Arimoto-Rényi Conditional Entropy and Bayesian M-ary Hypothesis Testing

Igal Sason (Technion) Sergio Verdú (Princeton)

2017 IEEE International Symposium on Information Theory

Aachen, Germany
June 25–30, 2017

Hypothesis Testing

- Bayesian M-ary hypothesis testing:
  - X is a random variable taking values on $\mathcal{X}$ with $|\mathcal{X}| = M$;
  - a prior distribution P_X on $\mathcal{X}$;
  - M hypotheses for the $\mathcal{Y}$-valued data $\{P_{Y|X=m}, m \in \mathcal{X}\}$.

Hypothesis Testing

- Bayesian M-ary hypothesis testing:
  - X is a random variable taking values on $\mathcal{X}$ with $|\mathcal{X}| = M$;
  - a prior distribution P_X on $\mathcal{X}$;
  - M hypotheses for the $\mathcal{Y}$-valued data $\{P_{Y|X=m}, m \in \mathcal{X}\}$.
- $\varepsilon_{X|Y}$: the minimum probability of error of X given Y
  - achieved by the *maximum-a-posteriori* (MAP) decision rule.

Interplay $\varepsilon_{X|Y} \longleftrightarrow$ information measures

- Bounds on $\varepsilon_{X|Y}$ involving information measures exist in the literature.

Interplay $\varepsilon_{X|Y} \longleftrightarrow$ information measures

- Bounds on $\varepsilon_{X|Y}$ involving information measures exist in the literature.
- Useful for
  - the analysis of M-ary hypothesis testing
  - proofs of coding theorems.

Interplay $\varepsilon_{X|Y} \longleftrightarrow$ information measures

- Bounds on $\varepsilon_{X|Y}$ involving information measures exist in the literature.
- Useful for
  - the analysis of M-ary hypothesis testing
  - proofs of coding theorems.
- In this talk, we introduce:

upper and lower bounds on $\varepsilon_{X|Y}$ in terms of the *Arimoto-Rényi* conditional entropy $H_\alpha(X|Y)$ of any order α.

The Rényi Entropy

Definition

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

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

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

The Binary Rényi Divergence

Definition

For $\alpha \in (0,1) \cup (1,\infty)$, the binary Rényi divergence of order α is given by

$$d_\alpha(p\|q) = \frac{1}{\alpha - 1}\,\log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Big). \tag{2}$$

The Binary Rényi Divergence

Definition

For $\alpha \in (0,1) \cup (1,\infty)$, the binary Rényi divergence of order α is given by

$$d_\alpha(p\|q) = \frac{1}{\alpha - 1}\, \log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Big). \qquad (2)$$

$$\lim_{\alpha \uparrow 1} d_\alpha(p\|q) = d(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \qquad (3)$$

Rényi Conditional Entropy ?

- If we mimic the definition of $H(X|Y)$ and define conditional Rényi entropy as

$$\sum_{y \in \mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y),$$

we find that, for $\alpha \neq 1$, the conditional version may be larger than $H_\alpha(X)$!

Rényi Conditional Entropy ?

- If we mimic the definition of $H(X|Y)$ and define conditional Rényi entropy as
$$\sum_{y \in \mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y),$$
we find that, for $\alpha \neq 1$, the conditional version may be larger than $H_\alpha(X)$!
- To remedy this situation, Arimoto introduced a notion of conditional Rényi entropy, $H_\alpha(X|Y)$ (named Arimoto-Rényi conditional entropy), which is upper bounded by $H_\alpha(X)$.

The Arimoto-Rényi Conditional Entropy (cont.)

Definition

Let P_{XY} be defined on $\mathcal{X} \times \mathcal{Y}$, where X is a discrete random variable.

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

$$H_\alpha(X|Y) = \tfrac{\alpha}{1-\alpha} \log \mathbb{E}\left[\left(\sum_{x \in \mathcal{X}} P^\alpha_{X|Y}(x|Y)\right)^{\frac{1}{\alpha}}\right] \quad (4)$$

The Arimoto-Rényi Conditional Entropy (cont.)

Definition

Let P_{XY} be defined on $\mathcal{X} \times \mathcal{Y}$, where X is a discrete random variable.

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

$$H_\alpha(X|Y) = \frac{\alpha}{1-\alpha} \log \mathbb{E}\left[\left(\sum_{x \in \mathcal{X}} P^\alpha_{X|Y}(x|Y)\right)^{\frac{1}{\alpha}}\right] \tag{4}$$

$$= \frac{\alpha}{1-\alpha} \log \sum_{y \in \mathcal{Y}} P_Y(y) \exp\left(\frac{1-\alpha}{\alpha} H_\alpha(X|Y=y)\right), \tag{5}$$

  where (5) applies if Y is a discrete random variable.
- Continuous extension at $\alpha = 0, 1, \infty$ with $H_1(X|Y) = H(X|Y)$.

Fano's Inequality

Let X take values in $|\mathcal{X}| = M$, then

$$H(X|Y) \leq h(\varepsilon_{X|Y}) + \varepsilon_{X|Y} \log(M-1) \tag{6}$$

Fano's Inequality

Let X take values in $|\mathcal{X}| = M$, then

$$H(X|Y) \leq h(\varepsilon_{X|Y}) + \varepsilon_{X|Y} \log(M-1) \tag{6}$$

$$= \log M - d\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big) \tag{7}$$

Fano's Inequality

Let X take values in $|\mathcal{X}| = M$, then

$$H(X|Y) \le h(\varepsilon_{X|Y}) + \varepsilon_{X|Y} \log(M-1) \tag{6}$$
$$= \log M - d\left(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\right) \tag{7}$$

- (7) is not nearly as popular as (6);
- (7) turns out to be the version that admits an elegant (although not immediate) generalization to the Arimoto-Rényi conditional entropy.

Generalization of Fano's Inequality

- It is easy to get Fano's inequality by averaging $H(X|Y=y)$ with respect to the observation y: $H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, H(X|Y=y)$.

Generalization of Fano's Inequality

- It is easy to get Fano's inequality by averaging $H(X|Y=y)$ with respect to the observation y: $H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, H(X|Y=y)$.
- This simple route is not viable in the case of $H_\alpha(X|Y)$ since it is not an average of Rényi entropies of conditional distributions:

$$H_\alpha(X|Y) \neq \sum_{y \in \mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y), \quad \alpha \neq 1. \tag{8}$$

Generalization of Fano's Inequality

- It is easy to get Fano's inequality by averaging $H(X|Y=y)$ with respect to the observation y: $H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, H(X|Y=y)$.
- This simple route is not viable in the case of $H_\alpha(X|Y)$ since it is not an average of Rényi entropies of conditional distributions:

$$H_\alpha(X|Y) \neq \sum_{y \in \mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y), \quad \alpha \neq 1. \tag{8}$$

- The standard proof of Fano's inequality, also fails for $H_\alpha(X|Y)$ of order $\alpha \neq 1$ since it does not satisfy the chain rule.

Generalization of Fano's Inequality

- It is easy to get Fano's inequality by averaging $H(X|Y=y)$ with respect to the observation y: $H(X|Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, H(X|Y=y)$.
- This simple route is not viable in the case of $H_\alpha(X|Y)$ since it is not an average of Rényi entropies of conditional distributions:

$$H_\alpha(X|Y) \neq \sum_{y \in \mathcal{Y}} P_Y(y)\, H_\alpha(X|Y=y), \quad \alpha \neq 1. \tag{8}$$

- The standard proof of Fano's inequality, also fails for $H_\alpha(X|Y)$ of order $\alpha \neq 1$ since it does not satisfy the chain rule.
- Before we generalize Fano's inequality by linking $\varepsilon_{X|Y}$ with $H_\alpha(X|Y)$ for $\alpha \in [0, \infty)$, note that for $\alpha = \infty$, the following equality holds:

$$\varepsilon_{X|Y} = 1 - \exp\bigl(-H_\infty(X|Y)\bigr). \tag{9}$$

Generalization of Fano's Inequality (cont.)

Lemma

Let $\alpha \in (0,1) \cup (1,\infty)$ *and* $(\beta,\gamma) \in (0,\infty)^2$. *Then,*

$$f_{\alpha,\beta,\gamma}(u) = \left(\gamma(1-u)^{\alpha} + \beta u^{\alpha}\right)^{\frac{1}{\alpha}}, \quad u \in [0,1] \tag{10}$$

is

- *strictly convex for* $\alpha \in (1,\infty)$;
- *strictly concave for* $\alpha \in (0,1)$.

$$f''_{\alpha,\beta,\gamma}(u) = (\alpha-1)\beta\gamma\Big(\gamma(1-u)^{\alpha} + \beta u^{\alpha}\Big)^{\frac{1}{\alpha}-2}\Big(u(1-u)\Big)^{\alpha-2} \tag{11}$$

which is strictly negative if $\alpha \in (0,1)$, and strictly positive if $\alpha \in (1,\infty)$.

Generalization of Fano's Inequality (cont.)

Theorem

Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M < \infty$. For all $\alpha \in (0, \infty)$,

$$H_\alpha(X|Y) \le \log M - d_\alpha\big(\varepsilon_{X|Y} \| 1 - \tfrac{1}{M}\big). \tag{12}$$

Equality holds in (12) if and only if, for all y,

$$P_{X|Y}(x|y) = \begin{cases} \frac{\varepsilon_{X|Y}}{M-1}, & x \neq \mathcal{L}^*(y) \\ 1 - \varepsilon_{X|Y}, & x = \mathcal{L}^*(y) \end{cases} \tag{13}$$

where $\mathcal{L}^* \colon \mathcal{Y} \to \mathcal{X}$ is a deterministic MAP decision rule.

Generalization of Fano's Inequality (cont.)

If X, Y are vectors of dimension n, then $\varepsilon_{X|Y} \to 0 \;\Rightarrow\; \frac{1}{n} H(X|Y) \to 0$.
However, the picture with $H_\alpha(X|Y)$ is more nuanced !

Generalization of Fano's Inequality (cont.)

If X, Y are vectors of dimension n, then $\varepsilon_{X|Y} \to 0 \;\Rightarrow\; \frac{1}{n} H(X|Y) \to 0$.
However, the picture with $H_\alpha(X|Y)$ is more nuanced !

Theorem

Assume

- $\{X_n\}$ *is a sequence of random variables;*
- X_n *takes values on* $\mathcal{X}_n$ *such that* $|\mathcal{X}_n| \leq M^n$ *for* $M \geq 2$ *and all* n*;*
- $\{Y_n\}$ *is a sequence of random variables, for which* $\varepsilon_{X_n|Y_n} \to 0$.

a) *If* $\alpha \in (1, \infty]$, *then* $H_\alpha(X_n|Y_n) \to 0$*;*

b) *If* $\alpha = 1$, *then* $\frac{1}{n} H(X_n|Y_n) \to 0$*;*

c) *If* $\alpha \in [0, 1)$, *then* $\frac{1}{n} H_\alpha(X_n|Y_n)$ *is upper bounded by* $\log M$*; nevertheless, it does not necessarily tend to 0.*

Lower Bound on $H_\alpha(X|Y)$

Theorem

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

$$\frac{\alpha}{1-\alpha} \log g_\alpha(\varepsilon_{X|Y}) \leq H_\alpha(X|Y), \tag{14}$$

with the piecewise linear function

$$g_\alpha(t) = \left(k(k+1)^{\frac{1}{\alpha}} - k^{\frac{1}{\alpha}}(k+1)\right)t + k^{\frac{1}{\alpha}+1} - (k-1)(k+1)^{\frac{1}{\alpha}} \tag{15}$$

on the interval $t \in \left[1-\frac{1}{k}, 1-\frac{1}{k+1}\right)$ *for* $k \in \{1,2,\ldots\}$.

- Not restricted to finite M.

Proof Outline

Lemma

Let X be a discrete random variable attaining maximal mass $p_{\max}$. Then, for $\alpha \in (0,1) \cup (1,\infty)$,

$$H_\alpha(X) \geq s_\alpha(\varepsilon_X) \tag{16}$$

where $\varepsilon_X = 1 - p_{\max}$ is the minimum error probability of guessing X, and $s_\alpha \colon [0,1) \to [0,\infty)$ is given by

$$s_\alpha(x) := \frac{1}{1-\alpha} \log\left(\left\lfloor \frac{1}{1-x} \right\rfloor (1-x)^\alpha + \left(1 - (1-x)\left\lfloor \frac{1}{1-x} \right\rfloor\right)^\alpha \right).$$

Equality holds in (16) if and only if P_X has $\left\lfloor \frac{1}{p_{\max}} \right\rfloor$ masses equal to $p_{\max}$.

The proof relies on the Schur-concavity of $H_\alpha(\cdot)$.

Proof Outline (cont.)

For every $y \in \mathcal{Y}$, the lemma yields $H_\alpha(X \,|\, Y = y) \geq s_\alpha\big(\varepsilon_{X|Y}(y)\big)$.

Proof Outline (cont.)

For every $y \in \mathcal{Y}$, the lemma yields $H_\alpha(X \,|\, Y = y) \geq s_\alpha\big(\varepsilon_{X|Y}(y)\big)$.

For $\alpha \in (0,1)$, let $f_\alpha \colon [0,1) \to [1,\infty)$ be defined as

$$f_\alpha(x) = \exp\left(\tfrac{1-\alpha}{\alpha}\, s_\alpha(x)\right)$$

- g_α is the piecewise linear function which coincides with f_α at all points $1 - \frac{1}{k}$ for $k \in \mathbb{N}$;
- g_α is the lower convex envelope of f_α;

$$\begin{aligned} H_\alpha(X|Y) &\geq \tfrac{\alpha}{1-\alpha} \log \mathbb{E}\left[f_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] \text{ (Lemma; } f_\alpha \text{ increasing)} \\ &\geq \tfrac{\alpha}{1-\alpha} \log \mathbb{E}\left[g_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] \;(g_\alpha \leq f_\alpha) \\ &\geq \tfrac{\alpha}{1-\alpha} \log\, g_\alpha(\varepsilon_{X|Y}) \text{ (Jensen)} \end{aligned}$$

Proof Outline (cont.)

For every $y \in \mathcal{Y}$, the lemma yields $H_\alpha(X \mid Y = y) \geq s_\alpha\big(\varepsilon_{X|Y}(y)\big)$.

For $\alpha \in (0,1)$, let $f_\alpha \colon [0,1) \to [1,\infty)$ be defined as

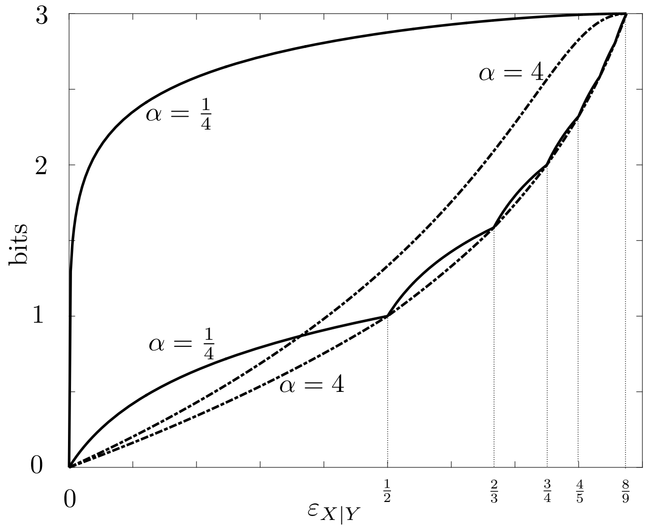
$$f_\alpha(x) = \exp\left(\tfrac{1-\alpha}{\alpha}\, s_\alpha(x)\right)$$

- g_α is the piecewise linear function which coincides with f_α at all points $1 - \frac{1}{k}$ for $k \in \mathbb{N}$;
- g_α is the lower convex envelope of f_α;

$$\begin{aligned} H_\alpha(X|Y) &\geq \tfrac{\alpha}{1-\alpha}\, \log \mathbb{E}\left[f_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] \text{ (Lemma; } f_\alpha \text{ increasing)} \\ &\geq \tfrac{\alpha}{1-\alpha}\, \log \mathbb{E}\left[g_\alpha\big(\varepsilon_{X|Y}(Y)\big)\right] \;(g_\alpha \leq f_\alpha) \\ &\geq \tfrac{\alpha}{1-\alpha}\, \log\, g_\alpha(\varepsilon_{X|Y}) \text{ (Jensen)} \end{aligned}$$

For $\alpha \in (1,\infty)$, $-g_\alpha$ is the lower convex envelope of $-f_\alpha$, and f_α is monotonically decreasing. Proof is similar.

$H_\alpha(X|Y) \longleftrightarrow \varepsilon_{X|Y}$



Asymptotic Tightness

Both upper and lower bounds on $\varepsilon_{X|Y}$ are asymptotically tight as $\alpha \to \infty$.

Asymptotic Tightness

Both upper and lower bounds on $\varepsilon_{X|Y}$ are asymptotically tight as $\alpha \to \infty$.

Special cases

As $\alpha \to 1$, we get existing bounds as special cases:

- Fano's inequality,
- Its counterpart by Kovalevsky ('68), and Tebbe and Dwyer ('68).

Asymptotic Tightness

Both upper and lower bounds on $\varepsilon_{X|Y}$ are asymptotically tight as $\alpha \to \infty$.

Special cases

As $\alpha \to 1$, we get existing bounds as special cases:

- Fano's inequality,
- Its counterpart by Kovalevsky ('68), and Tebbe and Dwyer ('68).

Upper bound on $\varepsilon_{X|Y}$

The most useful domain of applicability of the counterpart to the generalization of Fano's inequality is $\varepsilon_{X|Y} \in [0, \frac{1}{2}]$, in which case the lower bound specializes to $(k = 1)$

$$\frac{\alpha}{1-\alpha} \log\left(1 + \left(2^{\frac{1}{\alpha}} - 2\right)\varepsilon_{X|Y}\right) \leq H_\alpha(X|Y). \tag{17}$$

List Decoding

- Decision rule outputs a list of choices.
- The extension of Fano's inequality to list decoding, expressed in terms of the conditional Shannon entropy, was initiated by Ahlswede, Gacs and Körner ('66).
- Useful for proving converse results.

Generalization of Fano's Inequality for List Decoding (cont.)

Theorem (Fixed List Size)

Let P_{XY} *be a probability measure defined on* $\mathcal{X} \times \mathcal{Y}$ *where* $|\mathcal{X}| = M$*. Consider a decision rule*[a] $\mathcal{L} \colon \mathcal{Y} \to \binom{\mathcal{X}}{L}$*, and denote the decoding error probability by* $P_{\mathcal{L}} = \mathbb{P}\big[X \notin \mathcal{L}(Y)\big]$*. Then, for all* $\alpha \in (0,1) \cup (1,\infty)$*,*

$$H_\alpha(X|Y) \leq \log M - d_\alpha\big(P_{\mathcal{L}} \| 1 - \tfrac{L}{M}\big) \tag{18}$$

with equality in (18) *if and only if*

$$P_{X|Y}(x|y) = \begin{cases} \frac{P_{\mathcal{L}}}{M-L}, & x \notin \mathcal{L}(y) \\ \frac{1-P_{\mathcal{L}}}{L}, & x \in \mathcal{L}(y). \end{cases} \tag{19}$$

[a] $\binom{\mathcal{X}}{L}$ stands for the set of all subsets of $\mathcal{X}$ with cardinality L, with $L \leq |\mathcal{X}|$.

Further Results

- Explicit lower bounds on $\varepsilon_{X|Y}$ as a function of $H_\alpha(X|Y)$ for an arbitrary α (also, for $\alpha < 0$).
- Lower bounds on the list decoding error probability for fixed list size as a function of $H_\alpha(X|Y)$ for an arbitrary α (also, for $\alpha < 0$).
- New bounds on $\varepsilon_{X|Y}$ in terms of the Chernoff information and Rényi divergence.
- Application of $H_\alpha(X|Y)$-$\varepsilon_{X|Y}$ bounds: Analyzing the exponential decay of the Arimoto-Rényi conditional entropy of the message given the channel output for DMCs and random coding ensembles.

Journal Paper

I. Sason and S. Verdú, "Arimoto-Rényi conditional entropy and Bayesian M-ary hypothesis testing," submitted to the *IEEE Trans. on Information Theory* in September 2016, and revised in May 2017.
[Online]. Available at https://arxiv.org/abs/1701.01974.