Sample Complexity of Estimating Entropy

Himanshu Tyagi
Indian Institute of Science, Bangalore

Joint work with Jayadev Acharya, Ananda Theertha Suresh, and Alon Orlitsky
Measuring Randomness in Data

Estimating randomness of the observed data:

Neural signal processing

Feature selection for machine learning

Image Registration
Measuring Randomness in Data

Estimating randomness of the observed data:

Neural signal processing

Feature selection for machine learning

Image Registration

Approach: Estimate the “entropy” of the generating distribution
Measuring Randomness in Data

Estimating randomness of the observed data:

Neural signal processing  Feature selection for machine learning

Approach: Estimate the “entropy” of the generating distribution

Shannon entropy \( H(p) \overset{\text{def}}{=} \sum_x -p_x \log p_x \)
Estimating Shannon Entropy

For an (unknown) distribution \( p \) with a (unknown) support-size \( k \),

How many samples are needed for estimating \( H(p) \)?
Estimating Shannon Entropy

For an \textit{(unknown)} distribution \( p \) with a \textit{(unknown)} support-size \( k \),

How many samples are needed for estimating \( H(p) \)?

PAC Framework or Large Deviation Guarantees

Let \( X^n = X_1, \ldots, X_n \) denote \( n \) independent samples from \( p \)

Performance of an estimator \( \hat{H} \) is measured by

\[
S^{\hat{H}}(\delta, \epsilon, k) \overset{\text{def}}{=} \min \left\{ n : p^n \left( |\hat{H}(X^n) - H(p)| < \delta \right) > 1 - \epsilon, \forall p \text{ with support-size } k \right\}
\]

The sample complexity of estimating Shannon Entropy is defined as

\[
S(\delta, \epsilon, k) \overset{\text{def}}{=} \min_{\hat{H}} S^{\hat{H}}(\delta, \epsilon, k)
\]
Focus only on the dependence of $S(\delta, \epsilon, k)$ on $k$

- Asymptotically consistent and normal estimators: [Miller55], [Mokkadem89], [AntosK01]

- [Paninski03] For the empirical estimator $\hat{H}_e$, $S(\hat{H}_e(k)) \leq O(k)$

- [Paninski04] There exists an estimator $\hat{H}$ s.t. $S(\hat{H}(k)) \leq o(k)$

- [ValiantV11] $S(k) = \Theta(k/\log k)$
  - The proposed estimator is constructive and is based on a LP
  - See, also, [WuY14], [JiaoVW14] for new proofs
Focus only on the dependence of $S(\delta, \epsilon, k)$ on $k$

- Asymptotically consistent and normal estimators: [Miller55], [Mokkadem89], [AntosK01]

- [Paninski03] For the empirical estimator $\hat{H}_e$, $S^{\hat{H}_e}(k) \leq O(k)$

- [Paninski04] There exists an estimator $\hat{H}$ s.t. $S^{\hat{H}}(k) \leq o(k)$

- [ValiantV11] $S(k) = \Theta(k / \log k)$
  - The proposed estimator is constructive and is based on a LP
  - See, also, [WuY14], [JiaoVW14] for new proofs

But we can estimate the distribution itself using $O(k)$ samples.
Sample Complexity of Estimating Shannon Entropy

Focus only on the dependence of $S(\delta, \epsilon, k)$ on $k$

- Asymptotically consistent and normal estimators:
  [Miller55], [Mokkadem89], [AntosK01]

- [Paninski03] For the empirical estimator $\hat{H}_e$, $S(\hat{H}_e(k)) \leq O(k)$

- [Paninski04] There exists an estimator $\hat{H}$ s.t. $S(\hat{H})(k) \leq o(k)$

- [ValiantV11] $S(k) = \Theta(k/\log k)$
  - The proposed estimator is constructive and is based on a LP
  - See, also, [WuY14], [JiaoVW14] for new proofs

But we can estimate the distribution itself using $O(k)$ samples.

Is it easier to estimate some other entropy??
Estimating Rényi Entropy

**Definition.** The Rényi entropy of order $\alpha$, $0 < \alpha \neq 1$, for a distribution $p$ is given by

$$H_\alpha(p) = \frac{1}{1 - \alpha} \log \sum_x p_x^\alpha$$

**Sample Complexity of Estimating Rényi Entropy**

Performance of an estimator $\hat{H}$ is measured by

$$S_{\alpha}^{\hat{H}}(\delta, \epsilon, k) \overset{\text{def}}{=} \min \left\{ n : p^n \left( |\hat{H}(X^n) - H_\alpha(p)| < \delta \right) > 1 - \epsilon, \forall \ p \in \mathcal{P}_k \right\}$$

The sample complexity of estimating Rényi Entropy of order $\alpha$ is given by

$$S_{\alpha}(\delta, \epsilon, k) \overset{\text{def}}{=} \min_{\hat{H}} S_{\alpha}^{\hat{H}}(\delta, \epsilon, k)$$
Estimating Rényi Entropy

**Definition.** The *Rényi entropy* of order $\alpha$, $0 < \alpha \neq 1$, for a distribution $p$ is given by

$$H_\alpha(p) = \frac{1}{1 - \alpha} \log \sum_x p_x^\alpha$$

**Sample Complexity of Estimating Rényi Entropy**

Performance of an estimator $\hat{H}$ is measured by

$$S_{\hat{H}}(\delta, \epsilon, k) \equiv \min \left\{ n : p^n \left( |\hat{H}(X^n) - H_\alpha(p)| < \delta \right) > 1 - \epsilon, \forall p \in \mathcal{P}_k \right\}$$

The sample complexity of estimating Rényi Entropy of order $\alpha$ is given by

$$S_{\alpha}(\delta, \epsilon, k) \equiv \min_{\hat{H}} S_{\hat{H}}(\delta, \epsilon, k)$$

*We mainly seek to characterize the dependence of $S_{\alpha}(\delta, \epsilon, k)$ on $k$ and $\alpha$.*
Which Rényi Entropy is the Easiest to Estimate?

Notations:

\[ S_{\alpha}(k) \geq \Omega(k^{\beta}) \Rightarrow \text{for every } \eta > 0 \text{ and for all } \delta, \epsilon \text{ small,} \]

\[ S_{\alpha}(\delta, \epsilon, k) \geq k^{\beta - \eta}, \quad \text{for all } k \text{ large} \]

\[ S_{\alpha}(k) \leq O(k^{\beta}) \Rightarrow \text{there is a constant } c \text{ depending on } \delta, \epsilon \text{ s.t.} \]

\[ S_{\alpha}(\delta, \epsilon, k) \leq ck^{\beta}, \quad \text{for all } k \text{ large} \]

\[ S_{\alpha}(k) = \Theta(k^{\beta}) \Rightarrow \Omega(k^{\beta}) \leq S_{\alpha}(k) \leq O(k^{\beta}) \]
Which Rényi Entropy is the Easiest to Estimate?

**Notations:**

\[ S_\alpha(k) \geq \tilde{\Omega}(k^\beta) \Rightarrow \text{for every } \eta > 0 \text{ and for all } \delta, \epsilon \text{ small,} \]

\[ S_\alpha(\delta, \epsilon, k) \geq k^{\beta - \eta}, \quad \text{for all } k \text{ large} \]

\[ S_\alpha(k) \leq O(k^\beta) \Rightarrow \text{there is a constant } c \text{ depending on } \delta, \epsilon \text{ s.t.} \]

\[ S_\alpha(\delta, \epsilon, k) \leq ck^\beta, \quad \text{for all } k \text{ large} \]

\[ S_\alpha(k) = \Theta(k^\beta) \Rightarrow \Omega(k^\beta) \leq S_\alpha(k) \leq O(k^\beta) \]

**Theorem**

**For every** \(0 < \alpha < 1:**

\[ \tilde{\Omega}(k^{1/\alpha}) \leq S_\alpha(k) \leq O(k^{1/\alpha} / \log k) \]

**For every** \(1 < \alpha \notin \mathbb{N}:**

\[ \tilde{\Omega}(k) \leq S_\alpha(k) \leq O(k / \log k) \]
Which Rényi Entropy is the Easiest to Estimate?

Notations:

\[ S_\alpha(k) \geq \Omega(k^\beta) \Rightarrow \text{for every } \eta > 0 \text{ and for all } \delta, \epsilon \text{ small,} \]
\[ S_\alpha(\delta, \epsilon, k) \geq k^{\beta - \eta}, \quad \text{for all } k \text{ large} \]

\[ S_\alpha(k) \leq O(k^\beta) \Rightarrow \text{there is a constant } c \text{ depending on } \delta, \epsilon \text{ s.t.} \]
\[ S_\alpha(\delta, \epsilon, k) \leq ck^\beta, \quad \text{for all } k \text{ large} \]

\[ S_\alpha(k) = \Theta(k^\beta) \Rightarrow \Omega(k^\beta) \leq S_\alpha(k) \leq O(k^\beta) \]

Theorem

For every \(0 < \alpha < 1\):
\[ \Omega(k^{1/\alpha}) \leq S_\alpha(k) \leq O(k^{1/\alpha} / \log k) \]

For every \(1 < \alpha \notin \mathbb{N}\):
\[ \Omega(k) \leq S_\alpha(k) \leq O(k / \log k) \]

For every \(1 < \alpha \in \mathbb{N}\):
\[ S_\alpha(k) = \Theta(k^{1-1/\alpha}) \]
Which Rényi Entropy is the Easiest to Estimate?

Theorem

For every $0 < \alpha < 1$: \( \tilde{\Omega}(k^{1/\alpha}) \leq S_\alpha(k) \leq O(k^{1/\alpha}/\log k) \)

For every $1 < \alpha \notin \mathbb{N}$: \( \tilde{\Omega}(k) \leq S_\alpha(k) \leq O(k/\log k) \)

For every $1 < \alpha \in \mathbb{N}$: \( S_\alpha(k) = \Theta(k^{1-1/\alpha}) \)
The $\alpha$th power sum of a distribution $p$ is given by

$$P_\alpha(p) \overset{\text{def}}{=} \sum_x p_\alpha^x$$

Estimating Rényi entropy with small additive error is the same as estimating power sum with small multiplicative error

- [Bar-YossefKS01] Integer moments of frequencies in a sequence with multiplicative and additive accuracies
- [JiaoVW14] Estimating power sums with small additive error
The $\alpha$th power sum of a distribution $p$ is given by

$$P_\alpha(p) \overset{\text{def}}{=} \sum_x p_x^\alpha$$

Estimating Rényi entropy with small additive error is the same as estimating power sum with small multiplicative error

- [Bar-YossefKS01] Integer moments of frequencies in a sequence with multiplicative and additive accuracies
- [JiaoVW14] Estimating power sums with small additive error

For $\alpha < 1$: Additive and multiplicative accuracy estimation have roughly the same sample complexity

For $\alpha > 1$: Additive accuracy estimation requires only a constant number of samples
The Estimators
Empirical or Plug-in Estimator

Given \( n \) samples \( X_1, \ldots, X_n \),

Let \( N_x \) denote the empirical frequency of \( x \).

\[
\hat{p}_n(x) \overset{\text{def}}{=} \frac{N_x}{n}
\]
\[
\hat{H}_\alpha^e \overset{\text{def}}{=} \frac{1}{1 - \alpha} \log \sum \hat{p}_n(x)^\alpha
\]
Empirical or Plug-in Estimator

Given \( n \) samples \( X_1, \ldots, X_n \),

Let \( N_x \) denote the empirical frequency of \( x \).

\[
\hat{p}_n(x) \overset{\text{def}}{=} \frac{N_x}{n}
\]

\[
\hat{H}_\alpha^e \overset{\text{def}}{=} \frac{1}{1 - \alpha} \log \sum \hat{p}_n(x)^\alpha
\]

**Theorem**

*For \( \alpha > 1 \):*

\[
S_{\alpha}^{\hat{H}_\alpha^e} (\delta, \epsilon, k) \leq O \left( \frac{k}{\delta \max\{4, 1/(\alpha-1)\}} \log \frac{1}{\epsilon} \right)
\]

*For \( \alpha < 1 \):*

\[
S_{\alpha}^{\hat{H}_\alpha^e} (\delta, \epsilon, k) \leq O \left( \frac{k^{1/\alpha}}{\delta \max\{4, 2/\alpha\}} \log \frac{1}{\epsilon} \right)
\]
Empirical or Plug-in Estimator

Given $n$ samples $X_1, ..., X_n$,

Let $N_x$ denote the empirical frequency of $x$.

\[ \hat{p}_n(x) \overset{\text{def}}{=} \frac{N_x}{n} \]

\[ \hat{H}_\alpha^e \overset{\text{def}}{=} \frac{1}{1 - \alpha} \log \sum \hat{p}_n(x)^\alpha \]

Theorem

For $\alpha > 1$:

\[ S_{\alpha}^e \hat{H}_\alpha^e (\delta, \epsilon, k) \leq O \left( \frac{k}{\delta \max \{4, 1/(\alpha - 1)\}} \log \frac{1}{\epsilon} \right) \]

For $\alpha < 1$:

\[ S_{\alpha}^e \hat{H}_\alpha^e (\delta, \epsilon, k) \leq O \left( \frac{k^{1/\alpha}}{\delta \max \{4, 2/\alpha\}} \log \frac{1}{\epsilon} \right) \]

Proof??
Estimating Rényi entropy with small additive error is the same as estimating power sum with small multiplicative error.
Estimating Rényi entropy with small additive error is the same as estimating power sum with small multiplicative error.

Using a well-known sequence of steps, suffices to show that bias and variance of $\hat{p}_n$ are multiplicatively small.
The empirical frequencies $N_x$ are correlated.

Suppose $N \sim \text{Poi}(n)$ and $X_1, \ldots, X_N$ be independent samples from $p$.

Then,

1. $N_x \sim \text{Poi}(np_x)$

2. $\{N_x, x \in \mathcal{X}\}$ are mutually independent

3. For each estimator $\hat{H}$, there is a modified estimator $\hat{H}'$ such that

$$
P \left( |H_\alpha(p) - \hat{H}'(X^n)| > \delta \right) \leq P \left( |H_\alpha(p) - \hat{H}(X^N)| > \delta \right) + \frac{\epsilon}{2},
$$

where $N \sim \text{Poi}(n/2)$ and $n \geq 8 \log(2/\epsilon)$.

It suffices to bound the error probability under Poisson sampling.
Performance of the Empirical Estimator

For the empirical estimator \( \hat{p}_n \):

\[
\frac{1}{P_\alpha(p)} \left| \mathbb{E} \left[ \sum_x N_x \frac{X_x}{n^\alpha} \right] - P_\alpha(p) \right| \leq \begin{cases} 
  c_1 \max \left\{ \left( \frac{k}{n} \right)^{\alpha - 1}, \sqrt{\frac{k}{n}} \right\}, & \alpha > 1, \\
  c_2 \left( \frac{k^{1/\alpha}}{n} \right)^{\alpha}, & \alpha < 1 
\end{cases}
\]

\[
\frac{1}{P_\alpha(p)^2} \text{Var} \left[ \sum_x N_x \frac{X_x}{n^\alpha} \right] \leq \begin{cases} 
  c'_1 \max \left\{ \left( \frac{k}{n} \right)^{2\alpha - 1}, \sqrt{\frac{k}{n}} \right\}, & \alpha > 1, \\
  c'_2 \max \left\{ \left( \frac{k^{1/\alpha}}{n} \right)^{\alpha}, \sqrt{\frac{k}{n}}, \frac{1}{n^{2\alpha - 1}} \right\}, & \alpha < 1 
\end{cases}
\]

Theorem

For \( \alpha > 1 \):

\[
S_{\alpha}^{\hat{H}_e^c} (\delta, \epsilon, k) \leq O \left( \frac{k}{\delta_{\max\{4,1/(\alpha-1)\}} \log \frac{1}{\epsilon}} \right)
\]

For \( \alpha < 1 \):

\[
S_{\alpha}^{\hat{H}_e^c} (\delta, \epsilon, k) \leq O \left( \frac{k^{1/\alpha}}{\delta_{\max\{4,2/\alpha\}} \log \frac{1}{\epsilon}} \right)
\]
Consider an integer $\alpha > 1$

$$n^\alpha = n(n-1)\ldots(n-\alpha+1) = \alpha\text{th falling power of } n$$

**Claim:** For $X \sim \text{Poi}(\lambda)$, $\mathbb{E}[X^\alpha] = \lambda^\alpha$

Under Poisson sampling, an unbiased estimator of $P_\alpha(p)$ is

$$\hat{P}_n^u \overset{\text{def}}{=} \sum_x \frac{N_x^\alpha}{n^\alpha}$$

Our estimator for $H_\alpha(p)$ is

$$\hat{H}_n^u \overset{\text{def}}{=} \frac{1}{1-\alpha} \log \hat{P}_n^u$$
Performance of the Bias-Corrected Estimator

For the bias-corrected estimator $\hat{p}_n^u$ and an integer $\alpha > 1$

$$\frac{1}{P_\alpha(p)}^2 \text{Var}[\hat{p}_n^u] \leq \sum_{r=0}^{\alpha-1} \left( \frac{\alpha^2 k^{1-1/\alpha}}{n} \right)^{\alpha-r}$$

**Theorem**

*For integer $\alpha > 1$*

$$S_{\alpha}^{\hat{H}_n^u}(\delta, \epsilon, k) \leq O \left( \frac{k^{1-1/\alpha}}{\delta^2} \log \frac{1}{\epsilon} \right)$$
Performance of the Bias-Corrected Estimator

For the bias-corrected estimator \( \hat{p}_n^u \) and an integer \( \alpha > 1 \)

\[
\frac{1}{P_\alpha(p)^2} \text{Var}[\hat{p}_n^u] \leq \sum_{r=0}^{\alpha-1} \left( \frac{\alpha^2 k^{1-1/\alpha}}{n} \right)^{\alpha-r}
\]

Theorem

For integer \( \alpha > 1 \):

\[
S_{\alpha}^\hat{H}_n^u(\delta, \epsilon, k) \leq O \left( \frac{k^{1-1/\alpha}}{\delta^2} \log \frac{1}{\epsilon} \right)
\]

To summarize:

For every \( 0 < \alpha < 1 \): \( S_{\alpha}(k) \leq O(k^{1/\alpha}) \)

For every \( 1 < \alpha \notin \mathbb{N} \): \( S_{\alpha}(k) \leq O(k) \)

For every \( 1 < \alpha \in \mathbb{N} \): \( S_{\alpha}(k) \leq O(k^{1-1/\alpha}) \)
The rest of the paper is organized as follows. Section 1.6 Organization performance by reducing the bias in the empirical estimator.

The estimation algorithms are analyzed in Section 2.1 Bounds on power sums. In particular, for any $\alpha$, $1 < \alpha < 2$, we show results on the empirical bias-corrected estimator, in Section 3.2 Parameters of distributions and moments of Poisson random variables, which may be of independent interest.

For every $\alpha$, $1 < \alpha < 2$, and $k \geq 1$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$, using the bias-corrected estimator.

Considering a distribution $p$ over $\{0, 1\}$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$.

For every $\alpha$, $1 < \alpha < 2$, and $k \geq 1$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$.

Further, for $\alpha = 1$, $\alpha = 2$, and $\alpha > 2$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$.

For every $\alpha$, $1 < \alpha < 2$, and $k \geq 1$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$.

Consider a distribution $p$ over $\{0, 1\}$, we provide optimal results for integral $\alpha$ and non-integral $\alpha$. The case $\alpha = 1$ corresponds to the Shannon entropy of order 1. Finally, the lower bounds on the sample complexity $S$ can be bounded in terms of $H_{\alpha}(p)$ and $H_{\alpha-1}(p)$.
Lower Bounds
The General Approach

\( S_\alpha(\delta, \epsilon, k) \geq g(k) \) \text{ for all } \delta, \epsilon \text{ sufficiently small:}

Show that there exist two distributions \( p \) and \( q \) such that

1. Support-size of both \( p \) and \( q \) is \( k \);
2. \( |H_\alpha(p) - H_\alpha(q)| > \delta \);
3. For all \( n < g(k) \), the variation distance \( \|p^n - q^n\| \) is small.
The General Approach

\[ S_\alpha(\delta, \epsilon, k) \geq g(k) \] for all \( \delta, \epsilon \) sufficiently small:

Show that there exist two distributions \( p \) and \( q \) such that

1. Support-size of both \( p \) and \( q \) is \( k \);
2. \( |H_\alpha(p) - H_\alpha(q)| > \delta \); 
3. For all \( n < g(k) \), the variation distance \( \|p^n - q^n\| \) is small.

We can replace \( X^n \) with a sufficient statistic \( \psi(X^n) \) to replace (3) with:

For all \( n < g(k) \), the variation distance \( \|p_{\psi(X^n)} - q_{\psi(X^n)}\| \) is small.
Distance between Profile Distributions

**Definition.** Profile of $X^n$ refers $\Phi = (\Phi_1, ..., \Phi_n)$ where

$$\Phi_i = \text{number of symbols appearing } i \text{ times in } X^n$$

$$= \sum_x 1(N_x = i)$$

Two simple observations:

1. Profile is a sufficient statistic for the probability multiset of $p$

2. We can assume Poisson sampling without loss of generality

Let $p_\Phi$ and $q_\Phi$ denote the distribution of profiles under Poisson sampling
Distance between Profile Distributions

**Definition.** Profile of $X^n$ refers $\Phi = (\Phi_1, \ldots, \Phi_n)$ where

$$\Phi_i = \text{number of symbols appearing } i \text{ times in } X^n$$

$$= \sum_x 1(N_x = i)$$

Two simple observations:

1. Profile is a sufficient statistic for the probability multiset of $p$
2. We can assume Poisson sampling without loss of generality

Let $p_{\Phi}$ and $q_{\Phi}$ denote the distribution of profiles under Poisson sampling

**Theorem (Valiant08)**

Given distributions $p$ and $q$ such that $\max_x \max\{p_x; q_x\} \leq \frac{\epsilon}{40n}$, for Poisson sampling with $N \sim \text{Poi}(n)$, it holds that

$$\|p_{\Phi} - q_{\Phi}\| \leq \frac{\epsilon}{2} + 5 \sum a n^a |P_a(p) - P_a(q)|.$$
Derivation of our Lower Bounds

For distributions $p$ and $q$:

1. $\|p_\Phi - q_\Phi\| \lesssim 5 \sum_\alpha n^\alpha |P_\alpha(p) - P_\alpha(q)|$

2. $|H_\alpha(p) - H_\alpha(q)| = \frac{1}{1 - \alpha} \left| \log \frac{P_\alpha(p)}{P_\alpha(q)} \right|$

Choose $p$ and $q$ to be mixtures of $d$ uniform distributions as follows:

\[ p_{ij} = \frac{|x_i|}{k \|x\|_1}, \quad 1 \leq i \leq d, 1 \leq j \leq k \]

\[ q_{ij} = \frac{|y_i|}{k \|y\|_1}, \quad 1 \leq i \leq d, 1 \leq j \leq k \]
Derivation of our Lower Bounds

For distributions \( p \) and \( q \):

\[ \| p_\Phi - q_\Phi \| \lesssim 5 \sum_a n^a | P_a(p) - P_a(q) | \]

\[ | H_\alpha(p) - H_\alpha(q) | = \frac{1}{1 - \alpha} \left| \log \frac{P_\alpha(p)}{P_\alpha(q)} \right| \]

Choose \( p \) and \( q \) to be mixtures of \( d \) uniform distributions as follows:

\[ p_{ij} = \frac{|x_i|}{k \| x \|_1}, \quad 1 \leq i \leq d, 1 \leq j \leq k \]

\[ q_{ij} = \frac{|y_i|}{k \| y \|_1}, \quad 1 \leq i \leq d, 1 \leq j \leq k \]

Thus,

\[ \| p_\Phi - q_\Phi \| \lesssim 5 \sum_a \left( \frac{n}{k^{1 - 1/a}} \right)^a \left( \left( \frac{\| x \|_a}{\| x \|_1} \right)^a - \left( \frac{\| y \|_a}{\| y \|_1} \right)^a \right) \]

\[ | H_\alpha(p) - H_\alpha(q) | = \frac{\alpha}{(1 - \alpha) k^{\alpha - 1}} \left| \log \frac{\| x \|_\alpha}{\| y \|_\alpha} \cdot \frac{\| x \|_1}{\| y \|_1} \right| \]
Derivation of our Lower Bounds: Key Construction

Distributions with $\|x\|_r = \|y\|_r$, $\forall 1 \leq r \leq m - 1$ cannot be distinguished with fewer than $k^{1-1/m}$ samples

Distributions with $\|x\|_{\alpha} \neq \|y\|_{\alpha}$ have different $H_{\alpha}$
Distributions with $\|x\|_r = \|y\|_r$, $\forall 1 \leq r \leq m - 1$ cannot be distinguished with fewer than $k^{1-1/m}$ samples.

Distributions with $\|x\|_\alpha \neq \|y\|_\alpha$ have different $H_\alpha$.

**Lemma**

*For every $d \in \mathbb{N}$ and $\alpha$ not integer, there exist positive vectors $x, y \in \mathbb{R}^d$ such that*

$$
\|x\|_r = \|y\|_r, \quad 1 \leq r \leq d - 1, \\
\|x\|_d \neq \|y\|_d, \\
\|x\|_\alpha \neq \|y\|_\alpha.
$$
Derivation of our Lower Bounds: Key Construction

Lemma

For every \( d \in \mathbb{N} \) and \( \alpha \) not integer, there exist positive vectors \( x, y \in \mathbb{R}^d \) such that

\[
\|x\|_r = \|y\|_r, \quad 1 \leq r \leq d - 1, \\
\|x\|_d \neq \|y\|_d, \\
\|x\|_\alpha \neq \|y\|_\alpha.
\]
Derivation of our Lower Bounds: Key Construction

Lemma

For every $d \in \mathbb{N}$ and $\alpha$ not integer, there exist positive vectors $x, y \in \mathbb{R}^d$ such that

\[
\|x\|_r = \|y\|_r, \quad 1 \leq r \leq d - 1, \\
\|x\|_d \neq \|y\|_d, \\
\|x\|_\alpha \neq \|y\|_\alpha.
\]
Lemma

For every \( d \in \mathbb{N} \) and \( \alpha \) not integer, there exist positive vectors \( x, y \in \mathbb{R}^d \) such that

\[
\|x\|_r = \|y\|_r, \quad 1 \leq r \leq d - 1,
\]

\[
\|x\|_d \neq \|y\|_d,
\]

\[
\|x\|_\alpha \neq \|y\|_\alpha.
\]
In Closing ...
Rényi entropy of order 2 is the “easiest” entropy to estimate, requiring only $O(\sqrt{k})$ samples
Rényi entropy of order 2 is the “easiest” entropy to estimate, requiring only $O(\sqrt{k})$ samples

Sample complexity of estimating other information measures