Compression for Queries

Thomas Courtade

CSoI Workshop on Big Data

Joint work with: Amir Ingber, Tsachy Weissman
Also thanks to: Golan Yona, Sergio Verdú

March 19, 2013

Center for
Science of Information
NSF Science and Technology Center

The fundamental problem of communication is that of **reproducing at one point** *either exactly or approximately* **a message selected at another point**.

Claude E. Shannon, 1948

Transmission of Information

In modern data processing, objective is often not *reproduction* of a message

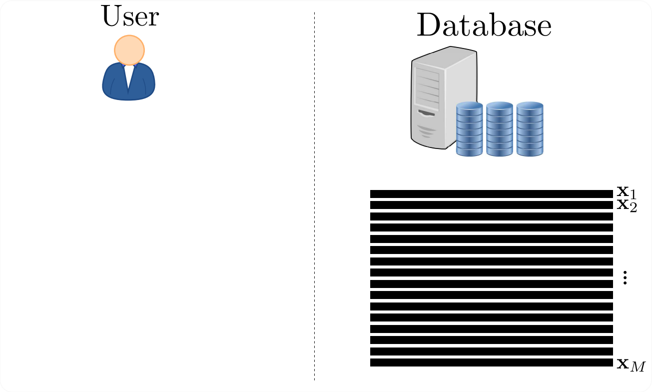
Transmission of Information

In modern data processing, objective is often not *reproduction* of a message

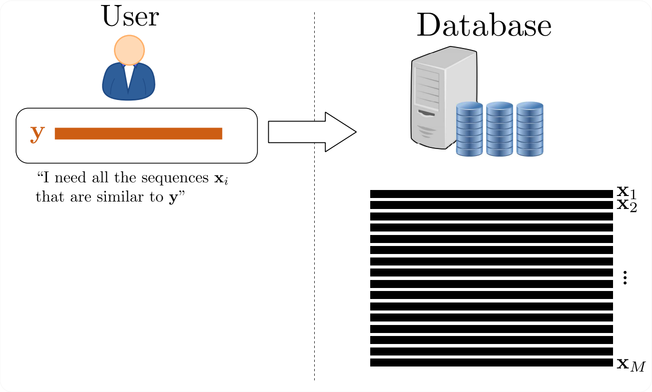
Today:

- "Compression for Queries"
- Compression – minimize space required to store database
- Compressed data does not represent the source itself – but rather "some useful information about the source"

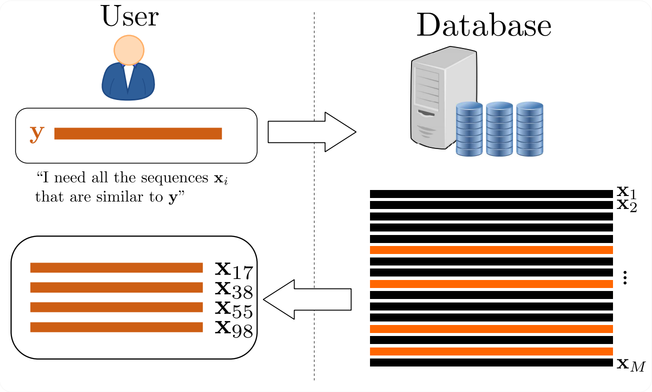
Similarity Queries in Databases



Similarity Queries in Databases



Similarity Queries in Databases



Applications

Any database with many long sequences and a similarity measure:

- Forensics: fingerprints
  - FBI: "Integrated automated fingerprint identification system (IAFIS)": data on more than 104M individuals [1]

[1] Source: www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis

[2] Source: NIH, www.ncbi.nlm.nih.gov/genbank.

[3] Source: Golan Yona, Dept. of Structural Biology, Stanford

Applications

Any database with many long sequences and a similarity measure:

- Forensics: fingerprints
  - FBI: "Integrated automated fingerprint identification system (IAFIS)": data on more than 104M individuals [1]
- Bioinformatics: DNA sequences
  - GenBank: 200M sequences[2]
  - Biozon: 100M records (DNA, proteins and more)[3]

---

[1] Source: www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis

[2] Source: NIH, www.ncbi.nlm.nih.gov/genbank.

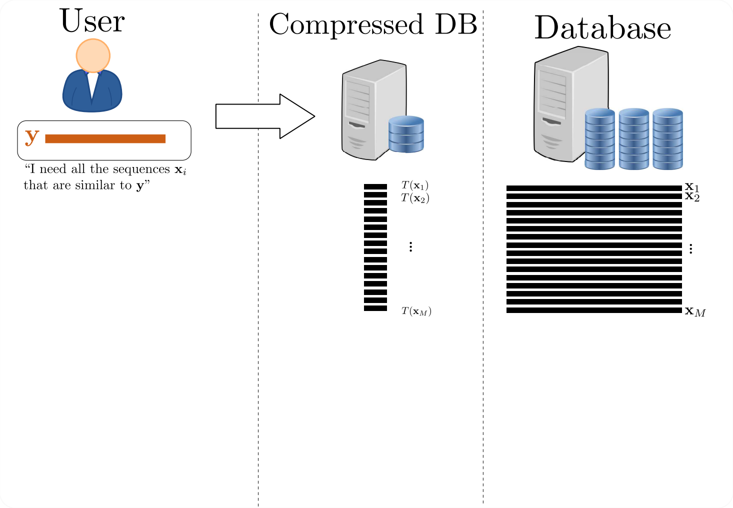
[3] Source: Golan Yona, Dept. of Structural Biology, Stanford

Similarity Queries on Compressed Data

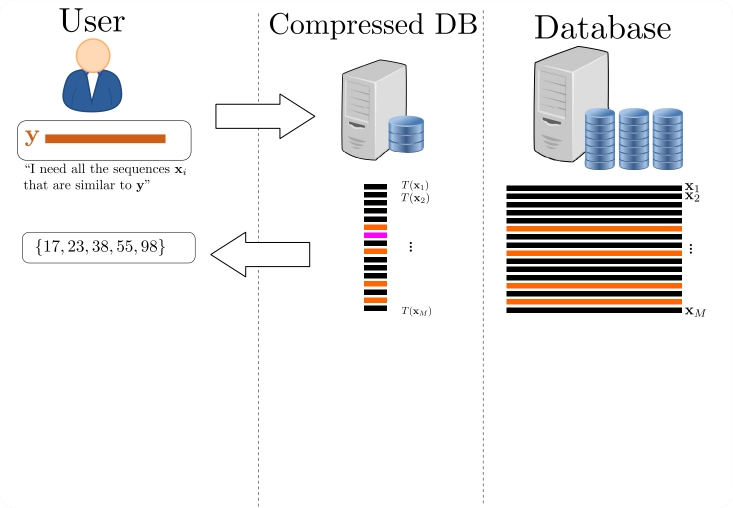
Today: detect similarity based on compressed data:

- For each sequence $\mathbf{x}$ in the database, store only a very small signature $T(\mathbf{x})$
- Need to decide whether $\mathbf{x}$ and $\mathbf{y}$ are similar given only $\mathbf{y}$, $T(\mathbf{x})$

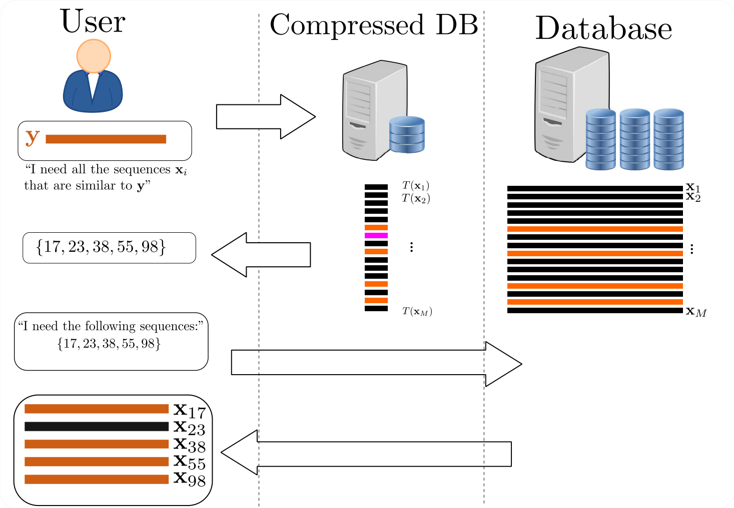
Similarity Queries on Compressed Data



Similarity Queries on Compressed Data



Similarity Queries on Compressed Data



Similarity Queries on Compressed Data: Remarks

- Not classical compression:
  - Original data not reproducible from compressed version
  - Compressed DB *does not replace the DB*

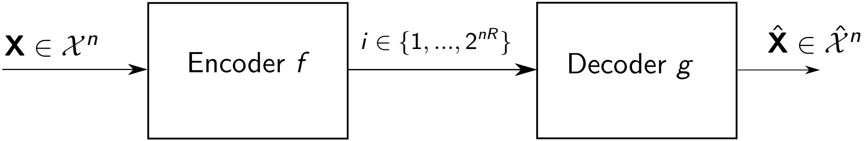
Similarity Queries on Compressed Data: Remarks

- Not classical compression:
  - Original data not reproducible from compressed version
  - Compressed DB *does not replace the DB*
- Beneficial when when access to full DB is costly, e.g. if
  - stored on slower media
  - stored in a remote location
  - full DB is used by many different users

Similarity Queries on Compressed Data: Remarks

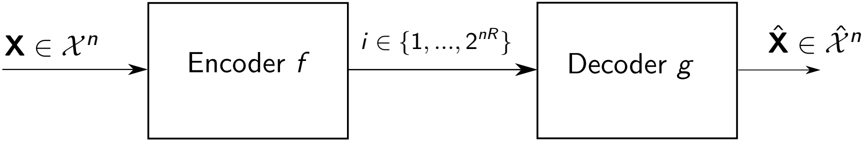
- Not classical compression:
  - Original data not reproducible from compressed version
  - Compressed DB *does not replace the DB*
- Beneficial when when access to full DB is costly, e.g. if
  - stored on slower media
  - stored in a remote location
  - full DB is used by many different users
- Queries answered w.r.t. compressed (i.e. partial) data are not always correct
  - False positive (FP)
  - False negatives (FN)

Compression



$$f : \mathcal{X}^n \to \{1, ..., 2^{nR}\}; \quad g : \{1, ..., 2^{nR}\} \to \hat{\mathcal{X}}^n$$

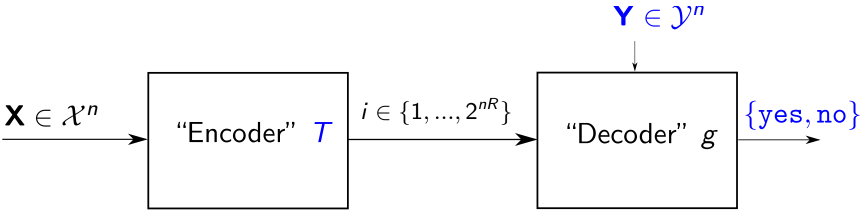
Compression



$$f : \mathcal{X}^n \to \{1, ..., 2^{nR}\}; \quad g : \{1, ..., 2^{nR}\} \to \hat{\mathcal{X}}^n$$

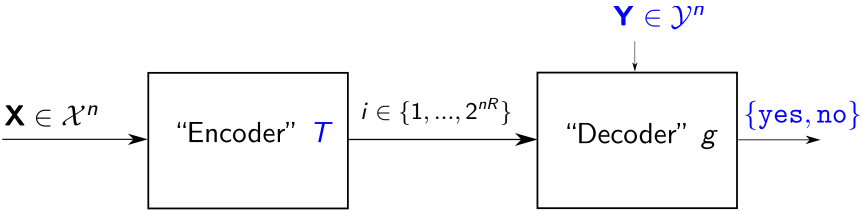
- Goal: Given $f(\mathbf{x})$, generate $\hat{\mathbf{x}}$ which is similar to $\mathbf{x}$.
  - (Nearly) Lossless Compression: $\Pr\{\mathbf{X} \neq \hat{\mathbf{X}}\} \to 0$
  - Lossy Compression: $\mathbb{E}[d(\mathbf{X}, \hat{\mathbf{X}})] \leq D$

Similarity Detection



$$T : \mathcal{X}^n \to \{1, ..., 2^{nR}\}; \quad g : \{1, ..., 2^{nR}\} \times \mathcal{Y}^n \to \{\texttt{yes}, \texttt{no}\}$$

Similarity Detection



$$T : \mathcal{X}^n \to \{1, ..., 2^{nR}\}; \quad g : \{1, ..., 2^{nR}\} \times \mathcal{Y}^n \to \{\texttt{yes}, \texttt{no}\}$$

- Goal: Given $\mathbf{y}$ and $T(\mathbf{x})$, determine whether $\mathbf{x}$ and $\mathbf{y}$ are similar.
  - "$\mathbf{x}$ and $\mathbf{y}$ are similar" $\Leftrightarrow d(\mathbf{x}, \mathbf{y}) \leq D$
  - A good scheme (T, g): the function g is correct "most of the time"

What makes a scheme "good"?

The errors $g(\cdot, \cdot)$ can make:

- False positives (FP): $g(T(\mathbf{x}), \mathbf{y}) = \texttt{yes}$ when $d(\mathbf{x}, \mathbf{y}) > D$
- False negative (FN): $g(T(\mathbf{x}), \mathbf{y}) = \texttt{no}$ when $d(\mathbf{x}, \mathbf{y}) \leq D$

What makes a scheme "good"?

The errors $g(\cdot, \cdot)$ can make:

- False positives (FP): $g(T(\mathbf{x}), \mathbf{y}) = \texttt{yes}$ when $d(\mathbf{x}, \mathbf{y}) > D$
- False negative (FN): $g(T(\mathbf{x}), \mathbf{y}) = \texttt{no}$ when $d(\mathbf{x}, \mathbf{y}) \leq D$

We focus on case where $\Pr\{\text{FN}\} = 0$.

- A FN causes an *undetected* error
- A FP does not incur an error *per se*, only increased computation / communication

Schemes with $\Pr\{\text{FN}\} = 0$ are said to be *admissible*.

What makes a scheme "good"?

The errors $g(\cdot,\cdot)$ can make:

- False positives (FP): $g(T(\mathbf{x}),\mathbf{y}) = \texttt{yes}$ when $d(\mathbf{x},\mathbf{y}) > D$
- False negative (FN): $g(T(\mathbf{x}),\mathbf{y}) = \texttt{no}$ when $d(\mathbf{x},\mathbf{y}) \le D$

We focus on case where $\Pr\{\text{FN}\} = 0$.

- A FN causes an *undetected* error
- A FP does not incur an error *per se*, only increased computation / communication

Schemes with $\Pr\{\text{FN}\} = 0$ are said to be *admissible*.

$\Rightarrow$ `no` means `no`; and `yes` means `maybe` !

What makes a scheme "good"?

The errors $g(\cdot,\cdot)$ can make:

- False positives (FP): $g(T(\mathbf{x}),\mathbf{y}) = \texttt{yes}$ when $d(\mathbf{x},\mathbf{y}) > D$
- False negative (FN): $g(T(\mathbf{x}),\mathbf{y}) = \texttt{no}$ when $d(\mathbf{x},\mathbf{y}) \leq D$

We focus on case where $\Pr\{\mathrm{FN}\} = 0$.

- A FN causes an *undetected* error
- A FP does not incur an error *per se*, only increased computation / communication

Schemes with $\Pr\{\mathrm{FN}\} = 0$ are said to be *admissible*.

$\Rightarrow$ `no` means `no`; and `yes` means `maybe` !

$$g : \{1, \ldots, 2^{nR}\} \times \mathcal{Y}^n \to \{\texttt{no}, \texttt{maybe}\}$$

A “good” scheme = low probability for `maybe`

Goal: Control the false positive probability

A “good” scheme = low probability for maybe

Goal: Control the false positive probability

$$
\begin{aligned}
&\Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \texttt{maybe}\} \\
&= \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \texttt{maybe} | d(\mathbf{X},\mathbf{Y}) \le D\} \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} \\
&\ + \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \texttt{maybe} | d(\mathbf{X},\mathbf{Y}) > D\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\}
\end{aligned}
$$

A "good" scheme = low probability for maybe

Goal: Control the false positive probability

$$
\begin{aligned}
&\Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe}\} \\
&= \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) \le D\} \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} \\
&\ + \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) > D\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\} \\
&= (1 - \Pr\{FN\}) \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} \\
&\ + \Pr\{FP\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\}
\end{aligned}
$$

A "good" scheme = low probability for maybe

Goal: Control the false positive probability

$$
\begin{aligned}
&\Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe}\} \\
&= \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) \le D\} \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} \\
&\quad + \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) > D\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\} \\
&= (1 - \Pr\{FN\}) \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} \\
&\quad + \Pr\{FP\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\} \\
&= \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} + \Pr\{FP\} \Pr\{d(\mathbf{X},\mathbf{Y}) > D\}.
\end{aligned}
$$

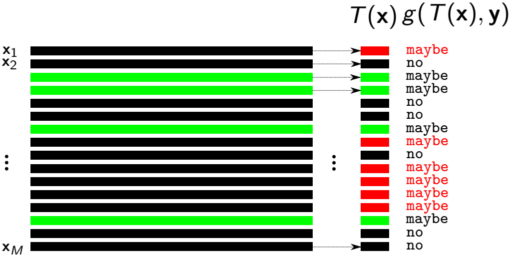
A "good" scheme = low probability for $\mathtt{maybe}$

Goal: Control the false positive probability

$$
\begin{aligned}
&\Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe}\}\\
&= \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) \le D\}\Pr\{d(\mathbf{X},\mathbf{Y}) \le D\}\\
&\ + \Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe} | d(\mathbf{X},\mathbf{Y}) > D\}\Pr\{d(\mathbf{X},\mathbf{Y}) > D\}\\
&= (1 - \Pr\{FN\})\Pr\{d(\mathbf{X},\mathbf{Y}) \le D\}\\
&\ + \Pr\{FP\}\Pr\{d(\mathbf{X},\mathbf{Y}) > D\}\\
&= \Pr\{d(\mathbf{X},\mathbf{Y}) \le D\} + \Pr\{FP\}\Pr\{d(\mathbf{X},\mathbf{Y}) > D\}.
\end{aligned}
$$

$\Pr\{g(T(\mathbf{X}),\mathbf{Y}) = \mathtt{maybe}\}$ minimized $\Leftrightarrow$ $\Pr\{FP\}$ minimized

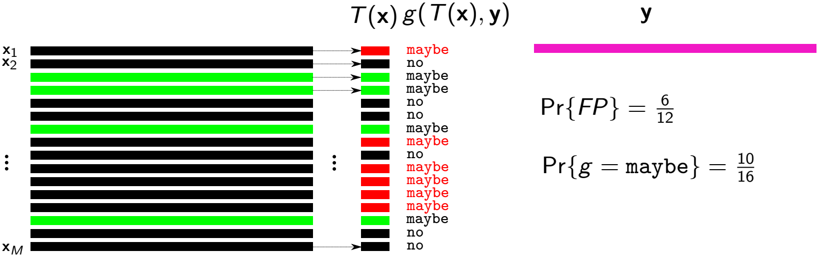
$\Pr\{g = \texttt{maybe}\}$: operational significance



$$\Pr\{FP\} = \frac{6}{12}$$

$$\Pr\{g = \texttt{maybe}\} = \frac{10}{16}$$

$\Pr\{g = \texttt{maybe}\}$: operational significance



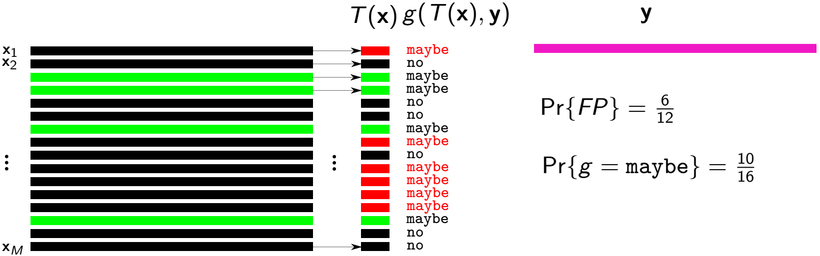
$\Pr\{FP\} = \frac{6}{12}$

$\Pr\{g = \texttt{maybe}\} = \frac{10}{16}$

$\Pr\{g = \texttt{maybe}\}$: the fraction of sequences retrieved from database

$\Rightarrow$ a proxy for complexity of answering a query

$\Pr\{g = \mathtt{maybe}\}$: operational significance



$\Pr\{FP\} = \frac{6}{12}$

$\Pr\{g = \mathtt{maybe}\} = \frac{10}{16}$

$\Pr\{g = \mathtt{maybe}\}$: the fraction of sequences retrieved from database $\Rightarrow$ a proxy for complexity of answering a query

We say that the query has been answered *reliably* if $\Pr\{g = \mathtt{maybe}\}$ is small.

Achievable Rates

$\mathbf{X} \sim$ i.i.d. $P_X(\cdot)$, $\mathbf{Y} \sim$ i.i.d. $P_Y(\cdot)$.
D is given (fixed) similarity threshold
– i.e. $\mathbf{x}, \mathbf{y}$ similar means $d(\mathbf{x}, \mathbf{y}) \leq D$.

Definition

Rate R is said to be *D-achievable* if there exists a sequence of rate-R admissible schemes $\left\{T^{(n)}, g^{(n)}\right\}$, s.t.

$$\lim_{n \to \infty} \Pr\left\{g^{(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \texttt{maybe}\right\} = 0.$$

Achievable Rates

$\mathbf{X} \sim$ i.i.d. $P_X(\cdot)$, $\mathbf{Y} \sim$ i.i.d. $P_Y(\cdot)$.
D is given (fixed) similarity threshold
– i.e. $\mathbf{x}, \mathbf{y}$ similar means $d(\mathbf{x}, \mathbf{y}) \leq D$.

Definition

Rate R is said to be *D-achievable* if there exists a sequence of rate-R admissible schemes $\left\{T^{(n)}, g^{(n)}\right\}$, s.t.

$$\lim_{n\to\infty} \Pr\left\{g^{(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \texttt{maybe}\right\} = 0.$$

Why does this model & definition make sense?

Identification Rate

Definition

For a similarity threshold D, the *identification rate* $R_{\mathrm{ID}}(D)$ is the infimum of D-achievable rates. That is,

$$R_{\mathrm{ID}}(D) \triangleq \inf\{R : R \text{ is } D\text{-achievable}\}.$$

Identification Rate

Definition

For a similarity threshold D, the *identification rate* $R_{\mathrm{ID}}(D)$ is the infimum of D-achievable rates. That is,

$$R_{\mathrm{ID}}(D) \triangleq \inf\{R : R \text{ is } D\text{-achievable}\}.$$

In other words, $R_{\mathrm{ID}}(D)$ is a *fundamental limit*. It is the degree to which we can compress the data, while retaining the ability to reliably answer similarity queries.

Identification Exponent

If $R > R_{\mathrm{ID}}(D)$, then $\Pr\{g = \mathtt{maybe}\}$ can be made arbitrarily small with n. How fast? (i.e., how precisely can we control the false-positive probability?)

Identification Exponent

If $R > R_{\mathrm{ID}}(D)$, then $\Pr\{g = \mathtt{maybe}\}$ can be made arbitrarily small with n. How fast? (i.e., how precisely can we control the false-positive probability?)

Definition

Fix $R > R_{\mathrm{ID}}(D)$. The *identification exponent* is defined as

$$\mathbf{E}_{\mathrm{ID}}(R) \triangleq \limsup_{n\to\infty} -\frac{1}{n}\log \Pr\left\{g^{(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \mathtt{maybe}\right\}$$

$g^{(n)}$, $T^{(n)}$: optimal schemes at rate R and length n.

Identification Exponent

If $R > R_{\mathrm{ID}}(D)$, then $\Pr\{g = \mathtt{maybe}\}$ can be made arbitrarily small with n. How fast? (i.e., how precisely can we control the false-positive probability?)

Definition

Fix $R > R_{\mathrm{ID}}(D)$. The *identification exponent* is defined as

$$\mathbf{E}_{\mathrm{ID}}(R) \triangleq \limsup_{n\to\infty} -\frac{1}{n} \log \Pr\left\{g^{(n)}\left(T^{(n)}(\mathbf{X}), \mathbf{Y}\right) = \mathtt{maybe}\right\}$$

$g^{(n)}$, $T^{(n)}$: optimal schemes at rate R and length n.

Can also pursue other directions

- e.g., finite blocklength bounds

The Quadratic-Gaussian case

- Quadratic distortion: $d(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\|\mathbf{x} - \mathbf{y}\|^2$
- Gaussian source: $\mathbf{X} \sim N(0, I\sigma^2)$, $\mathbf{Y} \sim N(0, I\sigma^2)$; $\mathbf{X}, \mathbf{Y}$ independent.

QG: the Identification Rate

Theorem (Ingber, Courtade, Weissman, DCC 2013)

Suppose $\mathbf{X} \sim N(0, I\sigma^2)$, $\mathbf{Y} \sim N(0, I\sigma^2)$; $\mathbf{X}, \mathbf{Y}$ *independent. Then*

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases}$$

Quadratic-Gaussian Case: Discussion

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases}$$

Quadratic-Gaussian Case: Discussion

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases}$$

- If $D > 2\sigma^2$,

  $\Rightarrow$ $\mathbf{X}$ and $\mathbf{Y}$ are naturally similar! [i.e. $d(\mathbf{X}, \mathbf{Y}) \leq D$ w.h.p.]

  $\Rightarrow$ $R_{\mathrm{ID}}(D) = \infty$,

Quadratic-Gaussian Case: Discussion

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases}$$

- If $D > 2\sigma^2$,
  $\Rightarrow$ **X** and **Y** are naturally similar! [i.e. $d(\mathbf{X}, \mathbf{Y}) \leq D$ w.h.p.]
  $\Rightarrow$ $R_{\mathrm{ID}}(D) = \infty$,
- If $D \to 0$, then asking "are $\mathbf{x}, \mathbf{y}$ similar?" is like asking whether $\mathbf{x} = \mathbf{y}$, so very little information is required to rule out most of the **x**'s

Quadratic-Gaussian Case: Discussion

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right) & \text{for } D < 2\sigma^2 \\ \infty & \text{for } D \geq 2\sigma^2. \end{cases}$$

- If $D > 2\sigma^2$,
  $\Rightarrow$ **X** and **Y** are naturally similar! [i.e. $d(\mathbf{X}, \mathbf{Y}) \leq D$ w.h.p.]
  $\Rightarrow$ $R_{\mathrm{ID}}(D) = \infty$,
- If $D \to 0$, then asking "are $\mathbf{x}, \mathbf{y}$ similar?" is like asking whether $\mathbf{x} = \mathbf{y}$, so very little information is required to rule out most of the **x**'s
- Similarity to classic rate distortion:

$$R(D) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) & \text{for } D < \sigma^2 \\ 0 & \text{for } D \geq \sigma^2. \end{cases}$$

Identification Rate vs Rate-Distortion

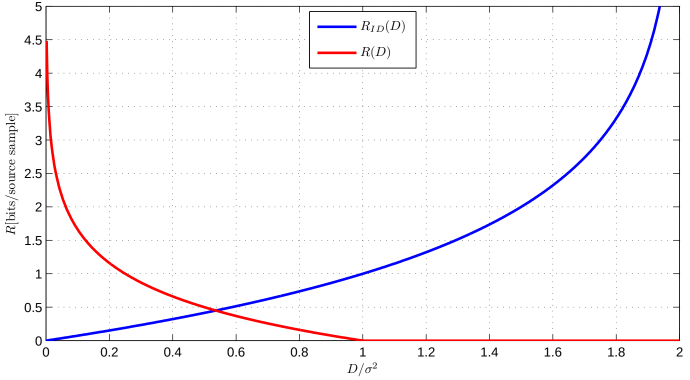


Figure: The rate distortion function $R(D)$ and the identification rate $R_{\mathrm{ID}}(D)$ of a Gaussian source with variance σ^2.

QG Identification Exponent

Theorem (Ingber, Courtade, Weissman, DCC 2013)

Suppose $\mathbf{X} \sim N(0, I\sigma^2)$, $\mathbf{Y} \sim N(0, I\sigma^2)$; $\mathbf{X}, \mathbf{Y}$ *independent.*
Then for $R > R_{\text{ID}}(D)$,

$$\mathbf{E}_{\text{ID}}(R) =$$

$$\min_{\rho \in (0,1]} 2\mathbf{E}_Z(\rho) - \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho - \frac{D}{2\sigma^2}}{\rho}, \frac{\pi}{2} \right]$$

where $\mathbf{E}_Z(\rho) \triangleq \frac{1}{\ln 2} \left[\frac{\rho}{2} - \frac{1}{2} - \frac{1}{2} \ln \rho \right]$.

QG Identification Exponent: Discussion

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho \in (0,1]} 2\mathbf{E}_Z(\rho) - \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho - \frac{D}{2\sigma^2}}{\rho}, \frac{\pi}{2} \right]$$

QG Identification Exponent: Discussion

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho \in (0,1]} 2\mathbf{E}_Z(\rho) - \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho - \frac{D}{2\sigma^2}}{\rho}, \frac{\pi}{2} \right]$$

- Only scalar minimization w.r.t. $\rho \Rightarrow$ easily computed

QG Identification Exponent: Discussion

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho \in (0,1]} 2\mathbf{E}_Z(\rho) - \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho - \frac{D}{2\sigma^2}}{\rho}, \frac{\pi}{2} \right]$$

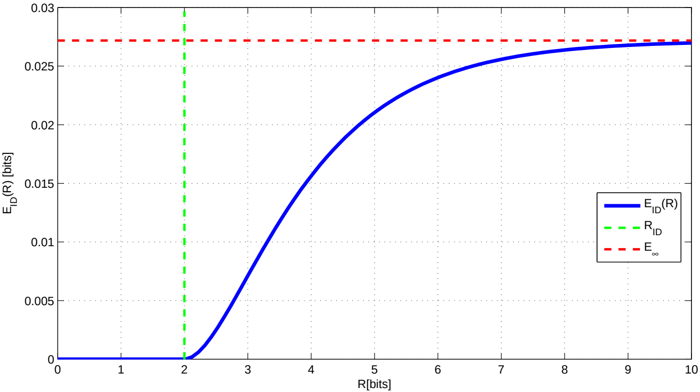
- Only scalar minimization w.r.t. $\rho \Rightarrow$ easily computed
- $\mathbf{E}_{\mathrm{ID}}(R_{\mathrm{ID}}(D)) = 0$, as expected

QG Identification Exponent: Discussion

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho \in (0,1]} 2\mathbf{E}_Z(\rho) - \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho - \frac{D}{2\sigma^2}}{\rho}, \frac{\pi}{2} \right]$$

- Only scalar minimization w.r.t. $\rho \Rightarrow$ easily computed
- $\mathbf{E}_{\mathrm{ID}}(R_{\mathrm{ID}}(D)) = 0$, as expected
- $\lim_{R \to \infty} \mathbf{E}_{\mathrm{ID}}(R)$ is given by the exponential decay factor of the event $\{d(\mathbf{X}, \mathbf{Y}) \leq D\}$.

$\mathbf{E}_{\mathrm{ID}}(R)$ for $R_{\mathrm{ID}}(D) = 2$ bits/sym



Different Variance

Suppose $\mathbf{X} \sim N(0, I\sigma_X^2)$, $\mathbf{Y} \sim N(0, I\sigma_Y^2)$; $\mathbf{X}, \mathbf{Y}$ independent. Then

Theorem

$$R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2) = \begin{cases} \log \frac{2\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2-D} & \textit{for } D < \sigma_X^2 + \sigma_Y^2 \\ \infty & \textit{for } D \geq \sigma_X^2 + \sigma_Y^2. \end{cases}$$

Different Variance

Suppose $\mathbf{X} \sim N(0, I\sigma_X^2)$, $\mathbf{Y} \sim N(0, I\sigma_Y^2)$; $\mathbf{X}, \mathbf{Y}$ independent. Then

Theorem

$$R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2) = \begin{cases} \log \frac{2\sigma_X\sigma_Y}{\sigma_X^2+\sigma_Y^2-D} & \text{for } D < \sigma_X^2 + \sigma_Y^2 \\ \infty & \text{for } D \geq \sigma_X^2 + \sigma_Y^2. \end{cases}$$

Theorem

For $R > R_{\mathrm{ID}}(D, \sigma_X^2, \sigma_Y^2)$,

$$\mathbf{E}_{\mathrm{ID}}(R) = \min_{\rho_X, \rho_Y > 0} \mathbf{E}_Z(\rho_X) + \mathbf{E}_Z(\rho_Y)$$
$$- \log \sin \min \left[\sin^{-1}(2^{-R}) + \cos^{-1} \frac{\rho_X\sigma_X^2 + \rho_Y\sigma_Y^2 - D}{2\sigma_X\sigma_Y\sqrt{\rho_X\rho_Y}}, \frac{\pi}{2} \right]$$

General Sources: Achievable Rate

Theorem

***X** and **Y** independent, $\sim$ i.i.d. P_X, finite second moment. Then*

$$R_{\mathrm{ID}}(D) \leq \inf_{P_{\hat{X}|X}} I(X; \hat{X})$$

inf is w.r.t. all test channels $P_{\hat{X}|X}$ satisfying

$$\sqrt{\mathbb{E}_{P_X \otimes P_{\hat{X}}}(X - \hat{X})^2} \geq \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X - \hat{X})^2} + \sqrt{D}$$

General Sources: About the Result

- Works for any $d(\cdot,\cdot)$ that satisfies the *triangle inequality*

  A version exists for general $d(\cdot,\cdot)$
- Easily extended to different P_X,P_Y

General Sources: About the Result

- Works for any $d(\cdot,\cdot)$ that satisfies the *triangle inequality*
  A version exists for general $d(\cdot,\cdot)$
- Easily extended to different P_X,P_Y
- Similar in spirit to [Ahlswede, Yang, Zhang '93]
  - study a related problem

Gaussian as an Extreme Case

Classical lossy source coding: *among all sources with the same variance, the Gaussian is the hardest to compress*.

Gaussian as an Extreme Case

Classical lossy source coding: *among all sources with the same variance, the Gaussian is the hardest to compress.*
In our case:

Theorem

If X is a random variable with finite variance σ^2, then

$$R_{ID}(D) \leq \log\left(\frac{1}{1-\frac{D}{2\sigma^2}}\right),$$

i.e. a Gaussian source X requires the largest identification rate for a given variance.

Gaussian as an Extreme Case: Proof #1

Take a distribution P_X (assume $E[X]=0$). Then $R_{\mathrm{ID}}(D) \leq \inf_{P_{\hat{X}|X}} I(X;\hat{X})$, where inf is w.r.t. $P_{\hat{X}|X}$ s.t. $\sqrt{\mathbb{E}_{P_X \otimes P_{\hat{X}}}(X-\hat{X})^2} \geq \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} + \sqrt{D}$.

Gaussian as an Extreme Case: Proof #1

Take a distribution P_X (assume $E[X]=0$). Then $R_{\mathrm{ID}}(D) \le \inf_{P_{\hat{X}|X}} I(X;\hat{X})$, where inf is w.r.t. $P_{\hat{X}|X}$ s.t. $\sqrt{\mathbb{E}_{P_X \otimes P_{\hat{X}}}(X-\hat{X})^2} \ge \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} + \sqrt{D}$.

Choose a channel $P_{\hat{X}|X}$: $\hat{X} = \rho X + Z$; $Z \sim N(0,\sigma_Z^2)$, ind. of X, and

$$\rho = \frac{(4\sigma^2 - D)}{(2\sigma^2)}; \quad \sigma_Z^2 = \frac{(4\sigma^2 - D)(2\sigma^2 - D)^2}{4\sigma^2 D}.$$

Constraints on $P_{\hat{X}|X}$ are satisfied.

Gaussian as an Extreme Case: Proof #1

Take a distribution P_X (assume $E[X]=0$). Then $R_{\mathrm{ID}}(D) \le \inf_{P_{\hat{X}|X}} I(X;\hat{X})$, where inf is w.r.t. $P_{\hat{X}|X}$ s.t. $\sqrt{\mathbb{E}_{P_X \otimes P_{\hat{X}}}(X-\hat{X})^2} \ge \sqrt{\mathbb{E}_{P_{X,\hat{X}}}(X-\hat{X})^2} + \sqrt{D}$.

Choose a channel $P_{\hat{X}|X}$: $\hat{X} = \rho X + Z$; $Z \sim N(0,\sigma_Z^2)$, ind. of X, and

$$\rho = \frac{(4\sigma^2 - D)}{(2\sigma^2)}; \quad \sigma_Z^2 = \frac{(4\sigma^2 - D)(2\sigma^2 - D)^2}{4\sigma^2 D}.$$

Constraints on $P_{\hat{X}|X}$ are satisfied.

$VAR[\hat{X}] = \rho^2\sigma^2 + \sigma_Z^2 \Rightarrow$

$$I(X;\hat{X}) = h(\hat{X}) - h(\hat{X}|X) \le \frac{1}{2}\log\frac{\rho^2\sigma^2 + \sigma_Z^2}{\sigma_Z^2} = \log\frac{1}{1 - D/(2\sigma^2)}$$

[since Gaussian maximizes diff. entropy for a given variance]

A Universal Scheme [+ Proof #2]

A scheme, that for any P_X, attains R_{ID} of a Gaussian:

A Universal Scheme [+ Proof #2]

A scheme, that for any P_X, attains R_{ID} of a Gaussian:

Assume $n = 2^\ell$. Let

$$\mathbb{X} = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)].$$

Now define

$$[\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \ldots, \tilde{\mathbf{X}}(n)] = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)] \times H_\ell$$

H_ℓ: a Hadamard matrix of order $n = 2^\ell$. Do the same with $\tilde{\mathbf{Y}}(i)$.

A Universal Scheme [+ Proof #2]

A scheme, that for any P_X, attains R_{ID} of a Gaussian:

Assume $n = 2^\ell$. Let
$$\mathbb{X} = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)].$$
Now define
$$[\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \ldots, \tilde{\mathbf{X}}(n)] = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)] \times H_\ell$$
H_ℓ: a Hadamard matrix of order $n = 2^\ell$. Do the same with $\tilde{\mathbf{Y}}(i)$.

- As n grows, the elements of each $\tilde{\mathbf{X}}(i)$ become Gaussian (CLT)
- The columns of $\mathbf{X}$ remain independent!
- Apply a length-n Gaussian scheme on each $\tilde{\mathbf{X}}(i)$.
- Union bound $\rightarrow$ vanishing $\Pr\{g = \mathtt{maybe}\}$!

A Universal Scheme [+ Proof #2]

A scheme, that for any P_X, attains R_{ID} of a Gaussian:

Assume $n = 2^\ell$. Let

$$\mathbb{X} = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)].$$

Now define

$$[\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \ldots, \tilde{\mathbf{X}}(n)] = [\mathbf{X}(1), \mathbf{X}(2), \ldots, \mathbf{X}(n)] \times H_\ell$$

H_ℓ: a Hadamard matrix of order $n = 2^\ell$. Do the same with $\tilde{\mathbf{Y}}(i)$.

- As n grows, the elements of each $\tilde{\mathbf{X}}(i)$ become Gaussian (CLT)
- The columns of $\mathbf{X}$ remain independent!
- Apply a length-n Gaussian scheme on each $\tilde{\mathbf{X}}(i)$.
- Union bound $\rightarrow$ vanishing $\Pr\{g = \texttt{maybe}\}$!

More than just another proof – this provides a scheme which is minimax optimal w.r.t. all sources with variance σ^2.

The Symmetric Binary-Hamming case

Suppose $\mathbf{X}, \mathbf{Y} \sim \mathrm{Ber}(\frac{1}{2})$ and distance is measured under Hamming distortion

Theorem

$$\begin{aligned} R_{\mathrm{ID}}(D) &= 1 - h\left(\tfrac{1}{2} - D\right) \\ &= D^2 \cdot 2\log e + o(D^2) \end{aligned}$$

- $h(\cdot)$: binary entropy function
- Classic rate distortion: $R(D) = 1 - h(D)$

General Sources under Hamming Distortion

Theorem

If $\mathbf{X}, \mathbf{Y}$ *are both drawn i.i.d. according to* P_X *and similarity is measured under Hamming loss,*

$$R_{\mathrm{ID}}(D) \geq D^2 \cdot 2 \log e.$$

General Sources under Hamming Distortion

Theorem

If $\mathbf{X}, \mathbf{Y}$ *are both drawn i.i.d. according to* P_X *and similarity is measured under Hamming loss,*

$$R_{\mathrm{ID}}(D) \geq D^2 \cdot 2 \log e.$$

- For $P_X = \mathrm{Ber}(\frac{1}{2})$, recall $R_{\mathrm{ID}}(D) = D^2 \cdot 2 \log e + o(D^2)$.

General Sources under Hamming Distortion

Theorem

If $\mathbf{X}, \mathbf{Y}$ *are both drawn i.i.d. according to* P_X *and similarity is measured under Hamming loss,*

$$R_{\mathrm{ID}}(D) \geq D^2 \cdot 2 \log e.$$

- For $P_X = \mathrm{Ber}(\frac{1}{2})$, recall $R_{\mathrm{ID}}(D) = D^2 \cdot 2 \log e + o(D^2)$.
- $\Rightarrow \mathrm{Ber}(\frac{1}{2})$ is nearly "easiest" to compress (in interesting regime of small D) of *all* sources when distortion measured under Hamming loss.

General Sources under Hamming Distortion

Theorem

If $\mathbf{X}, \mathbf{Y}$ *are both drawn i.i.d. according to* P_X *and similarity is measured under Hamming loss,*

$$R_{\mathrm{ID}}(D) \geq D^2 \cdot 2 \log e.$$

- For $P_X = \mathrm{Ber}(\frac{1}{2})$, recall $R_{\mathrm{ID}}(D) = D^2 \cdot 2 \log e + o(D^2)$.
- $\Rightarrow \mathrm{Ber}(\frac{1}{2})$ is nearly "easiest" to compress (in interesting regime of small D) of *all* sources when distortion measured under Hamming loss.
- Stark contrast to Quadratic-Gaussian setting!

Towards a general $R_{\mathrm{ID}}(D)$:

So far, we saw several examples:

- Quadratic-Gaussian
- Quadratic-general
- Symmetric Binary-Hamming
- General DMS & Hamming
- DMS (results depend on an aux. RV with unbounded card.)

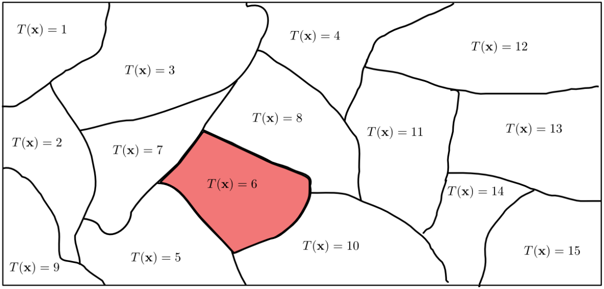
Towards a general $R_{\mathrm{ID}}(D)$:

So far, we saw several examples:

- Quadratic-Gaussian
- Quadratic-general
- Symmetric Binary-Hamming
- General DMS & Hamming
- DMS (results depend on an aux. RV with unbounded card.)

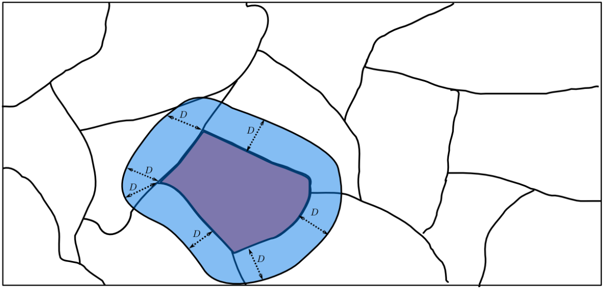
Why no general solution?

Identification schemes as Quantizers



- Size of quantization cell $\propto \Pr(T(\mathbf{X}) = i) \approx 2^{-nR}$ (symmetry)
- Expanded quantization cells: $\{\mathbf{y} : d(\mathbf{x}, \mathbf{y}) \leq D$ for **some** $\mathbf{x}$ in cell$\}$
- $\Pr(\texttt{maybe}) \propto$ size of expanded cell

Identification schemes as Quantizers



- Size of quantization cell $\propto \Pr(T(\mathbf{X}) = i) \approx 2^{-nR}$ (symmetry)
- Expanded quantization cells: $\{\mathbf{y} : d(\mathbf{x}, \mathbf{y}) \le D$ for **some** $\mathbf{x}$ in cell$\}$
- $\Pr(\texttt{maybe}) \propto$ size (i.e., measure) of expanded cell

Toward a converse:

Need to minimize size of expanded cell, for a given size of base cell

- A set A, its expansion $\Gamma^D(A)$
- What set A minimizes $|\Gamma^D(A)|$ for a fixed $|A|$?

Toward a converse:

Need to minimize size of expanded cell, for a given size of base cell

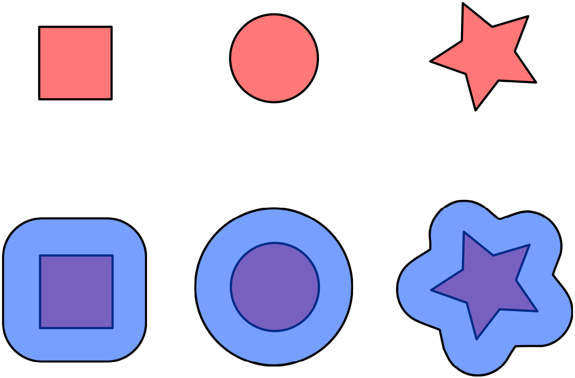
- A set A, its expansion $\Gamma^D(A)$
- What set A minimizes $|\Gamma^D(A)|$ for a fixed $|A|$?

$\Rightarrow$ an Isoperimetric Inequality!
What domain? The typical set!

- Where the probability is uniform
- Contains most of the probability mass

Isoperimetric Inequality in $\mathbb{R}^2$, Euclidean distance

$|\Gamma^D(A)|$ minimized when A is a **sphere**



Different Isoperimetric Inequalities

Domain	$d(\cdot,\cdot)$	Minimizer	When	Converse for
$\mathbb{R}^n$	Euclidean	n-sphere	late 1800's	–

Different Isoperimetric Inequalities

Domain	$d(\cdot,\cdot)$	Minimizer	When	Converse for
$\mathbb{R}^n$	Euclidean	n-sphere	late 1800's	–
n-dim. spherical shell	Euclidean/ Geodesic	Spherical cap	Levy '51	Quadratic-Gaussian

Different Isoperimetric Inequalities

Domain	$d(\cdot, \cdot)$	Minimizer	When	Converse for
$\mathbb{R}^n$	Euclidean	n-sphere	late 1800's	–
n-dim. spherical shell	Euclidean/ Geodesic	Spherical cap	Levy '51	Quadratic-Gaussian
Binary hypercube	Hamming	Hamming ball	Harper '66	Symmetric Binary-Hamming

Different Isoperimetric Inequalities

Domain	$d(\cdot,\cdot)$	Minimizer	When	Converse for
$\mathbb{R}^n$	Euclidean	n-sphere	late 1800's	–
n-dim. spherical shell	Euclidean/ Geodesic	Spherical cap	Levy '51	Quadratic-Gaussian
Binary hypercube	Hamming	Hamming ball	Harper '66	Symmetric Binary-Hamming
r-sets	Hamming	restricted Hamming ball ?	–	General Binary-Hamming
Type class	general	cond. type class ("V-shell")?	–	DMS and general $d(\cdot,\cdot)$

Different Isoperimetric Inequalities

Domain	$d(\cdot,\cdot)$	Minimizer	When	Converse for
$\mathbb{R}^n$	Euclidean	n-sphere	late 1800's	–
n-dim. spherical shell	Euclidean/ Geodesic	Spherical cap	Levy '51	Quadratic-Gaussian
Binary hypercube	Hamming	Hamming ball	Harper '66	Symmetric Binary-Hamming
r-sets	Hamming	restricted Hamming ball ?	–	General Binary-Hamming
Type class	general	cond. type class ("V-shell")?	–	DMS and general $d(\cdot,\cdot)$

$\Rightarrow$ an isoperimetric inequality implies a converse

- Might be too much to ask for
- But known in several special cases...

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative
- Quantities of interest: Identification rate and exponent

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative
- Quantities of interest: Identification rate and exponent
  - Complete solution for quadratic-Gaussian, symmetric binary-Hamming

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative
- Quantities of interest: Identification rate and exponent
  - Complete solution for quadratic-Gaussian, symmetric binary-Hamming
  - Achievability result for general sources, similarity metrics

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative
- Quantities of interest: Identification rate and exponent
  - Complete solution for quadratic-Gaussian, symmetric binary-Hamming
  - Achievability result for general sources, similarity metrics
  - "Universal" lower bound for Hamming loss

Summary

Compression for similarity queries

- Compression for purpose of answering queries reliably, rather than reproducing data
- Reliability $\triangleq$ vanishing probability of false positive, zero probability of false negative
- Quantities of interest: Identification rate and exponent
  - Complete solution for quadratic-Gaussian, symmetric binary-Hamming
  - Achievability result for general sources, similarity metrics
  - "Universal" lower bound for Hamming loss
  - A matching converse: implied by an appropriate *isoperimetric inequality*

What's next?

- Theory
  - Close the gap in the general case
  - Extensions: $\mathbf{X}, \mathbf{Y}$ non-i.i.d., but satisfying sparsity constraints

What's next?

- Theory
  - Close the gap in the general case
  - Extensions: $\mathbf{X}, \mathbf{Y}$ non-i.i.d., but satisfying sparsity constraints
- Applications:
  - Quadratic-Gaussian: spherical codes, lattices, wrapping
  - Symmetric Binary-Hamming: LDGM codes (already working on this...)
  - Bioinformatics (with Golan Yona, Stanford)

What's next?

- Theory
  - Close the gap in the general case
  - Extensions: $\mathbf{X}, \mathbf{Y}$ non-i.i.d., but satisfying sparsity constraints

- Applications:
  - Quadratic-Gaussian: spherical codes, lattices, wrapping
  - Symmetric Binary-Hamming: LDGM codes (already working on this...)
  - Bioinformatics (with Golan Yona, Stanford)

THANKS!