EC$(1 - \eta)$ erases input w.p. $1 - \eta$, and reproduces input w.p. $\eta$.

- **Prop [PW17]:**

  $$q$-EC$(1 - \eta) \preceq_{\text{in}} V \iff \forall P_X, Q_X, \eta D(P_X \parallel Q_X) \geq D(P_X V \parallel Q_X V).$$

SDPI $\iff \preceq_{\text{in}}$ domination by erasure channel
Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq \ln V$?
Main Question

Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq \ln V$?

Definition ($q$-ary Symmetric Channel)

Channel matrix:

$$
W_\delta \triangleq \begin{bmatrix}
1 - \delta & \frac{\delta}{q-1} & \cdots & \frac{\delta}{q-1} \\
\frac{\delta}{q-1} & 1 - \delta & \cdots & \frac{\delta}{q-1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\delta}{q-1} & \frac{\delta}{q-1} & \cdots & 1 - \delta
\end{bmatrix}
$$

where $\delta \in [0, 1]$ is the total crossover probability.
Main Question

Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in [0, \frac{q-1}{q}]$ such that $W_\delta \succeq_{\text{in}} V$?

Definition ($q$-ary Symmetric Channel)

Channel matrix:

$$W_\delta \triangleq \begin{bmatrix}
1 - \delta & \frac{\delta}{q-1} & \cdots & \frac{\delta}{q-1} \\
\frac{\delta}{q-1} & 1 - \delta & \cdots & \frac{\delta}{q-1} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\delta}{q-1} & \frac{\delta}{q-1} & \cdots & 1 - \delta
\end{bmatrix}$$

where $\delta \in [0, 1]$ is the total crossover probability.

Remark: For every channel $V$, $W_0 \succeq_{\text{in}} V$ and $V \succeq_{\text{in}} W_{(q-1)/q}$. 
Outline

1 Introduction

2 Contraction Coefficients and Strong Data Processing Inequalities

3 Extension using Comparison of Channels
   - Motivation and Main Results
   - Equivalent Characterizations of Less Noisy Preorder
   - Conditions for Less Noisy Domination by Symmetric Channels
   - Less Noisy Domination and Logarithmic Sobolev Inequalities

4 Modal Decomposition of Mutual $\chi^2$-Information

5 Information Contraction in Networks: Broadcasting on DAGs

6 Conclusion
Operator Convexity

\[ f : \mathbb{R} \rightarrow \mathbb{R} \] can be applied to an \( n \times n \) Hermitian matrix \( A \) via:

\[ f(A) = U \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)) U^H \]

where \( A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^H \), \( \lambda_i \) are eigenvalues, and \( U \) is unitary.
Operator Convexity

$f : \mathbb{R} \rightarrow \mathbb{R}$ can be applied to an $n \times n$ Hermitian matrix $A$ via:

$$f(A) = U \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)) U^H$$

where $A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^H$, $\lambda_i$ are eigenvalues, and $U$ is unitary.

**Definition (Operator Convexity)**

$f : \mathbb{R} \rightarrow \mathbb{R}$ is **operator convex** if for every $n$, every pair of $n \times n$ Hermitian matrices $A, B$, and every $\lambda \in [0, 1]$:

$$\lambda f(A) + (1 - \lambda) f(B) \succeq_{\text{PSD}} f(\lambda A + (1 - \lambda) B)$$

where $\succeq_{\text{PSD}}$ is the Löwner partial order.
Operator Convexity

\( f : \mathbb{R} \to \mathbb{R} \) can be applied to an \( n \times n \) Hermitian matrix \( A \) via:

\[
    f(A) = U \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)) U^H
\]

where \( A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^H \), \( \lambda_i \) are eigenvalues, and \( U \) is unitary.

**Definition (Operator Convexity)**

\( f : \mathbb{R} \to \mathbb{R} \) is operator convex if for every \( n \), every pair of \( n \times n \) Hermitian matrices \( A, B \), and every \( \lambda \in [0, 1] \):

\[
    \lambda f(A) + (1 - \lambda) f(B) \succeq_{\text{PSD}} f(\lambda A + (1 - \lambda) B)
\]

where \( \succeq_{\text{PSD}} \) is the Löwner partial order.

**Löwner-Heinz Theorem (Examples [Löw34, Hei51])**

- For every \( \alpha \in (0, 2] \setminus \{ 1 \} \), \( f : (0, \infty) \to \mathbb{R} \), \( f(t) = \frac{t^{\alpha-1}}{\alpha-1} \) is operator convex.
- \( f : (0, \infty) \to \mathbb{R} \), \( f(t) = t \log(t) \) is operator convex.
Operator Convexity

\( f : \mathbb{R} \to \mathbb{R} \) can be applied to an \( n \times n \) Hermitian matrix \( A \) via:

\[
f(A) = U \text{diag}(f(\lambda_1), \ldots, f(\lambda_n)) U^H
\]

where \( A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^H \), \( \lambda_i \) are eigenvalues, and \( U \) is unitary.

**Definition (Operator Convexity)**

\( f : \mathbb{R} \to \mathbb{R} \) is operator convex if for every \( n \), every pair of \( n \times n \) Hermitian matrices \( A, B \), and every \( \lambda \in [0, 1] \):

\[
\lambda f(A) + (1 - \lambda) f(B) \succeq_{\text{PSD}} f(\lambda A + (1 - \lambda) B)
\]

where \( \succeq_{\text{PSD}} \) is the Löwner partial order.

**Löwner-Heinz Theorem (Examples [Löw34, Hei51])**

- For every \( \alpha \in (0, 2] \setminus \{1\} \), \( f : (0, \infty) \to \mathbb{R}, f(t) = \frac{t^{\alpha-1}}{\alpha-1} \) is operator convex. (Hellinger divergence of order \( \alpha \), \( \chi^2 \)-divergence)

- \( f : (0, \infty) \to \mathbb{R}, f(t) = t \log(t) \) is operator convex. (KL divergence)
Theorem (Equivalent Characterizations of $\succeq_{\ln}$)

Given channels $W$ and $V$, and any non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$:

$$W \succeq_{\ln} V \iff \forall P_X, Q_X, \ D_f(P_X W \parallel Q_X W) \geq D_f(P_X V \parallel Q_X V)$$

Remarks:
Proof uses Löwner's integral representation [CRS94].
PSD characterization follows from [vDi97].
Characterization of Less Noisy using Operator Convexity

Theorem (Equivalent Characterizations of $\geq_{\ln}$)

Given channels $W$ and $V$, and any non-linear operator convex function $f : (0, \infty) \rightarrow \mathbb{R}$ such that $f(1) = 0$:

$$W \geq_{\ln} V \iff \forall P_X, Q_X, D_f(P_X W \| Q_X W) \geq D_f(P_X V \| Q_X V)$$

Remarks:

- Proof uses Löwner’s integral representation [CRS94].
Theorem (Equivalent Characterizations of $\succeq_{\ln}$)

Given channels $W$ and $V$, and any non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$:

$$W \succeq_{\ln} V \iff \forall P_X, Q_X, D_f(P_X W \parallel Q_X W) \geq D_f(P_X V \parallel Q_X V)$$
$$\iff \forall P_X, Q_X, \chi^2(P_X W \parallel Q_X W) \geq \chi^2(P_X V \parallel Q_X V)$$

Remarks:

- Proof uses Löwner’s integral representation [CRS94].
Theorem (Equivalent Characterizations of $\succeq_{\text{ln}}$)

Given channels $W$ and $V$, and any non-linear operator convex function $f : (0, \infty) \rightarrow \mathbb{R}$ such that $f(1) = 0$:

\[
W \succeq_{\text{ln}} V \iff \forall P_X, Q_X, D_f(P_X W \| Q_X W) \geq D_f(P_X V \| Q_X V)
\]
\[
\iff \forall P_X, Q_X, \chi^2(P_X W \| Q_X W) \geq \chi^2(P_X V \| Q_X V)
\]
\[
\iff \forall Q_X, W \text{diag}(Q_X W)^{-1} W^T \succeq_{\text{PSD}} V \text{diag}(Q_X V)^{-1} V^T
\]

Remarks:

- Proof uses Löwner’s integral representation [CRS94].
- Let $J_X = P_X - Q_X$. Then, we have:

\[
\chi^2(P_X W \| Q_X W) = J_X W \text{diag}(Q_X W)^{-1} W^T J_X^T.
\]
Theorem (Equivalent Characterizations of $\succeq_{\ln}$)

Given channels $W$ and $V$, and any non-linear operator convex function $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$:

$$W \succeq_{\ln} V \iff \forall P_X, Q_X, D_f(P_X W \| Q_X W) \geq D_f(P_X V \| Q_X V)$$

$$\iff \forall P_X, Q_X, \chi^2(P_X W \| Q_X W) \geq \chi^2(P_X V \| Q_X V)$$

$$\iff \forall Q_X, W \text{diag}(Q_X W)^{-1} W^T \succeq_{\text{PSD}} V \text{diag}(Q_X V)^{-1} V^T$$

Remarks:

- Proof uses Löwner’s integral representation [CRS94].
- PSD characterization follows from [vDi97].
1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels
   - Motivation and Main Results
   - Equivalent Characterizations of Less Noisy Preorder
   - Conditions for Less Noisy Domination by Symmetric Channels
   - Less Noisy Domination and Logarithmic Sobolev Inequalities

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Condition for Degradation by Symmetric Channels

Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq_{\ln} V$?
Given channel \( V \), find \( q \)-ary symmetric channel \( W_\delta \) with largest \( \delta \in [0, \frac{q-1}{q}] \) such that \( W_\delta \succeq \ln V \).

Definition (Degradation [Bla51, She51, Ste51, Cov72, Ber73]): \( V \) is degraded version of \( W \), denoted \( W \succeq_{\text{deg}} V \), if \( V = WA \) for some channel \( A \).
Condition for Degradation by Symmetric Channels

Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq_{\ln} V$?

- **Definition** (Degradation [Bla51, She51, Ste51, Cov72, Ber73]): $V$ is degraded version of $W$, denoted $W \succeq_{\text{deg}} V$, if $V = WA$ for some channel $A$.

- **Prop:** $W \succeq_{\text{deg}} V \implies W \succeq_{\ln} V$. 

---

Anuran Makur (MIT)  
Information Contraction & Decomposition  
15 May 2019  
23 / 64
Condition for Degradation by Symmetric Channels

Given channel $V$, find $q$-ary symmetric channel $W_\delta$ with largest $\delta \in \left[0, \frac{q-1}{q}\right]$ such that $W_\delta \succeq \ln V$?

- **Definition (Degradation [Bla51, She51, Ste51, Cov72, Ber73]):** $V$ is degraded version of $W$, denoted $W \succeq_{\text{deg}} V$, if $V = WA$ for some channel $A$.

- **Prop:** $W \succeq_{\text{deg}} V \implies W \succeq_{\ln} V$.

**Theorem (Degradation by Symmetric Channels)**

For channel $V$ with common input and output alphabet, and minimum probability entry $\nu = \min\{[V]_{i,j} : 1 \leq i, j \leq q\}$:

$$0 \leq \delta \leq \frac{\nu}{1 - (q - 1)\nu + \frac{\nu}{q-1}} \implies W_\delta \succeq_{\text{deg}} V.$$
Theorem (Degradation by Symmetric Channels)

For channel $V$ with common input and output alphabet, and minimum probability entry $\nu = \min\{[V]_{i,j} : 1 \leq i, j \leq q\}$:

$$0 \leq \delta \leq \frac{\nu}{1 - (q - 1)\nu + \frac{\nu}{q-1}} \implies W_\delta \succeq_{\text{deg}} V.$$ 

Remark: Condition is tight when no further information about $V$ known. For example, suppose:

$$V = \begin{bmatrix}
\nu & 1 - (q - 1)\nu & \nu & \cdots & \nu \\
1 - (q - 1)\nu & \nu & \nu & \cdots & \nu \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 - (q - 1)\nu & \nu & \nu & \cdots & \nu
\end{bmatrix}.$$ 

Then, $0 \leq \delta \leq \nu/(1 - (q - 1)\nu + \frac{\nu}{q-1}) \iff W_\delta \succeq_{\text{deg}} V$. 
Additive Noise Channels

- Fix Abelian group $(\mathcal{X}, \oplus)$ with order $q$ as alphabet.
Additive Noise Channels

- Fix Abelian group \((\mathcal{X}, \oplus)\) with order \(q\) as alphabet.
- Additive noise channel:

\[
Y = X \oplus Z, \quad X \perp \perp Z
\]

where \(X, Y, Z \in \mathcal{X}\) are input, output, and noise random variables.
Additive Noise Channels

- Fix Abelian group \((\mathcal{X}, \oplus)\) with order \(q\) as alphabet.
- Additive noise channel:
  \[
  Y = X \oplus Z, \quad X \perp Z
  \]
  where \(X, Y, Z \in \mathcal{X}\) are input, output, and noise random variables.
- Channel probabilities given by noise pmf \(P_Z\):
  \[
  \forall x, y \in \mathcal{X}, \quad P_{Y|X}(y|x) = P_Z(-x \oplus y).
  \]
Additive Noise Channels

- Fix Abelian group $(\mathcal{X}, \oplus)$ with order $q$ as alphabet.
- Additive noise channel:

\[ Y = X \oplus Z, \quad X \perp Z \]

where $X, Y, Z \in \mathcal{X}$ are input, output, and noise random variables.
- Channel probabilities given by noise pmf $P_Z$:

\[ \forall x, y \in \mathcal{X}, \quad P_{Y|X}(y|x) = P_Z(-x \oplus y). \]

- $P_Y$ is convolution of $P_X$ and $P_Z$:

\[ \forall y \in \mathcal{X}, \quad P_Y(y) = (P_X * P_Z)(y) \triangleq \sum_{x \in \mathcal{X}} P_X(x)P_Z(-x \oplus y). \]
Additive Noise Channels

- Fix Abelian group \((\mathcal{X}, \oplus)\) with order \(q\) as alphabet.
- Additive noise channel:

\[
Y = X \oplus Z, \quad X \perp Z
\]

where \(X, Y, Z \in \mathcal{X}\) are input, output, and noise random variables.

- Channel probabilities given by noise pmf \(P_Z\):

\[
\forall x, y \in \mathcal{X}, \quad P_{Y|X}(y|x) = P_Z(-x \oplus y).
\]

- \(P_Y\) is convolution of \(P_X\) and \(P_Z\):

\[
\forall y \in \mathcal{X}, \quad P_Y(y) = (P_X \ast P_Z)(y) \triangleq \sum_{x \in \mathcal{X}} P_X(x)P_Z(-x \oplus y).
\]

- \(q\)-ary symmetric channel: \(P_Z = \left(1 - \delta, \frac{\delta}{q-1}, \ldots, \frac{\delta}{q-1}\right)\) for \(\delta \in [0, 1]\)

\[
(\ast P_Z) = W_\delta
\]
Fix $q$-ary symmetric channel $W_\delta$ with $\delta \in [0, 1]$. 
Fix $q$-ary symmetric channel $W_\delta$ with $\delta \in [0, 1]$.

More noisy region of $W_\delta$ is:

$$more-noisy(W_\delta) \triangleq \{ P_Z : W_\delta \succeq_{\ln} (\cdot * P_Z) \}.$$
More Noisy and Degradation Regions

- Fix $q$-ary symmetric channel $W_\delta$ with $\delta \in [0, 1]$.

- More noisy region of $W_\delta$ is:

  $\text{more-noisy}(W_\delta) \triangleq \{ P_Z : W_\delta \succeq_{\ln} (\cdot \ast P_Z) \}$.

- Degradation region of $W_\delta$ is:

  $\text{degrade}(W_\delta) \triangleq \{ P_Z : W_\delta \succeq_{\text{deg}} (\cdot \ast P_Z) \}$. 
Theorem (More Noisy and Degradation Regions)

For $W_\delta$ with $\delta \in \left[0, \frac{q-1}{q}\right]$ and $q \geq 2$:

$$\text{degrade}(W_\delta) = \text{conv}(\text{rows of } W_\delta) \subseteq \text{conv}(\text{rows of } W_\delta \text{ and } W_\gamma) \subseteq \text{more-noisy}(W_\delta) \subseteq \{P_Z : \|P_Z - u\|_2 \leq \|w_\delta - u\|_2\}$$

where $\text{conv}(\cdot)$ denotes convex hull, $\gamma = (1 - \delta)/(1 - \delta + \frac{\delta}{(q-1)^2})$, $u$ is the uniform pmf, and $w_\delta$ is first row of $W_\delta$. 
Theorem (More Noisy and Degradation Regions)

For $W_\delta$ with $\delta \in \left[0, \frac{q-1}{q}\right]$ and $q \geq 2$:

$$\text{degrade}(W_\delta) = \text{conv}\left(\text{rows of } W_\delta\right) \subseteq \text{conv}\left(\text{rows of } W_\delta \text{ and } W_\gamma\right) \subseteq \text{more-noisy}(W_\delta) \subseteq \{P_Z : \|P_Z - u\|_2 \leq \|w_\delta - u\|_2\}$$

where $\text{conv}(\cdot)$ denotes convex hull, $\gamma = (1 - \delta)/(1 - \delta + \frac{\delta}{(q-1)^2})$, $u$ is the uniform pmf, and $w_\delta$ is first row of $W_\delta$.

Furthermore, $\text{more-noisy}(W_\delta)$ is closed, convex, and invariant under permutations corresponding to $(\mathcal{X}, \oplus)$. 
Illustration of the $q = 3$ case:
Illustration of the $q = 3$ case:
Illustration of the $q = 3$ case:

$d_{\text{degrade}}(W_\delta)$

$w_\delta = (1 - \delta, \frac{\delta}{2}, \frac{\delta}{2})$

$w_1 = (0, \frac{1}{2}, \frac{1}{2})$

$w_0 = (1, 0, 0)$

lower bound

symmetric channels

probability simplex
Illustration of the $q = 3$ case:
Illustration of the $q = 3$ case:
Outline

1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels
   - Motivation and Main Results
   - Equivalent Characterizations of Less Noisy Preorder
   - Conditions for Less Noisy Domination by Symmetric Channels
   - Less Noisy Domination and Logarithmic Sobolev Inequalities

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Consider irreducible Markov chain $V$ with uniform stationary pmf $u$ on state space of size $q$. 

Log-Sobolev inequality with constant $\alpha \geq 0$: 

For every $f \in \mathbb{R}^q$ such that $f^T f = q$: 

$$D(\|f\|^2 u || u) = \frac{1}{q} \sum_{i=1}^q f_i^2 \log(f_i^2) \leq \frac{1}{\alpha} E_V(f, f).$$ 

Log-Sobolev constant – largest $\alpha$ satisfying log-Sobolev inequality.
Consider irreducible Markov chain $V$ with uniform stationary pmf $u$ on state space of size $q$.

Dirichlet form $\mathcal{E}_V : \mathbb{R}^q \times \mathbb{R}^q \rightarrow [0, \infty)$

$$\mathcal{E}_V(f, f) \triangleq \frac{1}{q} f^T \left( I - \frac{V + V^T}{2} \right) f$$
Consider irreducible Markov chain $V$ with uniform stationary pmf $u$ on state space of size $q$.

Dirichlet form $\mathcal{E}_V : \mathbb{R}^q \times \mathbb{R}^q \to [0, \infty)$

$$\mathcal{E}_V(f, f) \triangleq \frac{1}{q} f^T \left( I - \frac{V + V^T}{2} \right) f$$

Log-Sobolev inequality with constant $\alpha \geq 0$:
For every $f \in \mathbb{R}^q$ such that $f^T f = q$:

$$D(f^2 u \| u) = \frac{1}{q} \sum_{i=1}^{q} f_i^2 \log(f_i^2) \leq \frac{1}{\alpha} \mathcal{E}_V(f, f).$$
Consider irreducible Markov chain $V$ with uniform stationary pmf $u$ on state space of size $q$.

Dirichlet form $\mathcal{E}_V : \mathbb{R}^q \times \mathbb{R}^q \rightarrow [0, \infty)$

$$\mathcal{E}_V(f, f) \triangleq \frac{1}{q} f^T \left( I - \frac{V + V^T}{2} \right) f$$

Log-Sobolev inequality with constant $\alpha \geq 0$:
For every $f \in \mathbb{R}^q$ such that $f^T f = q$:

$$D(f^2 u \| u) = \frac{1}{q} \sum_{i=1}^{q} f_i^2 \log(f_i^2) \leq \frac{1}{\alpha} \mathcal{E}_V(f, f).$$

Log-Sobolev constant – largest $\alpha$ satisfying log-Sobolev inequality.
Comparison of Dirichlet Forms

- **Standard Dirichlet form:**

\[
\mathcal{E}_{\text{std}}(f, f) \triangleq \nabla \mathbf{A} \nabla u(f) = \sum_{i=1}^{q} \frac{1}{q} f_i^2 - \left( \sum_{i=1}^{q} \frac{1}{q} f_i \right)^2
\]

For standard Dirichlet form, \( \mathcal{E}_{\text{std}}(f, f) \triangleq \nabla \mathbf{A} \nabla u(f) \), the log-Sobolev constant known [DSC96]:

\[
D(f^2_u | | u) \leq \left( q - 2 \right) \mathcal{E}_{\text{std}}(f, f)
\]

for all \( f \in \mathbb{R}^q \) with \( f^T f = q \).

**Theorem (Domination of Dirichlet Forms)**

For channels \( W_{\delta} \) and \( V \) with \( \delta \in \left[ 0, q - 1 \right] \) and stationary pmf \( u \):

\[
W_{\delta} \succeq \ln V \Rightarrow \mathcal{E}(V) \geq q \delta (q - 1) \mathcal{E}_{\text{std}} \text{ pointwise}
\]

\[
W_{\delta} \succeq \ln V \Rightarrow \text{log-Sobolev inequality for } V:
\]

\[
D(f^2_u | | u) \leq \left( q - 2 \right) \mathcal{E}(V)\left( q - \frac{1}{2} \right)
\]

for every \( f \in \mathbb{R}^q \) satisfying \( f^T f = q \).
For standard Dirichlet form, $\mathcal{E}_{\text{std}}(f, f) \triangleq \text{VAR}_u(f)$, log-Sobolev constant known [DSC96]:

$$D(f^2 u \parallel u) \leq \frac{q \log(q - 1)}{(q - 2)} \mathcal{E}_{\text{std}}(f, f)$$

for all $f \in \mathbb{R}^q$ with $f^T f = q$. 
Comparison of Dirichlet Forms

- For standard Dirichlet form, $\mathcal{E}_{\text{std}}(f, f) \triangleq \text{VAR}_u(f)$, log-Sobolev constant known [DSC96]:

$$D(f^2 u \| u) \leq \frac{q \log(q - 1)}{(q - 2)} \mathcal{E}_{\text{std}}(f, f)$$

for all $f \in \mathbb{R}^q$ with $f^T f = q$.

Theorem (Domination of Dirichlet Forms)

For channels $W_\delta$ and $V$ with $\delta \in \left[0, \frac{q-1}{q}\right]$ and stationary pmf $u$:

$$W_\delta \succeq_{\text{ln}} V \Rightarrow \mathcal{E}_V \geq \frac{q \delta}{q - 1} \mathcal{E}_{\text{std}} \text{ pointwise}.$$
Comparison of Dirichlet Forms

- For standard Dirichlet form, $\mathcal{E}_{\text{std}}(f, f) \triangleq \text{VAR}_u(f)$, log-Sobolev constant known [DSC96]:
  \[
  D(f^2 u \parallel u) \leq \frac{q \log(q - 1)}{(q - 2)} \mathcal{E}_{\text{std}}(f, f)
  \]
  for all $f \in \mathbb{R}^q$ with $f^T f = q$.

Theorem (Domination of Dirichlet Forms)

For channels $W_{\delta}$ and $V$ with $\delta \in \left[0, \frac{q-1}{q}\right]$ and stationary pmf $u$:

$W_{\delta} \preceq_{\ln} V \Rightarrow \mathcal{E}_V \geq \frac{q\delta}{q-1} \mathcal{E}_{\text{std}}$ pointwise.

- $W_{\delta} \preceq_{\ln} V \Rightarrow$ log-Sobolev inequality for $V$:
  \[
  D(f^2 u \parallel u) \leq \frac{(q - 1) \log(q - 1)}{\delta (q - 2)} \mathcal{E}_V(f, f)
  \]
  for every $f \in \mathbb{R}^q$ satisfying $f^T f = q$. 
Outline

1. Introduction
2. Contraction Coefficients and Strong Data Processing Inequalities
3. Extension using Comparison of Channels
4. Modal Decomposition of Mutual $\chi^2$-Information
   - Maximal Correlation and Conditional Expectation Operators
   - Embedding Data using Modal Decompositions
   - Algorithm for Information Decomposition
5. Information Contraction in Networks: Broadcasting on DAGs
6. Conclusion
Maximal Correlation and Contraction Coefficients

Definition (Maximal Correlation [Hir35, Geb41, Sar58, Rén59])

Maximal correlation between random variables \( X \in \mathcal{X} \) and \( Y \in \mathcal{Y} \) is:

\[
\rho_{\text{max}}(X; Y) \triangleq \max_{f, g} \mathbb{E}[f(X)g(Y)]
\]

where maximization is over all \( f : \mathcal{X} \rightarrow \mathbb{R} \) and \( g : \mathcal{Y} \rightarrow \mathbb{R} \) such that \( \mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0 \) and \( \mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1 \).
Maximal Correlation and Contraction Coefficients

**Definition (Maximal Correlation [Hir35, Geb41, Sar58, Rén59])**

Maximal correlation between random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ is:

$$\rho_{\text{max}}(X; Y) \triangleq \max_{f, g} \mathbb{E}[f(X)g(Y)]$$

where maximization is over all $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$.

**Prop (Maximal Correlation as Contraction Coefficient [Sar58])**

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\text{max}}(X; Y)^2$$
Maximal Correlation and Contraction Coefficients

**Definition (Maximal Correlation [Hir35, Geb41, Sar58, Rén59])**

Maximal correlation between random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ is:

$$\rho_{\text{max}}(X; Y) \triangleq \max_{f, g} \mathbb{E}[f(X)g(Y)] = \max_{f, g} \mathbb{E}[g(Y)\mathbb{E}[f(X) | Y]]$$

where maximization is over all $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Y} \rightarrow \mathbb{R}$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$.

**Prop (Maximal Correlation as Contraction Coefficient [Sar58])**

$$\eta_{\chi^2}(P_X, P_Y | X) = \rho_{\text{max}}(X; Y)^2$$

- $\rho_{\text{max}}(X; Y)$ is singular value of conditional expectation operator $\mathbb{E}[\cdot | Y]$ and optimizing functions are singular vectors [Hir35, Rén59].
Maximal Correlation and Contraction Coefficients

Definition (Maximal Correlation [Hir35, Geb41, Sar58, Rén59])

Maximal correlation between random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ is:

$$\rho_{\text{max}}(X; Y) \triangleq \max_{f, g} \mathbb{E}[f(X)g(Y)]$$

where maximization is over all $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$.

Prop (Maximal Correlation as Contraction Coefficient [Sar58])

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho_{\text{max}}(X; Y)^2$$

- $\rho_{\text{max}}(X; Y)$ is singular value of $\mathbb{E}[\cdot|Y]$ [Hir35, Rén59].
- SVD structure of $\mathbb{E}[\cdot|Y] \Rightarrow$ SDPI for $\chi^2$-divergence
Maximal Correlation and Contraction Coefficients

Definition (Maximal Correlation [Hir35, Geb41, Sar58, Rén59])

Maximal correlation between random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ is:

$$\rho_{\text{max}}(X; Y) \triangleq \max_{f, g} \mathbb{E}[f(X)g(Y)]$$

where maximization is over all $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Y} \rightarrow \mathbb{R}$ such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f(X)^2] = \mathbb{E}[g(Y)^2] = 1$.

Prop (Maximal Correlation as Contraction Coefficient [Sar58])

$$\eta_{\chi^2}(P_X, P_Y|X) = \rho_{\text{max}}(X; Y)^2$$

- $\rho_{\text{max}}(X; Y)$ is singular value of $\mathbb{E}[\cdot|Y]$ [Hir35, Rén59].
- SVD structure of $\mathbb{E}[\cdot|Y] \Rightarrow$ SDPI for $\chi^2$-divergence
- Singular vectors of $\mathbb{E}[\cdot|Y] \Rightarrow$ feature functions for embedding
Conditional Expectation Operators

Fix bivariate distribution $P_{X,Y}$ such that $P_X > 0$ and $P_Y > 0$. 
Fix bivariate distribution $P_{X,Y}$ such that $P_X > 0$ and $P_Y > 0$.

**Hilbert Spaces:**

\[ L^2(\mathcal{X}, P_X) \triangleq \left\{ f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}[f(X)^2] < +\infty \right\} \] with inner product:

\[ \forall f, f' \in L^2(\mathcal{X}, P_X), \quad \langle f, f' \rangle_{P_X} \triangleq \mathbb{E}[f(X)f'(X)]. \]
Conditional Expectation Operators

Fix bivariate distribution $P_{X,Y}$ such that $P_X > 0$ and $P_Y > 0$.

**Hilbert Spaces:**

$L^2(\mathcal{X}, P_X) \triangleq \{ f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}[f(X)^2] < +\infty \}$ with inner product:

$$\forall f, f' \in L^2(\mathcal{X}, P_X), \quad \langle f, f' \rangle_{P_X} \triangleq \mathbb{E}[f(X)f'(X)].$$

$L^2(\mathcal{Y}, P_Y) \triangleq \{ g : \mathcal{Y} \to \mathbb{R} \mid \mathbb{E}[g(Y)^2] < +\infty \}$ with inner product:

$$\forall g, g' \in L^2(\mathcal{Y}, P_Y), \quad \langle g, g' \rangle_{P_Y} \triangleq \mathbb{E}[g(Y)g'(Y)].$$
Conditional Expectation Operators

Fix bivariate distribution $P_{X,Y}$ such that $P_X > 0$ and $P_Y > 0$.

**Hilbert Spaces:**

$L^2(\mathcal{X}, P_X) \triangleq \{ f : \mathcal{X} \to \mathbb{R} \mid \mathbb{E}[f(X)^2] < +\infty \}$ with inner product:

$$\forall f, f' \in L^2(\mathcal{X}, P_X), \quad \langle f, f' \rangle_{P_X} \triangleq \mathbb{E}[f(X)f'(X)].$$

$L^2(\mathcal{Y}, P_Y) \triangleq \{ g : \mathcal{Y} \to \mathbb{R} \mid \mathbb{E}[g(Y)^2] < +\infty \}$ with inner product:

$$\forall g, g' \in L^2(\mathcal{Y}, P_Y), \quad \langle g, g' \rangle_{P_Y} \triangleq \mathbb{E}[g(Y)g'(Y)].$$

**Definition (Conditional Expectation Operator)**

$C : L^2(\mathcal{X}, P_X) \rightarrow L^2(\mathcal{Y}, P_Y)$ maps $f \in L^2(\mathcal{X}, P_X)$ to $C(f) \in L^2(\mathcal{Y}, P_Y)$:

$$(C(f))(y) \triangleq \mathbb{E}[f(X) \mid Y = y].$$
Singular Value Decomposition (SVD)

SVD of Conditional Expectation Operator: For $1 \leq i \leq \min\{|X|, |Y|\}$,

$$C(f_i) = \sigma_ig_i$$

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|X|, |Y|\}} \geq 0$ are singular values,
- $\{f_1, \ldots, f_{|X|}\} \subseteq L^2(X, P_X)$ are right singular vectors,
- $\{g_1, \ldots, g_{|Y|}\} \subseteq L^2(Y, P_Y)$ are left singular vectors.
Singular Value Decomposition (SVD)

SVD of Conditional Expectation Operator: For $1 \leq i \leq \min\{|X|, |Y|\}$,

$$C(f_i) = \sigma_i g_i$$

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|X|, |Y|\}} \geq 0$ are singular values,
- $\{f_1, \ldots, f_{|X|}\} \subseteq L^2(X, P_X)$ are right singular vectors,
- $\{g_1, \ldots, g_{|Y|}\} \subseteq L^2(Y, P_Y)$ are left singular vectors.

Theorem (SVD Structure)

- **Operator Norm**: $\|C\|_{\text{op}} = \sigma_1 = 1$, and corresponding singular vectors are $f_1 = 1$ and $g_1 = 1$. 
Singular Value Decomposition (SVD)

**SVD of Conditional Expectation Operator:** For $1 \leq i \leq \min\{|X|, |Y|\}$,

$$C(f_i) = \sigma_i g_i$$

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|X|, |Y|\}} \geq 0$ are singular values,
- $\{f_1, \ldots, f_{|X|}\} \subseteq L^2(X, P_X)$ are right singular vectors,
- $\{g_1, \ldots, g_{|Y|}\} \subseteq L^2(Y, P_Y)$ are left singular vectors.

### Theorem (SVD Structure)

- **Operator Norm:** $\|C\|_{\text{op}} = \sigma_1 = 1$, and corresponding singular vectors are $f_1 = 1$ and $g_1 = 1$.
- **Max Correlation [Hir35, Rén59]:** $\sigma_2 = \rho_{\max}(X; Y) = \mathbb{E}[f_2(X)g_2(Y)]$. 
Singular Value Decomposition (SVD)

**SVD of Conditional Expectation Operator:** For $1 \leq i \leq \min\{|X|, |Y|\}$,

$$C(f_i) = \sigma_i g_i$$

- $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{|X|, |Y|\}} \geq 0$ are singular values,
- $\{f_1, \ldots, f_{|X|}\} \subseteq \mathcal{L}^2(X, P_X)$ are right singular vectors,
- $\{g_1, \ldots, g_{|Y|}\} \subseteq \mathcal{L}^2(Y, P_Y)$ are left singular vectors.

**Theorem (SVD Structure)**

- **Operator Norm:** $\|C\|_{\text{op}} = \sigma_1 = 1$, and corresponding singular vectors are $f_1 = 1$ and $g_1 = 1$.
- **Max Correlation [Hir35, Rén59]:** $\sigma_2 = \rho_{\max}(X; Y) = \mathbb{E}[f_2(X)g_2(Y)]$.
- **Courant-Fischer-Weyl:** For $2 \leq k \leq \min\{|X|, |Y|\}$,

$$\sigma_k = \mathbb{E}[f_k(X)g_k(Y)] = \max_{f, g} \mathbb{E}[f(X)g(Y)]$$

where maximization is over unit-norm $f \in \text{span}(f_1, \ldots, f_{k-1})^\perp$ and $g \in \text{span}(g_1, \ldots, g_{k-1})^\perp$. 
Consider $C = \mathbb{E}_{P_{X|Y}}[\cdot|Y] : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ with operator norm:

$$\|C\|_{Q_X \to P_Y}^2 \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X)} \mathbb{E}_{P_Y}\left[\mathbb{E}_{P_{X|Y}}[f(X)|Y]^2\right]$$

$$\mathbb{E}_{Q_X}[f(X)^2] = 1$$
Consider $C = \mathbb{E}_{P_{X|Y}}[\cdot | Y] : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ with operator norm:

$$\|C\|_{Q_X \to P_Y}^2 \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X)} \mathbb{E}_{P_Y}\left[ \mathbb{E}_{P_{X|Y}}[f(X)|Y]^2 \right].$$

**Prop (Inner Product for Contraction Property)**

- $\min_{Q_X} \|C\|_{Q_X \to P_Y}^2 = \|C\|_{P_X \to P_Y}^2 = 1.$

**Remark:** $Q_X^* = P_X$ is *only* inner product that makes $C$ contractive.
Consider $C = \mathbb{E}_{P_{X|Y}}[\cdot | Y] : \mathcal{L}^2(\mathcal{X}, Q_X) \to \mathcal{L}^2(\mathcal{Y}, P_Y)$ with operator norm:

$$\| C \|_{Q_X \to P_Y}^2 \triangleq \max_{f \in \mathcal{L}^2(\mathcal{X}, Q_X): \mathbb{E}_{Q_X}[f(X)^2]=1} \mathbb{E}_{P_Y} \left[ \mathbb{E}_{P_{X|Y}}[f(X)|Y]^2 \right].$$

**Prop (Inner Product for Contraction Property)**

- $\min_{Q_X} \| C \|_{Q_X \to P_Y}^2 = \| C \|_{P_X \to P_Y}^2 = 1.$
- For all $Q_X$, $\| C \|_{Q_X \to P_Y}^2 - 1 \geq \chi^2(P_X || Q_X).$

**Remark:** $Q_X^* = P_X$ is *only* inner product that makes $C$ contractive.
Theorem (Modal Decomposition [Hir35, Lan58])

- Modal decomposition of bivariate distribution:

\[
P_{X,Y}(x,y) = P_X(x) P_Y(y) \left( 1 + \frac{\min\{|X|,|Y|\}}{\sum_{i=2}^{\min\{|X|,|Y|\}} \sigma_i f_i(x) g_i(y)} \right)
\]

where \(\{f_i\},\{g_i\}\) are singular vectors of \(C\), and \(\sigma_i = \mathbb{E}[f_i(X)g_i(Y)]\) are singular values.
Modal Decomposition

**Theorem (Modal Decomposition [Hir35, Lan58])**

- **Modal decomposition of bivariate distribution:**

\[
P_{X,Y}(x, y) = P_X(x) P_Y(y) \left(1 + \sum_{i=2}^{\min\{|X|,|Y|\}} \sigma_i f_i(x) g_i(y) \right)
\]

where \(\{f_i\},\{g_i\}\) are singular vectors of \(C\), and \(\sigma_i = \mathbb{E}[f_i(X)g_i(Y)]\) are singular values.

- **Modal decomposition of mutual \(\chi^2\)-information:**

\[
I_{\chi^2}(X; Y) \triangleq \chi^2(P_{X,Y}||P_X P_Y) = \sum_{i=2}^{\min\{|X|,|Y|\}} \sigma_i^2.
\]
Outline

1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels

4. Modal Decomposition of Mutual $\chi^2$-Information
   - Maximal Correlation and Conditional Expectation Operators
   - Embedding Data using Modal Decompositions
   - Algorithm for Information Decomposition

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$$X = \{x_1, x_2, x_3, \ldots\}$$

$$Y = \{ISIT, Allerton, ICASSP, ICML, \ldots\}$$

Want: Embed $X$ into Euclidean space $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g. clustering.

"Natural" Embedding: Represent each $x \in X$ using conditional distribution $P_{Y|X} = x \in \mathbb{R}^{|Y|}$.

Dimensionality Reduction: $|Y|$ is large! Reduce dimension of embedding.
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$$X = \{ \text{[Images]} \}$$

$$Y = \{ \text{ISIT, Allerton, ICASSP, ICML, \ldots} \}$$
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g. $X = \{\text{ISIT, Allerton, ICASSP, ICML, \ldots}\}$ and $Y = \{\}$.

**Want:** Embed $X$ into Euclidean space $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g. clustering.
Application: Embedding Data into Euclidean Space

Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$X = \{ \ldots \}$

$Y = \{ ISIT, \text{Allerton, ICASSP, ICML,} \ldots \}$

Want: Embed $X$ into Euclidean space $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g. clustering.

“Natural” Embedding: Represent each $x \in X$ using conditional distribution $P_{Y|X=x} \in \mathbb{R}^{\left| Y \right|}$. 

- Dimensionality Reduction: $\left| Y \right|$ is large!
- Reduce dimension of embedding.
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$$X = \{\text{person1}, \text{person2}, \text{person3}, \ldots\}$$

$$Y = \{\text{ISIT, Allerton, ICASSP, ICML}, \ldots\}$$

**Want**: Embed $X$ into Euclidean space $\mathbb{R}^k$ using knowledge of $P_{X,Y}$ for further processing, e.g. clustering.

**“Natural” Embedding**: Represent each $x \in X$ using conditional distribution $P_{Y|X=x} \in \mathbb{R}^{|Y|}$.

**Dimensionality Reduction**: $|Y|$ is large! Reduce dimension of embedding.
Application: Embedding Data into Euclidean Space

Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

\[ X = \left\{ \right. \]
\[ \begin{array}{c}
\text{\textbf{\textit{male}}}, \\
\text{\textbf{\textit{female}}}, \\
\text{\textbf{\textit{gender}}}, \\
\text{\textbf{\textit{...}}} \\
\end{array}
\]

\[ Y = \{ \text{ISIT}, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots \} \]

**Want:** Low-dimensional embedding of $X$ into Euclidean space $\mathbb{R}^k$. 

Modal Decomposition Embedding:

(when $\sigma_k + 2$ small)

$\zeta_k: X \rightarrow \mathbb{R}^k$, $\zeta_k(x) = [\sigma_2^2 f_2(x) \cdots \sigma_{k+1} f_{k+1}(x)]^T$

Diffusion Distance Preservation:

(similar to diffusion maps [CL06])

$D_{\text{diff}}(P_Y|X = x, P_Y|X = x') \approx \frac{1}{\min\{|X|, |Y|} - 1(x) - \frac{1}{\min\{|X|, |Y|} - 1(x') \approx \frac{1}{2} \|\zeta_k(x) - \zeta_k(x')\|_2^2$
Consider bivariate distribution \( P_{X,Y} \) on categorical variables \( X \) and \( Y \), e.g.

\[
X = \left\{ \begin{array}{c}
\text{, , , , , ,} \\
\end{array} \right\}
\]

\[
Y = \{ \text{ISIT, Allerton, ICASSP, ICML, \ldots} \}
\]

**Modal Decomposition Embedding:**

\[
P_{Y|X=x} = P_Y + \min\{|X|,|Y|\} \sum_{i=2}^{\min\{|X|,|Y|\}} \sigma_i f_i(x) (g_i \cdot P_Y)
\]
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$$X = \left\{\right.$$ 

$$\right.$$ 

$$\right.$$ 

$$Y = \{\text{ISIT, Allerton, ICASSP, ICML, \ldots}\}$$

**Modal Decomposition Embedding:** (when $\sigma_{k+2}$ small)

$$\zeta_k : \mathcal{X} \rightarrow \mathbb{R}^k, \quad \zeta_k(x) = [\sigma_2 f_2(x) \cdots \sigma_{k+1} f_{k+1}(x)]^T$$

![Diagram showing embedding process](image)
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$X = \{ \text{9x252} \}$,

$Y = \{ \text{ISIT, Allerton, ICASSP, ICML, \ldots} \}$

**Modal Decomposition Embedding:** (when $\sigma_{k+2}$ small)

$$
\zeta_k : \mathcal{X} \rightarrow \mathbb{R}^k, \quad \zeta_k(x) = [\sigma_2 f_2(x) \cdots \sigma_{k+1} f_{k+1}(x)]^T
$$

**Diffusion Distance Preservation:** (similar to diffusion maps [CL06])

$$
D_{\text{diff}}(P_{Y|X=x}, P_{Y|X=x'}) \triangleq \sum_{y \in \mathcal{Y}} \frac{(P_{Y|X}(y|x) - P_{Y|X}(y|x'))^2}{P_Y(y)}
$$
Consider bivariate distribution $P_{X,Y}$ on categorical variables $X$ and $Y$, e.g.

$$X = \{, , , , \ldots \}$$

$$Y = \{ISIT, \text{Allerton}, \text{ICASSP}, \text{ICML}, \ldots \}$$

**Modal Decomposition Embedding:** (when $\sigma_{k+2}$ small)

$$\zeta_k : \mathcal{X} \rightarrow \mathbb{R}^k, \quad \zeta_k(x) = [\sigma_2 f_2(x) \cdots \sigma_{k+1} f_{k+1}(x)]^T$$

**Diffusion Distance Preservation:** (similar to diffusion maps [CL06])

$$D_{\text{diff}}(P_{Y|X=x}, P_{Y|X=x'}) \triangleq \sum_{y \in Y} \frac{(P_{Y|X}(y|x) - P_{Y|X}(y|x'))^2}{P_Y(y)}$$

$$= \|\zeta_{\min\{|X|,|Y|}-1(x) - \zeta_{\min\{|X|,|Y|}-1(x')\|}_2^2$$

$$\approx \|\zeta_k(x) - \zeta_k(x')\|_2^2$$
Extended Alternating Conditional Expectations Algorithm

Require: joint pmf $P_{X,Y}$, number of dominant modes $k$

Remarks:
- Orthogonal iteration method [GvL96]
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_X,Y$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

Repeat:

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:

   $$
   \mathbb{E}[\hat{r}_k(X)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I.
   $$

3. Compute update $s_k: Y \rightarrow \mathbb{R}^k$:

   $$
   s_k(y) = \mathbb{E}[\hat{r}_k(X) | Y = y].
   $$

4. Center and whiten $s_k$ to obtain $\hat{s}_k$:

   $$
   \mathbb{E}[\hat{s}_k(Y)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{s}_k(Y)\hat{s}_k(Y)^T] = I.
   $$

5. Compute update $r_k$:

   $$
   r_k(x) = \mathbb{E}[\hat{s}_k(Y) | X = x].
   $$

Until $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ stops increasing.

**Remarks:**

- Orthogonal iteration method [GvL96]
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_{X,Y}$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

**Repeat:**

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:
   \[
   \mathbb{E}[\hat{r}_k(X)] = 0 \text{ and } \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I.
   \]

3. Compute update $s_k : \mathcal{Y} \rightarrow \mathbb{R}^k$: $s_k(y) = \mathbb{E}[\hat{r}_k(X)|Y = y]$.

**Remarks:**

- Orthogonal iteration method [GvL96]
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_{X,Y}$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

**Repeat:**

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:
   \[ \mathbb{E}[\hat{r}_k(X)] = 0 \text{ and } \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I. \]

3. Compute update $s_k : \mathcal{Y} \rightarrow \mathbb{R}^k$: $s_k(y) = \mathbb{E}[\hat{r}_k(X) | Y = y]$.

4. Center and whiten $s_k$ to obtain $\hat{s}_k$:
   \[ \mathbb{E}[\hat{s}_k(Y)] = 0 \text{ and } \mathbb{E}[\hat{s}_k(Y)\hat{s}_k(Y)^T] = I. \]

5. Compute update $r_k$: $r_k(x) = \mathbb{E}[\hat{s}_k(Y) | X = x]$.

**Remarks:**

- Orthogonal iteration: $C^* = \mathbb{E}[\cdot | X]$ is adjoint of $C = \mathbb{E}[\cdot | Y]$
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_{X,Y}$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

**Repeat:**

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:
   \[\mathbb{E}[\hat{r}_k(X)] = \mathbf{0} \text{ and } \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I.\]

3. Compute update $s_k : \mathcal{Y} \rightarrow \mathbb{R}^k$:
   \[s_k(y) = \mathbb{E}[\hat{r}_k(X) | Y = y].\]

4. Center and whiten $s_k$ to obtain $\hat{s}_k$:
   \[\mathbb{E}[\hat{s}_k(Y)] = \mathbf{0} \text{ and } \mathbb{E}[\hat{s}_k(Y)\hat{s}_k(Y)^T] = I.\]

5. Compute update $r_k$:
   \[r_k(x) = \mathbb{E}[\hat{s}_k(Y) | X = x].\]

**Until** $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ stops increasing.

**Remarks:**

- **Orthogonal iteration:** $C^* = \mathbb{E}[\cdot | X]$ is adjoint of $C = \mathbb{E}[\cdot | Y]$.
- **Termination:** $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ converges to Ky Fan $k$-norm $\sum_{i=2}^{k+1} \sigma_i$. 
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_{X,Y}$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

**Repeat:**

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:
   
   \[ \mathbb{E}[\hat{r}_k(X)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I. \]

3. Compute update $s_k : \mathcal{Y} \rightarrow \mathbb{R}^k$: 
   \[ s_k(y) = \mathbb{E}[\hat{r}_k(X)|Y = y]. \]

4. Center and whiten $s_k$ to obtain $\hat{s}_k$:
   
   \[ \mathbb{E}[\hat{s}_k(Y)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{s}_k(Y)\hat{s}_k(Y)^T] = I. \]

5. Compute update $r_k$: 
   \[ r_k(x) = \mathbb{E}[\hat{s}_k(Y)|X = x]. \]

**Until** $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ stops increasing.

**Remarks:**

- Orthogonal iteration: $\hat{r}_k, \hat{s}_k$ converge to $[f_2 \cdots f_{k+1}]^T, [g_2 \cdots g_{k+1}]^T$,
- Termination: $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ converges to Ky Fan $k$-norm $\sum_{i=2}^{k+1} \sigma_i$
Extended Alternating Conditional Expectations Algorithm

**Require:** joint pmf $P_{X,Y}$, number of dominant modes $k$

1. **Initialization:** Randomly choose $r_k : \mathcal{X} \rightarrow \mathbb{R}^k$.

**Repeat:**

2. Center and whiten $r_k$ to obtain $\hat{r}_k$:
   \[
   \mathbb{E}[\hat{r}_k(X)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{r}_k(X)\hat{r}_k(X)^T] = I.
   \]

3. Compute update $s_k : \mathcal{Y} \rightarrow \mathbb{R}^k$: $s_k(y) = \mathbb{E}[\hat{r}_k(X)|Y = y]$.

4. Center and whiten $s_k$ to obtain $\hat{s}_k$:
   \[
   \mathbb{E}[\hat{s}_k(Y)] = 0 \quad \text{and} \quad \mathbb{E}[\hat{s}_k(Y)\hat{s}_k(Y)^T] = I.
   \]

5. Compute update $r_k$: $r_k(x) = \mathbb{E}[\hat{s}_k(Y)|X = x]$.

**Until** $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ stops increasing.

**Remarks:**

- Orthogonal iteration: $\hat{r}_k, \hat{s}_k$ converge to $[f_2 \cdots f_{k+1}]^T, [g_2 \cdots g_{k+1}]^T$
- Termination: $\mathbb{E}[\hat{r}_k(X)^T\hat{s}_k(Y)]$ converges to Ky Fan $k$-norm $\sum_{i=2}^{k+1} \sigma_i$
- $k = 1$ case: alternating conditional expectations (ACE) algorithm for regression [BF85]
Sample Extended ACE Algorithm

- Suppose true $P_{X,Y}$ unknown.
Suppose true $P_{X,Y}$ unknown.

Observe i.i.d. training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{X,Y}$ with empirical joint pmf:

$$
\hat{P}_{X,Y}^n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y\}.
$$
Sample Extended ACE Algorithm

- Suppose true $P_{X,Y}$ unknown.
- Observe i.i.d. training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{X,Y}$ with empirical joint pmf:

$$\hat{P}_n^{X,Y}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y\}.$$ 

- Assume $P_X$ and $P_Y$ known (e.g. high-dimensional regime $\max\{|X|, |Y|\} \ll n \ll |X||Y|$, or additional “unlabeled” data).
Sample Extended ACE Algorithm

- Suppose true $P_{X,Y}$ unknown.
- Observe i.i.d. training samples $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{X,Y}$ with empirical joint pmf:

$$\hat{P}_X^n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y\}.$$ 

- Assume $P_X$ and $P_Y$ known (e.g. high-dimensional regime $\max\{|\mathcal{X}|, |\mathcal{Y}|\} \ll n \ll |\mathcal{X}| |\mathcal{Y}|$, or additional “unlabeled” data).

- **Sample Version:**
  Center and update steps use operator $\hat{C}_n : \mathcal{L}^2(\mathcal{X}, P_X) \rightarrow \mathcal{L}^2(\mathcal{Y}, P_Y)$ that maps $f \in \mathcal{L}^2(\mathcal{X}, P_X)$ to $\hat{C}_n(f) \in \mathcal{L}^2(\mathcal{Y}, P_Y)$:

$$\left(\hat{C}_n(f)\right)(y) \triangleq \frac{\hat{P}_Y^n(y)}{P_Y(y)} \mathbb{E}{\hat{P}_X^n}[f(X)|Y = y] - \mathbb{E}_{P_X}[f(X)].$$
Let $\hat{C}_n$ have singular values $\hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{\max\{|\mathcal{X}|,|\mathcal{Y}|\}+1} \geq 0$ with right singular vectors $\{\hat{f}_2, \ldots, \hat{f}_{|\mathcal{X}|+1}\} \subseteq L^2(\mathcal{X}, P_X)$. 
Sample Complexity Analysis

- Let \( \hat{C}_n \) have singular values \( \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{\max\{|X|,|Y|\}+1} \geq 0 \) with right singular vectors \( \{\hat{f}_2, \ldots, \hat{f}_{|X|+1}\} \subseteq L^2(\mathcal{X}, P_X) \).

- \( \hat{C}_n \) is “empirical version” of \( C \) with leading singular vector removed, i.e. \( \tilde{C} \triangleq C - \mathbb{E}_{P_X}[\cdot] \).
Let \( \hat{C}_n \) have singular values \( \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{\max\{|\mathcal{X}|,|\mathcal{Y}|\}+1} \geq 0 \) with right singular vectors \( \{\hat{f}_2, \ldots, \hat{f}_{|\mathcal{X}|+1}\} \subseteq L^2(\mathcal{X}, P_X) \).

\( \hat{C}_n \) is “empirical version” of \( C \) with leading singular vector removed, i.e. \( \tilde{C} \triangleq C - \mathbb{E}_{P_X} [.] \).

**Convergence of Ky Fan \( k \)-norm (termination condition):**

\[
\left\| \hat{C}_n \right\|_{(k)} = \sum_{i=2}^{k+1} \hat{\sigma}_i \xrightarrow{P} \left\| \tilde{C} \right\|_{(k)} = \sum_{i=2}^{k+1} \sigma_i
\]
Sample Complexity Analysis

- Let $\hat{C}_n$ have singular values $\hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{\max\{|X|,|Y|\}+1} \geq 0$ with right singular vectors $\{\hat{f}_2, \ldots, \hat{f}_{|X|+1}\} \subseteq L^2(X, P_X)$.
- $\hat{C}_n$ is “empirical version” of $C$ with leading singular vector removed, i.e. $\tilde{C} \triangleq C - \mathbb{E}_{P_X}[\cdot]$.
- **Convergence of Ky Fan $k$-norm (termination condition):**
  \[
  \|\hat{C}_n\|_{(k)} = \sum_{i=2}^{k+1} \hat{\sigma}_i \xrightarrow{P} \|\tilde{C}\|_{(k)} = \sum_{i=2}^{k+1} \sigma_i
  \]
- **Convergence of “rank $k$ approximation” of $\chi^2$-information:**
  \[
  \sum_{i=2}^{k+1} \mathbb{E}_{P_Y}[(\tilde{C}(\hat{f}_i))(Y)^2] \xrightarrow{P} \sum_{i=2}^{k+1} \sigma_i^2
  \]
Sample Complexity Analysis

Fix $\delta > 0$ such that $P_X, P_Y \geq \delta$. 

Fix $\delta > 0$ such that $P_X, P_Y \geq \delta$.

**Theorem (Consistency)**

- **Ky Fan $k$-Norm Estimation**: For every $0 \leq t \leq \frac{1}{\delta} \sqrt{\frac{k}{2}}$:
  \[
  \Pr \left( \left\| \mathbf{\hat{C}}_n^{(k)} - \mathbf{\tilde{C}}^{(k)} \right\| \geq t \right) \leq \exp \left( \frac{1}{4} - \frac{n\delta^2 t^2}{8k} \right)
  \]

- **Singular Vector Estimation**: For every $0 \leq t \leq 4k$:
  \[
  \Pr \left( \left\| \sum_{i=2}^{k+1} \mathbb{E}_{P_Y} [(\mathbf{\tilde{C}}(\mathbf{\hat{f}}_i))(Y)^2] - \sum_{i=2}^{k+1} \sigma_i^2 \right\| \geq t \right) \leq (|\mathcal{X}| + |\mathcal{Y}|) \exp \left( -\frac{n\delta t^2}{64k^2} \right)
  \]
Fix $\delta > 0$ such that $P_X, P_Y \geq \delta$.

**Theorem (Consistency)**

- **Ky Fan $k$-Norm Estimation:** For every $0 \leq t \leq \frac{1}{\delta} \sqrt{\frac{k}{2}}$:
  \[
P\left( \left| \| \hat{\mathcal{C}}_n \|_k - \| \mathcal{C} \|_k \right| \geq t \right) \leq \exp \left( \frac{1}{4} - \frac{n\delta^2 t^2}{8k} \right)
  \]

- **Singular Vector Estimation:** For every $0 \leq t \leq 4k$:
  \[
P\left( \left| \sum_{i=2}^{k+1} \mathbb{E}_{P_Y} \left[ (\mathcal{C}(\hat{f}_i))(Y)^2 \right] - \sum_{i=2}^{k+1} \sigma_i^2 \right| \geq t \right) \leq (|X| + |Y|) \exp \left( -\frac{n\delta t^2}{64k^2} \right)
  \]

**Remark:** $n$ grows with $k$
1 Introduction

2 Contraction Coefficients and Strong Data Processing Inequalities

3 Extension using Comparison of Channels

4 Modal Decomposition of Mutual $\chi^2$-Information

5 Information Contraction in Networks: Broadcasting on DAGs
   ● Problem and Motivation
   ● Results on Random DAGs
   ● Results on 2D Regular Grids

6 Conclusion
Fix infinite directed acyclic graph (DAG) with single source node.
Broadcasting on Bounded Indegree DAGs

- Fix infinite DAG with single source node.
- $X_{k,j} \in \{0, 1\}$ – node random variable at $j$th position in level $k$

![DAG Diagram]

- $X_{0,0}$
- $X_{1,0}, X_{1,1}, X_{1,2}$
- $X_{2,0}, X_{2,1}, X_{2,2}, X_{2,3}$
- $\cdots$
- $X_{k,0}, X_{k,1}, X_{k,L_k-2}, X_{k,L_k-1}$
Broadcasting on Bounded Indegree DAGs

- Fix infinite DAG with single source node.
- $X_{k,j} \in \{0, 1\}$ – node random variable at $j$th position in level $k$
- $L_k$ – number of nodes at level $k$

- Every edge is independent BSC with crossover probability $\delta \in (0, \frac{1}{2})$.
- Nodes combine inputs with $d$-ary Boolean functions.

This defines joint distribution of $\{X_{k,j}\}$.

---

**Diagram:**

- **Level 0:** $X_{0,0}$, $L_0 = 1$
- **Level 1:** $X_{1,0}, X_{1,1}, X_{1,2}$, $L_1 = 3$
- **Level 2:** $X_{2,0}, X_{2,1}, X_{2,2}, X_{2,3}$, $L_2 = 4$
- **Level $k$:** $X_{k,0}, X_{k,1}, \ldots, X_{k,L_k-2}, X_{k,L_k-1}$, $L_k$ vertices

- $X_{0,0}$, $X_{1,0}, X_{1,1}, X_{1,2}$, $X_{2,0}, X_{2,1}, X_{2,2}, X_{2,3}$, $\ldots$, $X_{k,0}, X_{k,1}, \ldots, X_{k,L_k-2}, X_{k,L_k-1}$
Broadcasting on Bounded Indegree DAGs

- Fix infinite DAG with single source node.
- $X_{k,j} \in \{0, 1\}$ – node random variable at $j$th position in level $k$
- $L_k$ – number of nodes at level $k$
- $d$ – indegree of each node

![Diagram of DAG levels with nodes and levels labeled]

- $X_{0,0}$
- $L_0 = 1$
- $X_{1,0}, X_{1,1}, X_{1,2}$
- $L_1 = 3$
- $X_{2,0}, X_{2,1}, X_{2,2}, X_{2,3}$
- $d = 2$
- $L_2 = 4$
- ... 
- ... 
- $X_{k,0}, X_{k,1}, \ldots, X_{k,L_k-2}, X_{k,L_k-1}$
- $L_k$ vertices

Anuran Makur (MIT) | Information Contraction & Decomposition | 15 May 2019 | 47 / 64
Broadcasting on Bounded Indegree DAGs

- Fix infinite DAG with single source node.
- $X_{k,j} \in \{0, 1\}$ – node random variable at $j$th position in level $k$
- $L_k$ – number of nodes at level $k$
- $d$ – indegree of each node

- $X_{0,0} \sim \text{Bernoulli}(\frac{1}{2})$
- Every edge is independent BSC with crossover probability $\delta \in (0, \frac{1}{2})$. 

![Diagram of DAG levels with labeled nodes and indegree $d = 2$]
Broadcasting on Bounded Indegree DAGs

- Fix infinite DAG with single source node.
- $X_{k,j} \in \{0, 1\}$ – node random variable at $j$th position in level $k$
- $L_k$ – number of nodes at level $k$
- $d$ – indegree of each node

$X_{0,0} \sim \text{Bernoulli}\left(\frac{1}{2}\right)$

- Every edge is independent BSC with crossover probability $\delta \in (0, \frac{1}{2})$.
- Nodes combine inputs with $d$-ary Boolean functions.
- This defines joint distribution of $\{X_{k,j}\}$.
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.

- **Can we decode $X_0$ from $X_k$ as $k \to \infty$?**

![Diagram of level 0, 1, 2, and $k$ vertices](image)

- Level 0: $X_{0,0}$, $L_0 = 1$
- Level 1: $X_{1,0}, X_{1,1}, X_{1,2}$, $L_1 = 3$
- Level 2: $X_{2,0}, X_{2,1}, X_{2,2}, X_{2,3}$, $L_2 = 4$
- Level $k$: $X_{k,0}, X_{k,1}, \ldots, X_{k,L_k-2}, X_{k,L_k-1}$, $L_k$ vertices
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.
- Can we decode $X_0$ from $X_k$ as $k \to \infty$?

Binary Hypothesis Testing: Let $\hat{X}_{ML}^k(X_k) \in \{0, 1\}$ be maximum likelihood (ML) decoder with probability of error:

$$P_{ML}^{(k)} \triangleq \mathbb{P}(\hat{X}_{ML}^k(X_k) \neq X_{0,0})$$.
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.
- **Can we decode $X_0$ from $X_k$ as $k \to \infty$?**
- **Binary Hypothesis Testing:** Let $\hat{X}^{k}_{ML}(X_k) \in \{0, 1\}$ be maximum likelihood (ML) decoder with probability of error:

$$P^{(k)}_{ML} \triangleq \mathbb{P}(\hat{X}^{k}_{ML}(X_k) \neq X_{0,0}) = \frac{1}{2} \left(1 - \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV}\right).$$
Let \( X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1}) \).

**Can we decode \( X_0 \) from \( X_k \) as \( k \to \infty \)?**

**Binary Hypothesis Testing:** Let \( \hat{X}_k^{ML}(X_k) \in \{0, 1\} \) be maximum likelihood (ML) decoder with probability of error:

\[
P^{(k)}_{ML} \triangleq \mathbb{P}\left( \hat{X}_k^{ML}(X_k) \neq X_{0,0} \right) = \frac{1}{2} \left( 1 - \| P_{X_k|X_0=1} - P_{X_k|X_0=0} \|_{TV} \right).
\]

By DPI, **TV distance contracts** as \( k \) increases.

For which \( \delta, d, \{L_k\} \), and Boolean processing functions is reconstruction possible?
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.
- Can we decode $X_0$ from $X_k$ as $k \to \infty$?

**Binary Hypothesis Testing:** Let $\hat{X}_{ML}^k(X_k) \in \{0, 1\}$ be maximum likelihood (ML) decoder with probability of error:

$$P_{ML}^{(k)} \triangleq P\left(\hat{X}_{ML}^k(X_k) \neq X_{0,0}\right) = \frac{1}{2} \left(1 - \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV}\right).$$

- By DPI, TV distance contracts as $k$ increases.

**Broadcasting/Reconstruction possible if:**

$$\lim_{k \to \infty} P_{ML}^{(k)} < \frac{1}{2} \iff \lim_{k \to \infty} \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV} > 0$$

and **Broadcasting/Reconstruction impossible if:**

$$\lim_{k \to \infty} P_{ML}^{(k)} = \frac{1}{2} \iff \lim_{k \to \infty} \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV} = 0.$$
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.
- **Can we decode** $X_0$ from $X_k$ as $k \to \infty$?
- **Binary Hypothesis Testing:** Let $\hat{X}_{ML}^k(X_k) \in \{0, 1\}$ be maximum likelihood (ML) decoder with probability of error:

  $$P_{ML}^{(k)} \triangleq \mathbb{P}(\hat{X}_{ML}^k(X_k) \neq X_{0,0}) = \frac{1}{2} \left(1 - \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV}\right).$$

- By DPI, TV distance contracts as $k$ increases.
- Broadcasting/Reconstruction possible iff:

  $$\lim_{k \to \infty} P_{ML}^{(k)} < \frac{1}{2} \iff \lim_{k \to \infty} \|P_{X_k|X_0=1} - P_{X_k|X_0=0}\|_{TV} > 0.$$

**For which $\delta$, $d$, $\{L_k\}$, and Boolean processing functions is reconstruction possible?**
Broadcasting Question

- Let $X_k \triangleq (X_{k,0}, \ldots, X_{k,L_k-1})$.
- **Can we decode $X_0$ from $X_k$ as $k \to \infty$?**
- **Binary Hypothesis Testing:** Let $\hat{X}_{\text{ML}}^k(X_k) \in \{0, 1\}$ be maximum likelihood (ML) decoder with probability of error:
  \[
  P_{\text{ML}}^{(k)} \triangleq \mathbb{P}\left(\hat{X}_{\text{ML}}^k(X_k) \neq X_{0,0}\right) = \frac{1}{2} \left(1 - \left\|P_{X_k|X_0=1} - P_{X_k|X_0=0}\right\|_{\text{TV}}\right).
  \]
- By DPI, TV distance contracts as $k$ increases.
- Broadcasting/Reconstruction possible iff:
  \[
  \lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2} \iff \lim_{k \to \infty} \left\|P_{X_k|X_0=1} - P_{X_k|X_0=0}\right\|_{\text{TV}} > 0.
  \]
- Broadcasting $\Leftrightarrow$ TV distance contraction.

**For which $\delta$, $d$, $\{L_k\}$, and Boolean processing functions is reconstruction possible?**
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])

If $\lim_{k \to \infty} P_k < \frac{1}{2}$, then reconstruction possible:

If $\lim_{k \to \infty} P_k = \frac{1}{2}$, then reconstruction impossible:

Idea: Contract $\eta_{KL}(\text{BSC}(\delta))$ along $\text{br}(T)$ paths [ES99].

Observations:

$L_k$ sub-exponential $\Rightarrow \text{br}(T) = 1$ and reconstruction impossible

$d > 1 \Rightarrow$ information fusion at nodes

Can we broadcast with sub-exponential $L_k$ when $d > 1$?

Yes, we can broadcast with $L_k = \Theta(\log(k))$!
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00]):

If $d > 1$, then reconstruction possible: $\lim_{k \to \infty} P(k)_{\text{ML}} < \frac{1}{2}$.

If $d < 1$, then reconstruction impossible: $\lim_{k \to \infty} P(k)_{\text{ML}} = \frac{1}{2}$.

Idea:

Contract $\eta_{KL}(\text{BSC}(\delta)) = (1 - 2^d)^{-2}$ along $\text{br}(T)^k$ paths [ES99].

Observations:

$L_k$ sub-exponential $\Rightarrow$ $\text{br}(T) = 1$ and reconstruction impossible.

$d > 1$ $\Rightarrow$ information fusion at nodes.

Can we broadcast with sub-exponential $L_k$ when $d > 1$?

Yes, we can broadcast with $L_k = \Theta(\log(k))$!
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

**Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])**

- If $\delta < \frac{1}{2} - \frac{1}{2\sqrt{\text{br}(T)}}$, then reconstruction possible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}$.
- If $\delta > \frac{1}{2} - \frac{1}{2\sqrt{\text{br}(T)}}$, then reconstruction impossible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}$.

---

**Idea:** Contract $\eta$ $\text{KL}(\text{BSC}(\delta))$ $k = (1 - 2\delta)^2$ along $\text{br}(T)$ $k$ paths [ES99].

---

**Observations:** $L_k$ sub-exponential $\Rightarrow$ $\text{br}(T) = 1$ and reconstruction impossible $d > 1$ $\Rightarrow$ information fusion at nodes

---

Can we broadcast with sub-exponential $L_k$ when $d > 1$?

Yes, we can broadcast with $L_k = \Theta(\log(k))$!
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

**Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])**

- If $(1 - 2\delta)^2 \text{br}(T) > 1$, then reconstruction possible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}$.
- If $(1 - 2\delta)^2 \text{br}(T) < 1$, then reconstruction impossible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}$.

**Idea:** Contract $\eta_{KL}(\text{BSC}(\delta))^k = (1 - 2\delta)^2k$ along $\text{br}(T)^k$ paths [ES99].

- $L_0 = 1$
- $L_1 = 2$
- $\text{br}(T) = 2$
- $L_2 = 2^2$
- $L_k = \text{br}(T)^k$
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

**Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])**

- If $(1 - 2\delta)^2 \cdot \text{br}(T) > 1$, then reconstruction possible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}$.
- If $(1 - 2\delta)^2 \cdot \text{br}(T) < 1$, then reconstruction impossible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}$.

**Idea:** Contract $\eta_{\text{KL}}(\text{BSC}(\delta))^k = (1 - 2\delta)^{2k}$ along $\text{br}(T)^k$ paths [ES99].

**Observations:**

- $L_k$ sub-exponential $\Rightarrow$ $\text{br}(T) = 1$ and reconstruction impossible
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

**Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])**

- If $(1 - 2\delta)^2 \text{br}(T) > 1$, then reconstruction possible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}$.
- If $(1 - 2\delta)^2 \text{br}(T) < 1$, then reconstruction impossible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}$.

Idea: Contract $\eta_{\text{KL}}(\text{BSC}(\delta))^k = (1 - 2\delta)^{2k}$ along $\text{br}(T)^k$ paths [ES99].

Observations:
- $L_k$ sub-exponential $\Rightarrow$ $\text{br}(T) = 1$ and reconstruction impossible
- $d > 1 \Rightarrow$ information fusion at nodes

Can we broadcast with sub-exponential $L_k$ when $d > 1$?
Motivation: Broadcasting on Trees

Fix tree $T$ with $d = 1$, identity processing, and branching number $\text{br}(T)$.

**Theorem (Phase Transition for Trees [KS66, BRZ95, EKPS00])**

- If $(1 - 2\delta)^2 \text{br}(T) > 1$, then reconstruction possible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}$.
- If $(1 - 2\delta)^2 \text{br}(T) < 1$, then reconstruction impossible: $\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}$.

**Idea:** Contract $\eta_{\text{KL}}(\text{BSC}(\delta))^k = (1 - 2\delta)^{2k}$ along $\text{br}(T)^k$ paths [ES99].

**Observations:**
- $L_k$ sub-exponential $\Rightarrow$ $\text{br}(T) = 1$ and reconstruction impossible
- $d > 1 \Rightarrow$ information fusion at nodes

**Can we broadcast with sub-exponential $L_k$ when $d > 1$?**

Yes, we can broadcast with $L_k = \Theta(\log(k))$!
1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs
   - Problem and Motivation
   - Results on Random DAGs
   - Results on 2D Regular Grids

6. Conclusion
Random DAG Model

- Fix $\{L_k\}$ and $d > 1$. 
Random DAG Model

- Fix \( \{L_k\} \) and \( d > 1 \).
- For each node \( X_{k,j} \), randomly and independently select \( d \) parents from level \( k - 1 \) (with repetition).
- This defines random DAG \( G \).
Random DAG Model

- Fix \( \{L_k\} \) and \( d > 1 \).
- For each node \( X_{k,j} \), randomly and independently select \( d \) parents from level \( k - 1 \) (with repetition).
- This defines random DAG \( G \).
- Let \( P_{\text{ML}}^{(k)}(G) \) be ML decoding probability of error for DAG \( G \), and define \( \sigma_k \triangleq \frac{1}{L_k} \sum_j X_{k,j} \) which is sufficient statistic of \( X_k \) for \( \sigma_0 = X_{0,0} \).
Random DAG with Majority Processing

**Theorem (Phase Transition for \( d \geq 3 \))**

Consider random DAG model with \( d \geq 3 \) and majority processing (with ties broken randomly). Let \( \delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\lceil d/2 \rceil (\lceil d/2 \rceil)} \).

Suppose \( \delta \in (0, \delta_{maj}) \). Then, there exists \( C(\delta, d) > 0 \) such that if \( L_k \geq C(\delta, d) \log(k) \), then reconstruction possible:

\[
\limsup_{k \to \infty} P(\hat{S}_k \neq X_0, 0) < \frac{1}{2}
\]

where \( \hat{S}_k \triangleq 1\{\sigma_k \geq \frac{1}{2} \} \) is majority decoder.

Suppose \( \delta \in (\delta_{maj}, \frac{1}{2}) \). Then, there exists \( D(\delta, d) > 1 \) such that if \( L_k = o(D(\delta, d) k) \), then reconstruction impossible:

\[
\lim_{k \to \infty} P(k)_{ML}(G) = \frac{1}{2} G - a.s.
\]
Consider random DAG model with \( d \geq 3 \) and majority processing (with ties broken randomly). Let \( \delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\lfloor d/2 \rfloor(d/2)} \).

- Suppose \( \delta \in (0, \delta_{maj}) \). Then, there exists \( C(\delta, d) > 0 \) such that if \( L_k \geq C(\delta, d) \log(k) \), then reconstruction possible:

\[
\limsup_{k \to \infty} \Pr(\hat{S}_k \neq X_{0,0}) < \frac{1}{2}
\]

where \( \hat{S}_k \triangleq \mathbb{I}\{\sigma_k \geq \frac{1}{2}\} \) is majority decoder.
Random DAG with Majority Processing

Theorem (Phase Transition for \( d \geq 3 \))

Consider random DAG model with \( d \geq 3 \) and majority processing (with ties broken randomly). Let \( \delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\lceil \frac{d}{2} \rceil \lceil \frac{d}{2} \rceil} \).

- Suppose \( \delta \in (0, \delta_{maj}) \). Then, there exists \( C(\delta, d) > 0 \) such that if \( L_k \geq C(\delta, d) \log(k) \), then reconstruction possible:

\[
\lim_{k \to \infty} \mathbb{E}\left[ P_{ML}^{(k)}(G) \right] \leq \limsup_{k \to \infty} \mathbb{P}(\hat{S}_k \neq X_{0,0}) < \frac{1}{2}
\]

where \( \hat{S}_k \triangleq \mathbb{I}\{\sigma_k \geq \frac{1}{2}\} \) is majority decoder.
Theorem (Phase Transition for $d \geq 3$)

Consider random DAG model with $d \geq 3$ and majority processing (with ties broken randomly). Let

$$\delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\lceil d/2 \rceil \lceil d/2 \rceil}.$$ 

- Suppose $\delta \in (0, \delta_{maj})$. Then, there exists $C(\delta, d) > 0$ such that if $L_k \geq C(\delta, d) \log(k)$, then reconstruction possible:

$$\lim_{k \to \infty} \mathbb{E} \left[ P^{(k)}_{\text{ML}}(G) \right] \leq \limsup_{k \to \infty} \mathbb{P} \left( \hat{S}_k \neq X_{0,0} \right) < \frac{1}{2}$$

where $\hat{S}_k \triangleq \mathbb{1} \left\{ \sigma_k \geq \frac{1}{2} \right\}$ is majority decoder.

- Suppose $\delta \in (\delta_{maj}, \frac{1}{2})$. Then, there exists $D(\delta, d) > 1$ such that if $L_k = o(D(\delta, d)^k)$, then reconstruction impossible:

$$\lim_{k \to \infty} P^{(k)}_{\text{ML}}(G) = \frac{1}{2} \quad \text{G-a.s.}$$
Random DAG with Majority Processing

**Theorem (Phase Transition for $d \geq 3$)**

Consider random DAG model with $d \geq 3$ and majority processing (with ties broken randomly). Let $\delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\left\lceil d/2 \right\rceil(\left\lceil d/2 \right\rceil)}$.

- Suppose $\delta \in (0, \delta_{maj})$. Then, there exists $C(\delta, d) > 0$ such that if $L_k \geq C(\delta, d) \log(k)$, then $\lim_{k \to \infty} \mathbb{E}\left[P_{ML}^{(k)}(G)\right] < \frac{1}{2}$.

- Suppose $\delta \in (\delta_{maj}, \frac{1}{2})$. Then, there exists $D(\delta, d) > 1$ such that if $L_k = o\left(D(\delta, d)^k\right)$, then $\lim_{k \to \infty} P_{ML}^{(k)}(G) = \frac{1}{2}$ G-a.s.

**Remarks:**

- $\delta_{maj} = \frac{1}{6}$ for $d = 3$ appears in reliable computation [vNe56, HW91].
- $\delta_{maj}$ for odd $d \geq 3$ also relevant in reliable computation [ES03].
- $\delta_{maj}$ for $d \geq 3$ relevant in recursive reconstruction on trees [Mos98].
Random DAG with Majority Processing

Theorem (Phase Transition for $d \geq 3$)

Consider random DAG model with $d \geq 3$ and majority processing (with ties broken randomly). Let $\delta_{maj} \triangleq \frac{1}{2} - \frac{2^{d-2}}{\lceil d/2 \rceil \lceil d/2 \rceil}$. 

- Suppose $\delta \in (0, \delta_{maj})$. Then, there exists $C(\delta, d) > 0$ such that if $L_k \geq C(\delta, d) \log(k)$, then $\lim_{k \to \infty} \mathbb{E} \left[ P_{ML}^{(k)}(G) \right] < \frac{1}{2}$.

- Suppose $\delta \in (\delta_{maj}, \frac{1}{2})$. Then, there exists $D(\delta, d) > 1$ such that if $L_k = o(D(\delta, d)^k)$, then $\lim_{k \to \infty} P_{ML}^{(k)}(G) = \frac{1}{2}$ G-a.s.

Questions:

- Broadcasting possible with sub-logarithmic $L_k$?
- Broadcasting possible when $\delta > \delta_{maj}$ with other processing functions?
- What about $d = 2$?
Optimality of Logarithmic Layer Size Growth

Broadcasting possible with sub-logarithmic $L_k$?

Prop (Layer Size Impossibility Result)
For any deterministic DAG, if:

$$L_k \leq \frac{\log(k)}{d \log\left(\frac{1}{2\delta}\right)},$$

then reconstruction impossible for all processing functions:

$$\lim_{k \to \infty} P_{ML}^{(k)} = \frac{1}{2}.$$
Optimality of Logarithmic Layer Size Growth

Broadcasting possible with sub-logarithmic $L_k$?

**Prop (Layer Size Impossibility Result)**

For any deterministic DAG, if:

$$L_k \leq \frac{\log(k)}{d \log\left(\frac{1}{2\delta}\right)},$$

then reconstruction impossible for all processing functions:

$$\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}.$$

No, broadcasting impossible with sub-logarithmic $L_k$!
Broadcasting possible when $\delta > \delta_{maj}$ with other processing functions?

Prop (Single Vertex Reconstruction)

Consider random DAG model with $d \geq 3$.

- If $\delta \in (0, \delta_{maj})$, $L_k \geq C(\delta, d) \log(k)$, and processing functions are majority, then single vertex reconstruction possible:

  $$\limsup_{k \to \infty} \mathbb{P}(X_{k,0} \neq X_{0,0}) < \frac{1}{2}.$$
Partial Converse Results

Broadcasting possible when $\delta > \delta_{\text{maj}}$ with other processing functions?

Prop (Single Vertex Reconstruction)

Consider random DAG model with $d \geq 3$.

- If $\delta \in (0, \delta_{\text{maj}})$, $L_k \geq C(\delta, d) \log(k)$, and processing functions are majority, then single vertex reconstruction possible:
  \[
  \limsup_{k \to \infty} \mathbb{P}(X_{k,0} \neq X_{0,0}) < \frac{1}{2}.
  \]

- If $\delta \in [\delta_{\text{maj}}, \frac{1}{2})$, $d$ is odd, $\lim_{k \to \infty} L_k = \infty$, and $\inf_{n \geq k} L_n = O(d^{2k})$, then single vertex reconstruction impossible for all processing functions:
  \[
  \lim_{k \to \infty} \mathbb{E} \left[ \left\| P_{X_{k,0}|G,X_{0,0}=1} - P_{X_{k,0}|G,X_{0,0}=0} \right\|_{\text{TV}} \right] = 0.
  \]

Remark: Converse uses reliable computation results [HW91, ES03].
Partial Converse Results

Broadcasting possible when $\delta > \delta_{\text{maj}}$ with other processing functions?

Remark: Converse uses reliable computation results \[HW91, \text{ES03}\].

Prop (Information Percolation [ES99])

For any deterministic DAG, if:

$$\delta > \frac{1}{2} - \frac{1}{2\sqrt{d}}$$

and

$$L_k = o\left(\frac{1}{((1 - 2\delta)^2d)^k}\right)$$

then reconstruction impossible for all processing functions:

$$\lim_{k \to \infty} P_{\text{ML}}^{(k)} = \frac{1}{2}.$$
Partial Converse Results

Broadcasting possible when $\delta > \delta_{\text{maj}}$
with other processing functions?

Remark: Converse uses reliable computation results \[WH91, ES03]\.

Prop (Information Percolation [ES99])

For any deterministic DAG, if:

$$\delta > \frac{1}{2} - \frac{1}{2\sqrt{d}} > \delta_{\text{maj}} \quad \text{and} \quad L_k = o\left(\frac{1}{((1 - 2\delta)^2d)^k}\right)$$

then reconstruction impossible for all processing functions:

$$\lim_{k \to \infty} P^{(k)}_{\text{ML}} = \frac{1}{2}.$$
What about \( d = 2 \)?

**Theorem (Phase Transition for \( d = 2 \))**

Consider random DAG model with \( d = 2 \) and **NAND processing** functions. Let \( \delta_{\text{nand}} \triangleq \frac{3 - \sqrt{7}}{4} \).

- Suppose \( \delta \in (0, \delta_{\text{nand}}) \). Then, there exist \( C(\delta) > 0 \) and \( t(\delta) \in (0, 1) \) such that if \( L_k \geq C(\delta) \log(k) \), then reconstruction possible:
  
  \[
  \lim_{k \to \infty} \mathbb{E} \left[ P_{\text{ML}}^{(k)}(G) \right] \leq \limsup_{k \to \infty} \mathbb{P} \left( \hat{T}_{2k} \neq X_{0,0} \right) < \frac{1}{2}
  \]
  
  where \( \hat{T}_k \triangleq \mathbb{1} \{ \sigma_k \geq t(\delta) \} \) is thresholding decoder.

- Suppose \( \delta \in (\delta_{\text{nand}}, \frac{1}{2}) \). Then, there exist \( D(\delta), E(\delta) > 1 \) such that if \( L_k = o\left( D(\delta)^k \right) \) and \( \liminf_{k \to \infty} L_k > E(\delta) \), then reconstruction impossible:
  
  \[
  \lim_{k \to \infty} P_{\text{ML}}^{(k)}(G) = \frac{1}{2} \quad \text{G-a.s.}
  \]
Random DAG with NAND Processing

What about $d = 2$?

**Theorem (Phase Transition for $d = 2$)**

Consider random DAG model with $d = 2$ and NAND processing functions. Let $\delta_{\text{nand}} \triangleq \frac{3 - \sqrt{7}}{4}$.

- Suppose $\delta \in (0, \delta_{\text{nand}})$. Then, there exist $C(\delta) > 0$ and $t(\delta) \in (0, 1)$ such that if $L_k \geq C(\delta) \log(k)$, then reconstruction possible:

$$\lim_{k \to \infty} \mathbb{E} \left[ P^{(k)}_{\text{ML}}(G) \right] \leq \limsup_{k \to \infty} \mathbb{P} \left( \hat{T}_{2k} \neq X_{0,0} \right) < \frac{1}{2}$$

where $\hat{T}_k \triangleq \mathbb{1}\{\sigma_k \geq t(\delta)\}$ is thresholding decoder.

- Suppose $\delta \in (\delta_{\text{nand}}, \frac{1}{2})$. Then, there exist $D(\delta), E(\delta) > 1$ such that if $L_k = o\left(D(\delta)^k\right)$ and $\liminf_{k \to \infty} L_k > E(\delta)$, then reconstruction impossible:

$$\lim_{k \to \infty} P^{(k)}_{\text{ML}}(G) = \frac{1}{2} \quad \text{G-a.s.}$$

**Remark:** $\delta_{\text{nand}}$ appears in reliable computation [EP98, Ung07].
Existence of DAGs where Broadcasting is Possible

Probabilistic Method:

Random DAG broadcasting $\Rightarrow$ DAG where reconstruction possible exists.
**Existence of DAGs where Broadcasting is Possible**

**Probabilistic Method:**

Random DAG broadcasting $\Rightarrow$ DAG where reconstruction possible exists. For example:

**Corollary (Existence of Deterministic Broadcasting DAGs)**

For every $d \geq 3$, $\delta \in (0, \delta_{\text{maj}})$, and $L_k \geq C(\delta, d) \log(k)$, there exists DAG with majority processing functions such that reconstruction possible:

$$\lim_{k \to \infty} P_{\text{ML}}^{(k)} < \frac{1}{2}.$$
Outline

1. Introduction
2. Contraction Coefficients and Strong Data Processing Inequalities
3. Extension using Comparison of Channels
4. Modal Decomposition of Mutual $\chi^2$-Information
5. Information Contraction in Networks: Broadcasting on DAGs
   - Problem and Motivation
   - Results on Random DAGs
   - Results on 2D Regular Grids
6. Conclusion
DAG is 2D regular grid with $L_k = k + 1$. 

Motivation:

"Positive rates conjecture" on ergodicity of simple 1D probabilistic cellular automata.
2D Regular Grid Model

- DAG is 2D regular grid with $L_k = k + 1$.
- Side nodes use identity processing.
- Other nodes use common Boolean processing function.

Conjecture: For all $\delta \in (0, 1/2)$ and common processing functions, reconstruction impossible on 2D regular grid model.

Motivation: "Positive rates conjecture" on ergodicity of simple 1D probabilistic cellular automata.

Anuran Makur (MIT)
2D Regular Grid Model

- DAG is 2D regular grid with $L_k = k + 1$.
- Side nodes use identity processing.
- Other nodes use common Boolean processing function.

Conjecture: For all $\delta \in (0, \frac{1}{2})$ and common processing functions, reconstruction impossible on 2D regular grid model.

Motivation: “Positive rates conjecture” on ergodicity of simple 1D probabilistic cellular automata.
Impossibility of Broadcasting

Theorem (2D Regular AND Grid)

For all $\delta \in (0, \frac{1}{2})$, reconstruction impossible on 2D regular grid model with AND processing:

$$\lim_{k \to \infty} P_{ML}^{(k)} = \frac{1}{2}.$$

Theorem (2D Regular XOR Grid)

For all $\delta \in (0, \frac{1}{2})$, reconstruction impossible on 2D regular grid model with XOR processing:

$$\lim_{k \to \infty} P_{ML}^{(k)} = \frac{1}{2}.$$
1. Introduction

2. Contraction Coefficients and Strong Data Processing Inequalities

3. Extension using Comparison of Channels

4. Modal Decomposition of Mutual $\chi^2$-Information

5. Information Contraction in Networks: Broadcasting on DAGs

6. Conclusion
Main Contributions:

- Properties of contraction coefficients
Main Contributions:

- Properties of contraction coefficients
- Characterization of $\ln$ via operator convexity
Main Contributions:

- Properties of contraction coefficients
- Characterization of $\geq_{\ln}$ via operator convexity
- Extending SDPIs: Conditions for $\geq_{\ln}$ domination by symmetric channels
- $\geq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequalities
Main Contributions:

- Properties of contraction coefficients
- Characterization of $\geq_{\ln}$ via operator convexity
- Extending SDPIs: Conditions for $\geq_{\ln}$ domination by symmetric channels
- $\geq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequalities
- Maximal correlation functions as embeddings of categorical data
- Structure of conditional expectation operators
Main Contributions:
- Properties of contraction coefficients
- Characterization of $\geq_{\ln}$ via operator convexity
- Extending SDPIs: Conditions for $\geq_{\ln}$ domination by symmetric channels
- $\geq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequalities
- Maximal correlation functions as embeddings of categorical data
- Structure of conditional expectation operators
- Extended ACE algorithm and sample complexity analysis
Main Contributions:

- Properties of contraction coefficients
- Characterization of $\preceq_{\ln}$ via operator convexity
- Extending SDPIs: Conditions for $\preceq_{\ln}$ domination by symmetric channels
- $\preceq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequalities
- Maximal correlation functions as embeddings of categorical data
- Structure of conditional expectation operators
- Extended ACE algorithm and sample complexity analysis
- **Broadcasting in random DAGs with** $d \geq 3$ **and majority processing**
- **Broadcasting in random DAGs with** $d = 2$ **and NAND processing**
Conclusion

Main Contributions:

- Properties of contraction coefficients
- Characterization of $\succeq_{\ln}$ via operator convexity
- Extending SDPIs: Conditions for $\succeq_{\ln}$ domination by symmetric channels
- $\succeq_{\ln}$ domination $\Rightarrow$ log-Sobolev inequalities
- Maximal correlation functions as embeddings of categorical data
- Structure of conditional expectation operators
- Extended ACE algorithm and sample complexity analysis
- Broadcasting in random DAGs with $d \geq 3$ and majority processing
- Broadcasting in random DAGs with $d = 2$ and NAND processing
- Broadcasting impossible in 2D regular grids with AND/XOR processing
Acknowledgments

- **Family:** Anamitra, Anindita, and Anyatama Makur
- **Doctoral Advisers:** Yury Polyanskiy and Lizhong Zheng
- **Research Guidance:** Elchanan Mossel and Gregory Wornell
- **Other Professors:** Venkat Anantharam, Afonso Bandeira, Guy Bresler, Alan Edelman, Muriel Médard, Alan Oppenheim, and Devavrat Shah
- **Admin:** Rachel Cohen, Molly Kruko, and Michael Lewy
Thank You!