Probabilities, Rényi Entropy and Guessing

Igal Sason
(EE Department, Technion)

Conference Prague Stochastics 2019

Workshop in Memory of František Matúš

Prague, Czechia

August 19–23, 2019

Probabilities, Rényi Entropy and Guessing

Igal Sason
(EE Department, Technion)

Conference Prague Stochastics 2019

Workshop in Memory of František Matúš

Prague, Czechia

August 19–23, 2019

Motivation

- Majorization: a simple & productive concept in the theory of inequalities.

Motivation

- Majorization: a simple & productive concept in the theory of inequalities.
- Cicalese *et al.* (IEEE T-IT, April '18):

  If X is a RV taking n possible values, and the support of $f(X)$ is equal to m with $m < n$, how close $H\big(f(X)\big)$ can be to $H(X)$?

Motivation

- Majorization: a simple & productive concept in the theory of inequalities.
- Cicalese *et al.* (IEEE T-IT, April '18):

  If X is a RV taking n possible values, and the support of $f(X)$ is equal to m with $m < n$, how close $H\big(f(X)\big)$ can be to $H(X)$?

- Their goal: computing

$$\max_f H\big(f(X)\big) = \max_f \Big\{ H\big(f(X)\big) - H\big(f(X)|X\big) \Big\} = \max_f I\big(X; f(X)\big)$$

  with max. over all functions mapping a set of cardinality n to a set of cardinality $m < n$.

Motivation

- Majorization: a simple & productive concept in the theory of inequalities.
- Cicalese *et al.* (IEEE T-IT, April '18):

  If X is a RV taking n possible values, and the support of $f(X)$ is equal to m with $m < n$, how close $H\big(f(X)\big)$ can be to $H(X)$?

- Their goal: computing

$$\max_f H\big(f(X)\big) = \max_f \Big\{ H\big(f(X)\big) - H\big(f(X)|X\big) \Big\} = \max_f I\big(X; f(X)\big)$$

  with max. over all functions mapping a set of cardinality n to a set of cardinality $m < n$.

- Useful in the context of data clustering.

Motivation (Cont.)

- Rényi measures are very useful, so how about ...

  Generalizing this question to $H_\alpha\big(f(X)\big)$ for any $\alpha > 0$ (not trivial).

Motivation (Cont.)

- Rényi measures are very useful, so how about ...

  Generalizing this question to $H_\alpha\big(f(X)\big)$ for any $\alpha > 0$ (not trivial).

- Possible Applications:
  - Guessing (Arikan '96);
  - Lossless compression problems (Campbell '65).

The Rényi Entropy

Let P_X be a probability distribution on a discrete set $\mathcal{X}$. The Rényi entropy of order $\alpha \in (0,1) \cup (1,\infty)$ of X is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X^\alpha(x)$$

By its continuous extension, $H_1(X) = H(X)$ where

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$$

is the Shannon entropy.

Setting

Let

- $\alpha > 0$;
- $\mathcal{X}$ and $\mathcal{Y}$ be finite sets of cardinalities

$$|\mathcal{X}| = n, \quad |\mathcal{Y}| = m, \quad n > m \geq 2;$$

without any loss of generality, let

$$\mathcal{X} = \{1, \ldots, n\}, \quad \mathcal{Y} = \{1, \ldots, m\};$$

- $\mathcal{P}_n$ $(n \geq 2)$ be the set of probability mass functions (pmf) on $\mathcal{X}$;
- X be a RV taking values on $\mathcal{X}$ with a pmf $P_X \in \mathcal{P}_n$;
- $\mathcal{F}_{n,m}$ be the set of deterministic functions $f \colon \mathcal{X} \to \mathcal{Y}$;
- $f \in \mathcal{F}_{n,m}$ is not one-to-one since $m < n$.

Bad News

For an arbitrary $\alpha > 0$, the maximization problem

$$\max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) \qquad (2 \leq m < n)$$

is strongly NP-hard.

- Unless P = NP, there is no poly. time algorithm which, for any $\varepsilon > 0$, computes an admissible deterministic function $f_\varepsilon \in \mathcal{F}_{n,m}$ such that

$$H_\alpha\big(f_\varepsilon(X)\big) \geq (1-\varepsilon) \max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big).$$

Good News

We can efficiently construct (by the use of Huffman algorithm) an admissible function $f^* \in \mathcal{F}_{n,m}$ s.t.

$$H_\alpha\big(f^*(X)\big) \geq \max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) - v(\alpha), \quad \alpha > 0$$

where

$$v(\alpha) := \begin{cases} \log\left(\dfrac{\alpha-1}{2^\alpha - 2}\right) - \dfrac{\alpha}{\alpha-1}\log\left(\dfrac{\alpha}{2^\alpha-1}\right), & \alpha \neq 1, \\ \log\left(\dfrac{2}{e\,\ln 2}\right) \approx 0.08607 \text{ bits}, & \alpha = 1. \end{cases}$$

$v\colon (0,\infty) \to (0, \log 2)$ is monotonically increasing, continuous, and

$$\lim_{\alpha \downarrow 0} v(\alpha) = 0, \qquad \lim_{\alpha \to \infty} v(\alpha) = \log 2 \ (1 \text{ bit}).$$

Plot

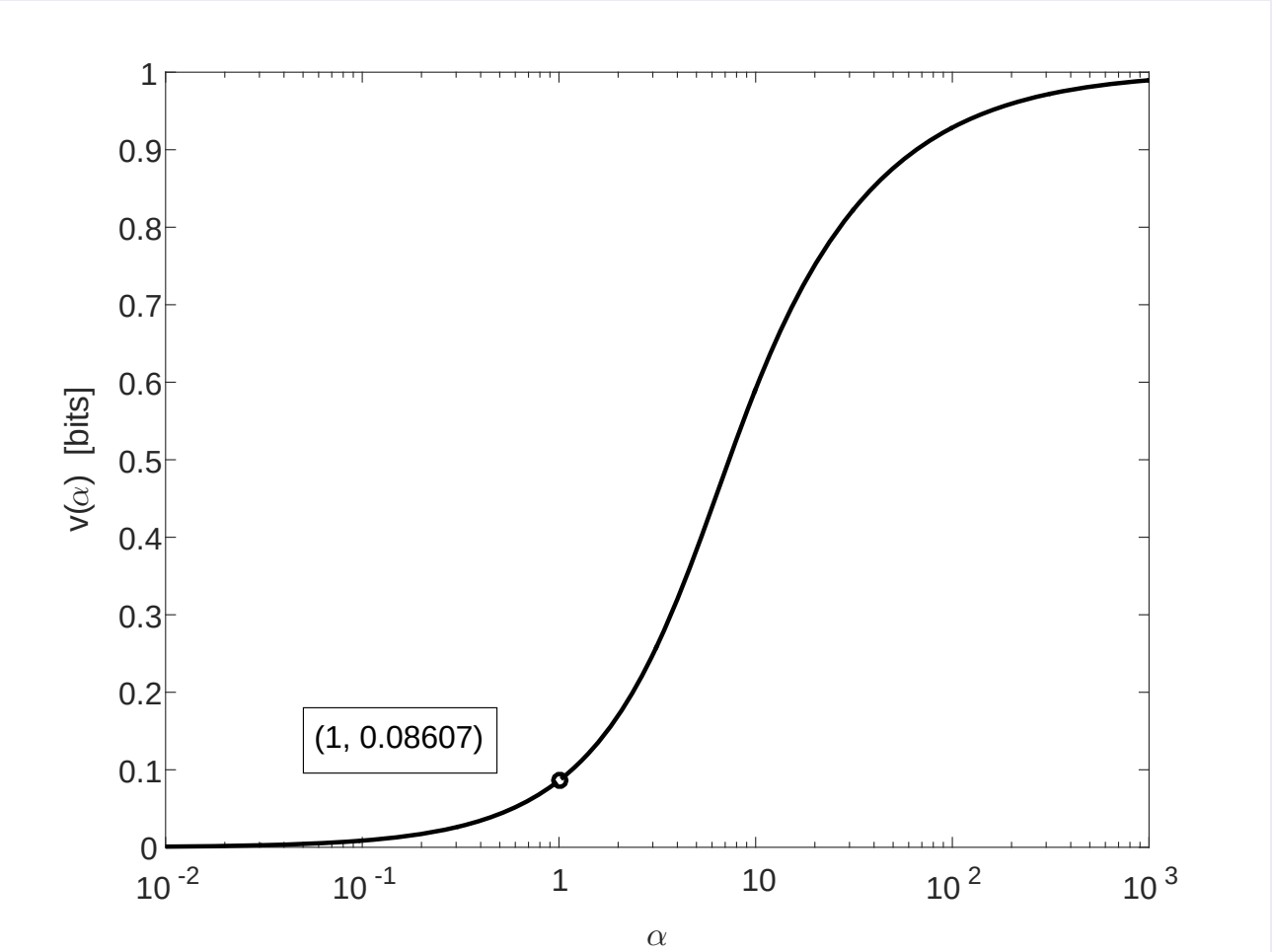


Figure: A plot of $v(\alpha)$ as a function of $\alpha > 0$.

Upper Bounding $\max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big)$

Let

- X be a discrete RV with pmf P_X, which takes n possible values, and assume that
$$P_X(1) \geq P_X(2) \geq \ldots \geq P_X(n).$$
- $f \in \mathcal{F}_{n,m}$;
- Q_X be the pmf of $f(X)$; assume that

$$Q_X(1) \geq P_X(2) \geq \ldots \geq Q_X(m),$$
$$Q_X(m+1) = \ldots = Q_X(n) = 0.$$

Then, P_X is majorized by Q_X:

$$P_X \prec Q_X \quad \left(\sum_{i=1}^{k} P_X(i) \leq \sum_{i=1}^{k} Q_X(i), \ \forall\, k \in \{1, \ldots, n\} \right).$$

Upper Bounding $\max_{f \in \mathcal{F}_{n,m}} H_\alpha\big(f(X)\big)$ (Cont.)

Since the Rényi entropy is Schur-concave, it follows that

$$H_\alpha(X) = H_\alpha(P_X) \geq H_\alpha(Q_X) = H_\alpha\big(f(X)\big).$$

Upper Bounding $\max\limits_{f\in\mathcal{F}_{n,m}} H_\alpha\big(f(X)\big)$ (Cont.)

Since the Rényi entropy is Schur-concave, it follows that

$$H_\alpha(X) = H_\alpha(P_X) \geq H_\alpha(Q_X) = H_\alpha\big(f(X)\big).$$

Let $\mathcal{P}_n$ $(n \geq 2)$ be the set of prob. mass functions (pmfs) on $\mathcal{X}$ $(|\mathcal{X}| = n)$. Since the pmf of $f(X)$ majorizes the pmf of X, then

$$\max_{f\in\mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) \leq \max_{Q\in\mathcal{P}_m : P_X \prec Q} H_\alpha(Q).$$

Upper Bounding $\max\limits_{f\in\mathcal{F}_{n,m}} H_\alpha\big(f(X)\big)$ (Cont.)

Since the Rényi entropy is Schur-concave, it follows that

$$H_\alpha(X) = H_\alpha(P_X) \geq H_\alpha(Q_X) = H_\alpha\big(f(X)\big).$$

Let $\mathcal{P}_n$ $(n \geq 2)$ be the set of prob. mass functions (pmfs) on $\mathcal{X}$ $(|\mathcal{X}| = n)$. Since the pmf of $f(X)$ majorizes the pmf of X, then

$$\max_{f\in\mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) \leq \max_{Q\in\mathcal{P}_m : P_X \prec Q} H_\alpha(Q).$$

- The max. in the right side of the last ineq. is explicitly calculated.
- A function $f^* \in \mathcal{F}_{n,m}$ is efficiently constructed such that

$$H_\alpha\big(f^*(X)\big) \geq \max_{Q\in\mathcal{P}_m : P_X \prec Q} H_\alpha(Q) - v(\alpha) \geq \max_{f\in\mathcal{F}_{n,m}} H_\alpha\big(f(X)\big) - v(\alpha).$$

Solving the Maximum Rényi Entropy Problem

$$\max_{Q \in \mathcal{P}_m : P_X \prec Q} H_\alpha(Q)$$

with $X \in \{1, \ldots, n\}$, $m < n$, and $\alpha > 0$.

Solution: $R_m(P_X)$

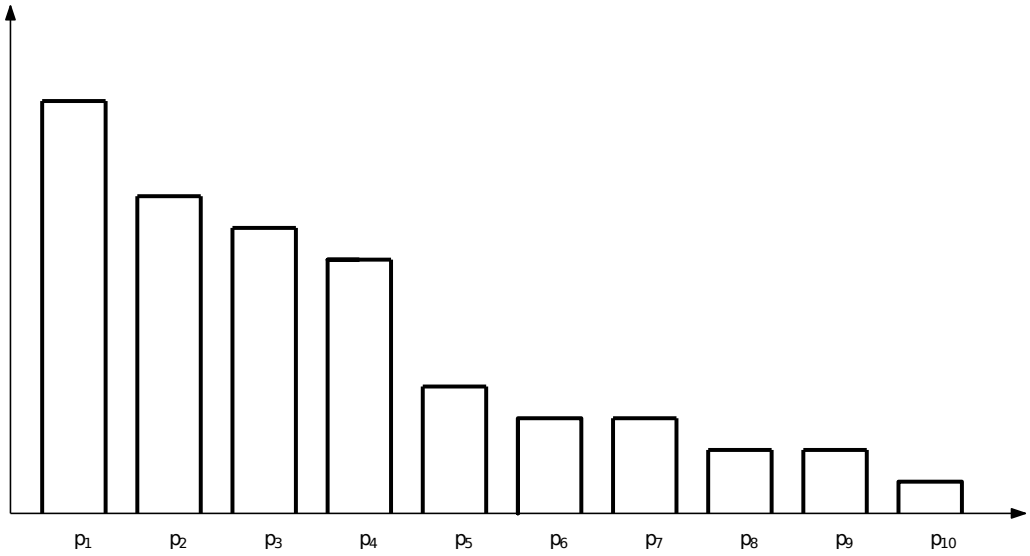
- If $P_X(1) < \frac{1}{m}$, then $R_m(P_X)$ is the equiprobable dist. on $\{1, \ldots, m\}$;
- Otherwise, $R_m(P_X) := Q_X \in \mathcal{P}_m$ with

$$Q_X(i) = \begin{cases} P_X(i), & i \in \{1, \ldots, n^*\}, \\ \dfrac{1}{m - n^*} \displaystyle\sum_{j=n^*+1}^{n} P_X(j), & i \in \{n^*+1, \ldots, m\}, \end{cases}$$

where n^* is the max. integer i s.t. $P_X(i) \geq \frac{1}{m-i} \sum_{j=i+1}^{n} P_X(j)$.

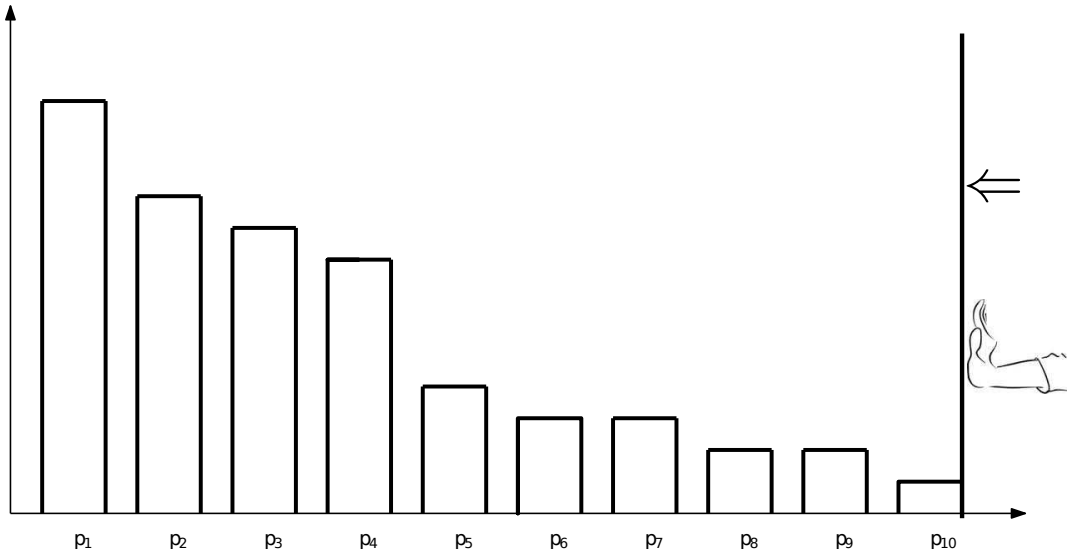
Intuitively

p



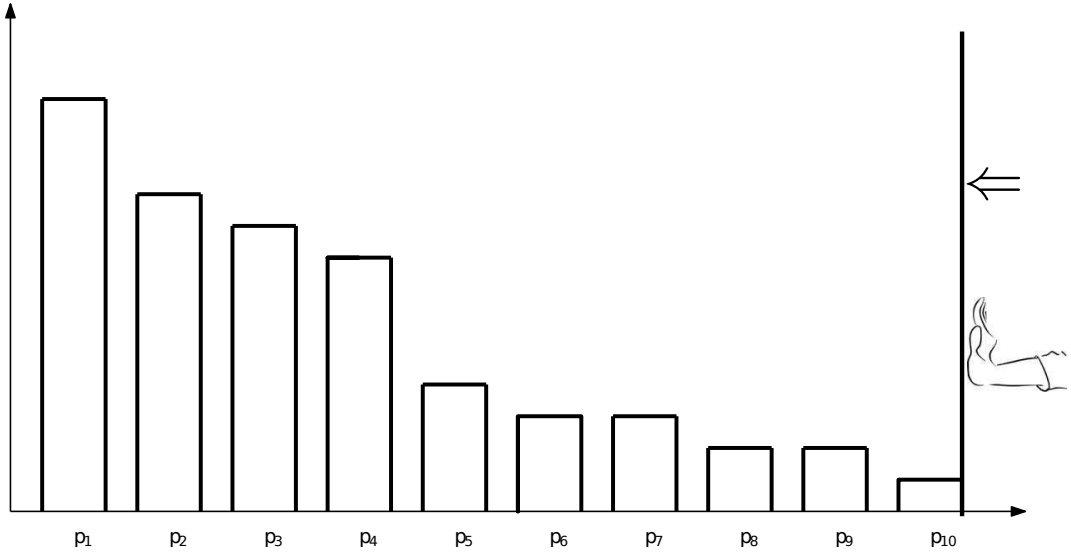
Intuitively

p

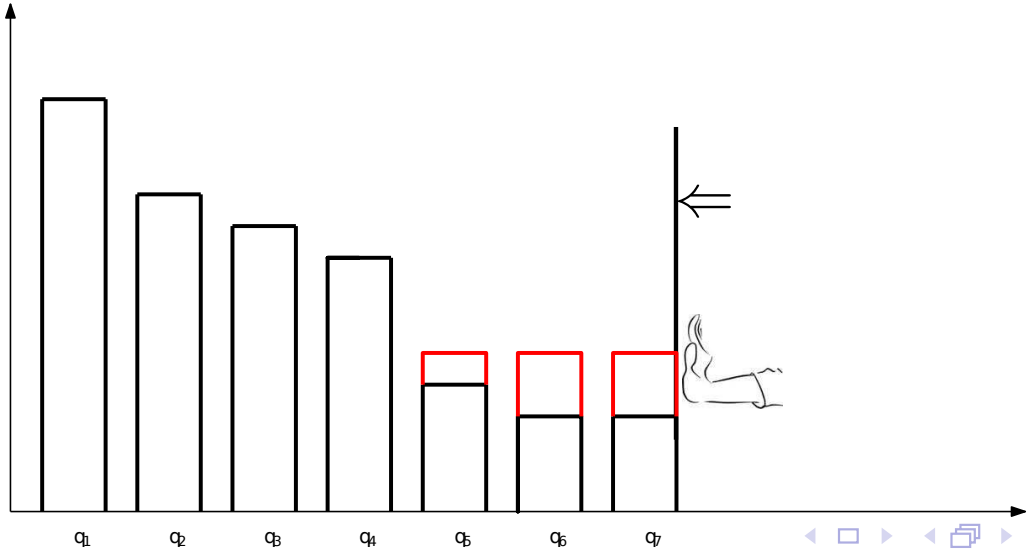


Intuitively

$\mathbf{p}$



$R(\mathbf{p})$



The Algorithm by Huffman Coding

1. Start from the pmf $P_X \in \mathcal{P}_n$ with $P_X(1) \geq \ldots \geq P_X(n)$;
2. Merge successively pairs of probability masses by applying the Huffman algorithm;
3. Stop the process in Step 2 when a probability mass function $Q \in \mathcal{P}_m$ is obtained (with $Q(1) \geq \ldots \geq Q(m)$);
4. Construct the deterministic function $f^* \in \mathcal{F}_{n,m}$ by setting $f^*(k) = j \in \{1, \ldots, m\}$ for all probability masses $P_X(k)$, with $k \in \{1, \ldots, n\}$, being merged in Steps 2–3 into the node of $Q(j)$.

Property - Link to the Next Auxiliary Result

Let $i \in \{0, \ldots, m-1\}$ the max. index s.t. $P_X(j) = Q(j)$ for all $j \leq i$. Then,

$$Q(i+1) \leq 2\, Q(m).$$

Notation for an Auxiliary Result

Let

- P be a probability mass function defined on a finite set $\mathcal{X}$,
- $p_{\max}$ and $p_{\min}$ be the max. and min. positive masses of P,
- $G_P(k)$ be the sum of the k largest masses of P for $k \in \{1, \ldots, |\mathcal{X}|\}$,
- $\mathcal{P}_n(\rho)$, for $\rho \geq 1$ and integer $n \geq 2$, be the subset of all $P \in \mathcal{P}_n$ s.t.

$$\frac{p_{\max}}{p_{\min}} \leq \rho.$$

Theorem (an Auxiliary Result)

For $\rho > 1$ and $\alpha > 0$, let

$$c_\alpha^{(n)}(\rho) := \log n - \min_{P \in \mathcal{P}_n(\rho)} H_\alpha(P), \quad n = 2, 3, \ldots$$

with $c_\alpha^{(1)}(\rho) := 0$. Then, for every $n \in \mathbb{N}$,

$$0 \le c_\alpha^{(n)}(\rho) \le \log \rho,$$
$$c_\alpha^{(n)}(\rho) \le c_\alpha^{(2n)}(\rho),$$

and $c_\alpha^{(n)}(\rho)$ is monotonically increasing in $\alpha \in [0, \infty]$.

The following limit exists for all $\alpha > 0$ and $\rho > 1$:

$$c_\alpha^{(\infty)}(\rho) := \lim_{n \to \infty} c_\alpha^{(n)}(\rho).$$

Theorem (Cont.)

Let $\rho > 1$, $\alpha > 0$, $n \geq 2$ be an integer, and $\beta \in \left[\frac{1}{1+(n-1)\rho}, \frac{1}{n}\right] := \Gamma_\rho^{(n)}$.

Let $Q_\beta \in \mathcal{P}_n(\rho)$ be defined on $\mathcal{X} = \{1, \ldots, n\}$ as follows:

$$Q_\beta(j) = \begin{cases} \rho\beta, & j \in \{1, \ldots, i_\beta\}, \\ 1 - (n + i_\beta\,\rho - i_\beta - 1)\beta, & j = i_\beta + 1, \\ \beta, & j \in \{i_\beta + 2, \ldots, n\} \end{cases}$$

where $i_\beta := \left\lfloor \frac{1-n\beta}{(\rho-1)\beta} \right\rfloor$. Then, for every $\alpha > 0$,

$$c_\alpha^{(n)}(\rho) = \log n - \min_{\beta \in \Gamma_\rho^{(n)}} H_\alpha(Q_\beta).$$

Theorem (Cont.)

1. If $\alpha \in (0,1) \cup (1,\infty)$, then

$$c_\alpha^{(\infty)}(\rho) = \frac{1}{\alpha-1}\log\left(1+\frac{1+\alpha\,(\rho-1)-\rho^\alpha}{(1-\alpha)(\rho-1)}\right) - \frac{\alpha}{\alpha-1}\log\left(1+\frac{1+\alpha\,(\rho-1)-\rho^\alpha}{(1-\alpha)(\rho^\alpha-1)}\right),$$

and

$$\lim_{\alpha\to\infty} c_\alpha^{(\infty)}(\rho) = \log\rho, \quad v(\alpha) = c_\alpha^{(\infty)}(2)$$

2. If $\alpha = 1$, then

$$c_1^{(\infty)}(\rho) = \lim_{\alpha\to 1} c_\alpha^{(\infty)}(\rho) = \frac{\rho\log\rho}{\rho-1} - \log\left(\frac{\mathrm{e}\rho\log_{\mathrm{e}}\rho}{\rho-1}\right).$$

Plot

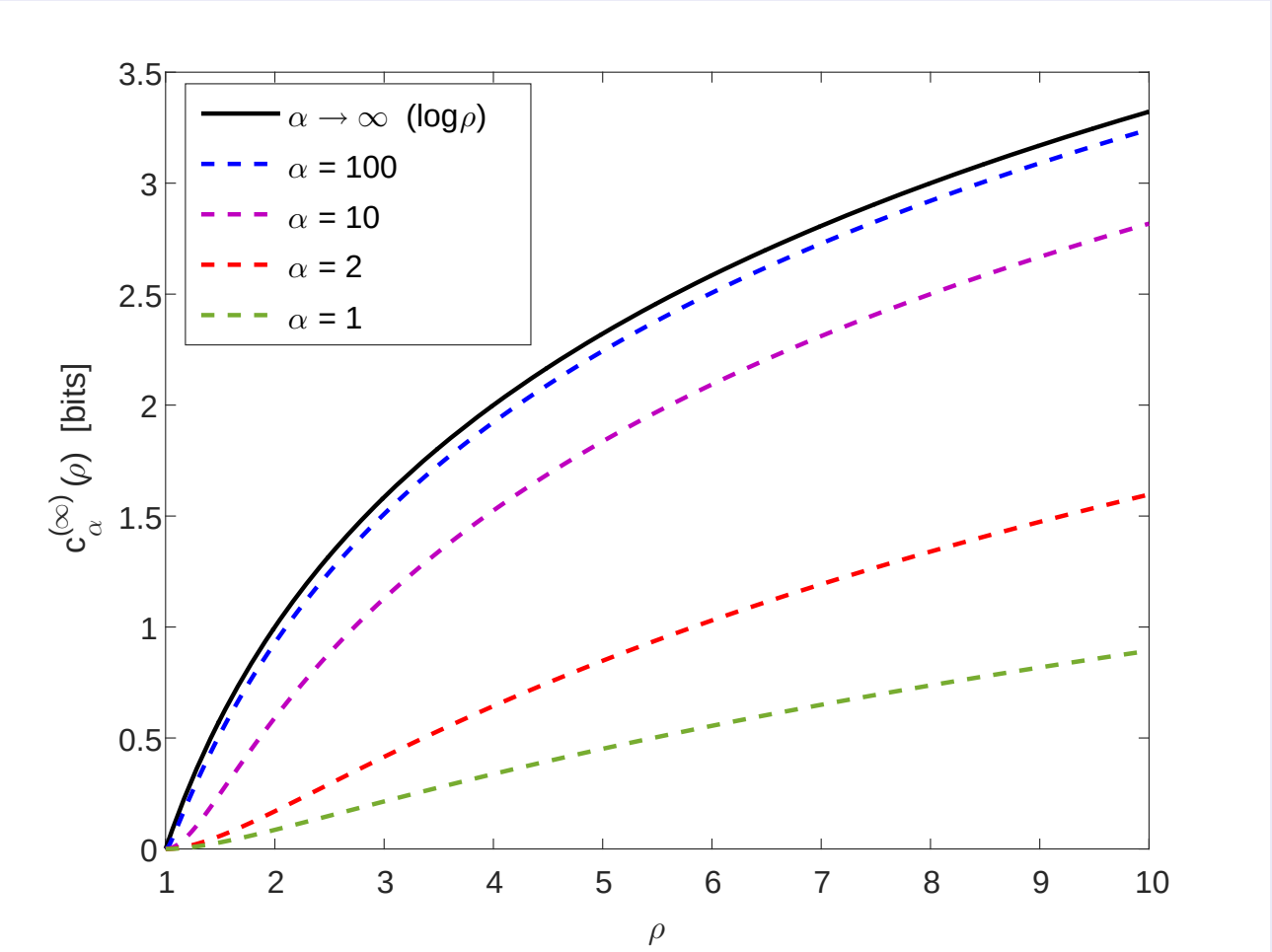


Figure: A plot of $c_\alpha^{(\infty)}(\rho)$ as a function of ρ.

Plot

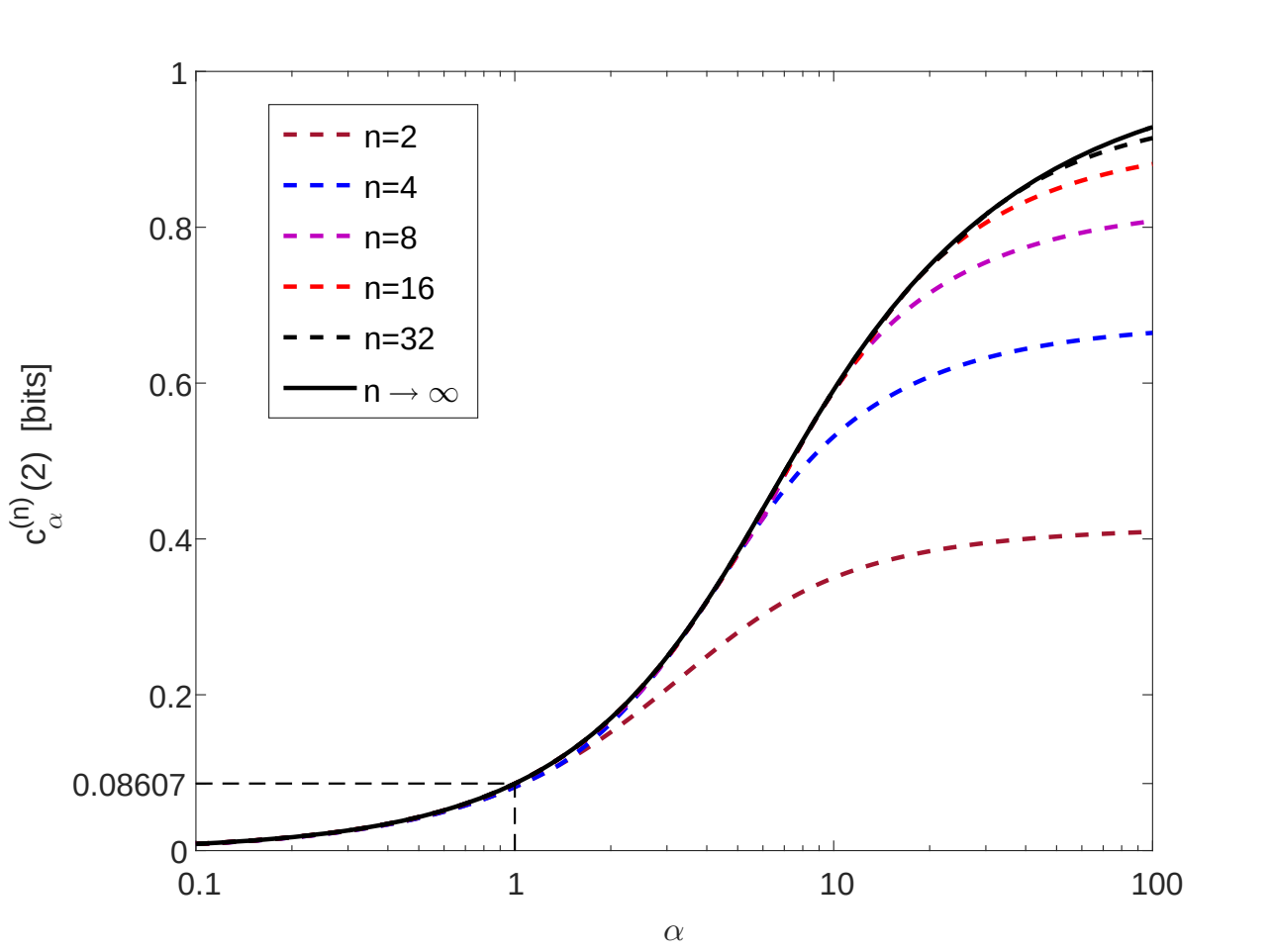


Figure: A plot of $c_\alpha^{(n)}(2)$ ($\log$ as a function of $\alpha > 0$, for several values of $n \geq 2$.

Guessing

The problem of guessing discrete random variables has found a variety of applications in

- Shannon theory,
- coding theory,
- cryptography,
- searching and sorting algorithms,

etc.

The central object of interest:

The distribution of the number of guesses required to identify a realization of a random variable, taking values on a finite or countably infinite set.

Guessing and Ranking functions

- X is a discrete random variable taking values on a finite or countably infinite set $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$.

Guessing and Ranking functions

- X is a discrete random variable taking values on a finite or countably infinite set $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$.
- One wishes to guess the value of X by repeatedly asking questions of the form "Is X equal to x ?" until X is guessed correctly.

Guessing and Ranking functions

- X is a discrete random variable taking values on a finite or countably infinite set $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$.
- One wishes to guess the value of X by repeatedly asking questions of the form "Is X equal to x ?" until X is guessed correctly.
- A guessing function is a 1-to-1 function $g\colon \mathcal{X} \to \mathcal{X}$ where the number of guesses is equal to $g(x)$ if $X = x \in \mathcal{X}$.

Guessing and Ranking functions

- X is a discrete random variable taking values on a finite or countably infinite set $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$.
- One wishes to guess the value of X by repeatedly asking questions of the form "Is X equal to x ?" until X is guessed correctly.
- A guessing function is a 1-to-1 function $g\colon \mathcal{X} \to \mathcal{X}$ where the number of guesses is equal to $g(x)$ if $X = x \in \mathcal{X}$.
- For $\rho > 0$, $\mathbb{E}[g^{\rho}(X)]$ is minimized by selecting g to be a ranking function g_X, for which $g_X(x) = k$ if $P_X(x)$ is the k-th largest mass.

$H_\alpha(X)$ and Guessing Moments

Theorem (Arikan '96)

Let X be a discrete random variable taking values on $\mathcal{X} = \{1, \ldots, M\}$. Let $g_X(\cdot)$ be a ranking function of X. Then, for $\rho > 0$,

$$\frac{1}{\rho} \log \mathbb{E}\left[g_X^\rho(X)\right] \geq H_{\frac{1}{1+\rho}}(X) - \log(1 + \log_{\mathrm{e}} M),$$

$$\frac{1}{\rho} \log \mathbb{E}\left[g_X^\rho(X)\right] \leq H_{\frac{1}{1+\rho}}(X).$$

$H_\alpha(X)$ and Guessing Moments

Theorem (Arikan '96)

Let X be a discrete random variable taking values on $\mathcal{X} = \{1, \ldots, M\}$. Let $g_X(\cdot)$ be a ranking function of X. Then, for $\rho > 0$,

$$\frac{1}{\rho} \log \mathbb{E}\big[g_X^\rho(X)\big] \geq H_{\frac{1}{1+\rho}}(X) - \log(1 + \log_{\mathrm{e}} M),$$

$$\frac{1}{\rho} \log \mathbb{E}\big[g_X^\rho(X)\big] \leq H_{\frac{1}{1+\rho}}(X).$$

Arikan's result yields an asymptotically tight error exponent:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[g_{X^n}^\rho(X^n)\big] = \rho H_{\frac{1}{1+\rho}}(X), \quad \forall\, \rho > 0$$

when $X_1, \ldots, X_n$ are i.i.d. $\quad [X^n := (X_1, \ldots, X_n)]$.

Reminder: $R_m(P_X)$

Let $X \sim P_X$ take n possible values, and let $m \in \{2, \ldots, n-1\}$.

- If $P_X(1) < \frac{1}{m}$, then $R_m(P_X)$ is the equiprobable dist. on $\{1, \ldots, m\}$;
- Otherwise, $R_m(P_X) := Q_X \in \mathcal{P}_m$ with

$$Q_X(i) = \begin{cases} P_X(i), & i \in \{1, \ldots, n^*\}, \\ \dfrac{1}{m-n^*} \displaystyle\sum_{j=n^*+1}^{n} P_X(j), & i \in \{n^*+1, \ldots, m\}, \end{cases}$$

where n^* is the max. integer i s.t. $P_X(i) \geq \frac{1}{m-i} \sum_{j=i+1}^{n} P_X(j)$.

Notation

for $m \in \{2, \ldots, n\}$, let

$$X_m \sim R_m(P_X).$$

Theorem: Guessing Moments

Let

- $\{X_i\}_{i=1}^k$ be i.i.d. with $X_1 \sim P_X$ taking values on a set $\mathcal{X}$, $|\mathcal{X}| = n$;
- $Y_i = f(X_i)$, for every $i \in \{1, \ldots, k\}$, where $f \in \mathcal{F}_{n,m}$ is a deterministic function with $m < n$;
- $$g_{X^k} \colon \mathcal{X}^k \to \{1, \ldots, n^k\}, \quad g_{Y^k} \colon \mathcal{Y}^k \to \{1, \ldots, m^k\}$$

  be, respectively, ranking functions of the random vectors

$$X^k := (X_1, \ldots, X_k), \quad Y^k := (Y_1, \ldots, Y_k).$$

Theorem: Guessing Moments (Cont.)

Then, for every $\rho > 0$,

1. For every deterministic function $f \in \mathcal{F}_{n,m}$

$$\frac{1}{k} \log \frac{\mathbb{E}\left[g^{\rho}_{X^k}(X^k)\right]}{\mathbb{E}\left[g^{\rho}_{Y^k}(Y^k)\right]} \geq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right] - \frac{\rho \log(1 + k \ln n)}{k}.$$

Theorem: Guessing Moments (Cont.)

Then, for every $\rho > 0$,

1. For every deterministic function $f \in \mathcal{F}_{n,m}$

$$\frac{1}{k} \log \frac{\mathbb{E}\left[g^{\rho}_{X^k}(X^k)\right]}{\mathbb{E}\left[g^{\rho}_{Y^k}(Y^k)\right]} \geq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right] - \frac{\rho \log(1 + k \ln n)}{k}.$$

2. For the deterministic function $f^* \in \mathcal{F}_{n,m}$, whose construction relies on the Huffman algorithm, with $Y_i = f^*(X_i)$ for all $i \in \{1, \ldots, k\}$,

$$\frac{1}{k} \log \frac{\mathbb{E}\left[g^{\rho}_{X^k}(X^k)\right]}{\mathbb{E}\left[g^{\rho}_{Y^k}(Y^k)\right]}$$
$$\leq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) + v\left(\frac{1}{1+\rho}\right) \right] + \frac{\rho \log(1 + k \ln m)}{k}.$$

Theorem: Guessing Moments (Cont.)

For every $\rho > 0$,

3) The gap between the universal lower bound and the upper bound, for $f = f^*$, is at most

$$\rho\, v\left(\frac{1}{1+\rho}\right) + \frac{2\rho \log(1 + k \log_{\mathrm{e}} n)}{k}$$

$$\approx \frac{0.08607\,\rho}{1+\rho} + O\left(\frac{\log k}{k}\right) \text{ bits.}$$

Letting $k \to \infty$, the gap is less than 0.08607 bits for all $\rho > 0$, and the construction of the function $f^* \in \mathcal{F}_{n,m}$ does not depend on ρ.

Application II:

Non-Asymptotic Bounds for Optimal Fixed-to-Variable Lossless Compression Codes

We rely on Campbell's work (1965), providing bounds on the cumulant generating function which are expressed in terms of Rényi entropies.

The paper refers to lossless data compression. However, due to time constraints, this talk needs to be a lossy compression of the journal paper, and I'll skip the details here.

Journal Paper

I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression," *Entropy*, vol. 20, no. 12, paper 896, pp. 1–25, November 2018.

Summary

- Tight bounds on the Rényi entropy of a function of a random variable with finite number of possible values where the function is not 1-to-1, and the cardinality of the image of the function is fixed.

Summary

- Tight bounds on the Rényi entropy of a function of a random variable with finite number of possible values where the function is not 1-to-1, and the cardinality of the image of the function is fixed.
- A tight lower bound on the Rényi entropy of a discrete random variable with a finite support is derived as a function of the size of the support, and the ratio of the maximal to minimal probability masses.

Summary

- Tight bounds on the Rényi entropy of a function of a random variable with finite number of possible values where the function is not 1-to-1, and the cardinality of the image of the function is fixed.
- A tight lower bound on the Rényi entropy of a discrete random variable with a finite support is derived as a function of the size of the support, and the ratio of the maximal to minimal probability masses.
- Inspired by the recent work by Cicalese *et al.*, which was focused on the Shannon entropy. It strengthens and generalizes the results of that paper to Rényi entropies of arbitrary positive orders.

Summary

- Tight bounds on the Rényi entropy of a function of a random variable with finite number of possible values where the function is not 1-to-1, and the cardinality of the image of the function is fixed.
- A tight lower bound on the Rényi entropy of a discrete random variable with a finite support is derived as a function of the size of the support, and the ratio of the maximal to minimal probability masses.
- Inspired by the recent work by Cicalese *et al.*, which was focused on the Shannon entropy. It strengthens and generalizes the results of that paper to Rényi entropies of arbitrary positive orders.
- In view of the generalized bounds and the works by Arikan and Campbell, non-asymptotic bounds are derived for guessing moments and lossless data compression of discrete memoryless sources.