Tight Bounds on the Rényi Entropy via Majorization with an Application to Guessing

Igal Sason
(EE Department, Technion, Israel)

2020 Information Theory and Applications Workshop

San-Diego, California, USA

February 2–7, 2020

Motivation

- 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

- 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

- 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.)

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

Motivation (Cont.)

- Generalizing this question to $H_\alpha(f(X))$ for any $\alpha > 0$ (not trivial).
- Possible Applications to the Rényi Entropy of order α:
  - Guessing (Arikan '96);
  - Lossless compression problems (Campbell '65).

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

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

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

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

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

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

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

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

Plot

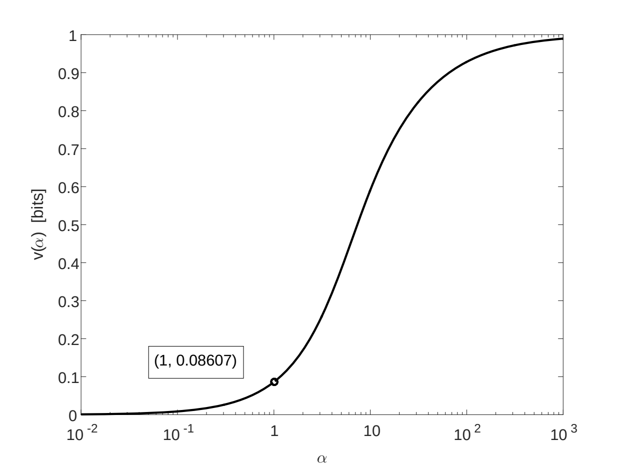

Figure: A plot of $v(\alpha)$ as a function of $\alpha > 0$.

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 $Q(j)$.

A Maximum Rényi Entropy Problem

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

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

Solution: $Q = R_m(P_X)$

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

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

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

Application I: Guessing and Ranking functions

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

Application I: 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.

Application I: Guessing and Ranking functions

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

Application I: 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 $\rho > 0$,

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

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

$H_\alpha(X)$ 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. $\quad [X^n := (X_1, \ldots, X_n)]$.

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

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

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

Notation

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

$$\widetilde{X}_m \sim R_m(P_X).$$

Theorem: Guessing Moments (Cont.)

1. The mutual information (in bits) satisfies

$$\max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big) - 0.08607 \le I\big(X;\, f^*(X)\big) \le \max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big).$$

Theorem: Guessing Moments (Cont.)

1. The mutual information (in bits) satisfies

$$\max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big) - 0.08607 \leq I\big(X;\, f^*(X)\big) \leq \max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big).$$

2. For every deterministic function $f \in \mathcal{F}_{n,m}$, for all $\rho > 0$,

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

Theorem: Guessing Moments (Cont.)

1. The mutual information (in bits) satisfies

$$\max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big) - 0.08607 \leq I\big(X;\, f^*(X)\big) \leq \max_{f \in \mathcal{F}_{n,m}} I\big(X;\, f(X)\big).$$

2. For every deterministic function $f \in \mathcal{F}_{n,m}$, for all $\rho > 0$,

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

3. For $f^* \in \mathcal{F}_{n,m}$, with $Y_i = f^*(X_i)$ for all $i \in \{1, \dots, k\}$, for all $\rho > 0$,

$$\frac{1}{k} \log \frac{\mathbb{E}\big[g^{\rho}_{X^k}(X^k)\big]}{\mathbb{E}\big[g^{\rho}_{Y^k}(Y^k)\big]} \leq \rho \left[H_{\frac{1}{1+\rho}}(X) - H_{\frac{1}{1+\rho}}(\widetilde{X}_m) \right] + \frac{0.08607\, \rho}{1 + \rho} + \frac{\rho \log(1 + k \ln m)}{k}.$$

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.

Journal Paper

I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression," *Entropy*, vol. 20, paper 896, pp. 1–25, November 2018.

Follow-up Journal Paper

I. S., "On data-processing and majorization inequalities for f-divergences," *Entropy*, vol. 21, paper 1022, pp. 1–80, October 2019.

To be presented in part at IZS '20.