Operational interpretation of Rényi conditional mutual information via composite hypothesis testing against Markov distributions

Marco Tomamichel[1], Masahito Hayashi[2]

[1]School of Physics, The University of Sydney
[2]Graduate School of Mathematics, Nagoya University, and
Centre for Quantum Technologies, National University of Singapore

THE UNIVERSITY OF SYDNEY

ISIT 2016, Barcelona
(arXiv: 1511.04874)

Simple Hypothesis testing

- Binary hypothesis testing (HT) is fundamental in statistics and information theory.

(Simple) Binary HT

- Sequence of random variables $X^n = (X_1, X_2, \ldots, X_n)$ with X_i taking values in $\mathcal{X}$.
- Two distributions $P, Q \in \mathcal{P}(\mathcal{X})$.

$$\text{null hypothesis: } X^n \sim P^{\times n}\,,$$
$$\text{alternative hypothesis: } X^n \sim Q^{\times n}.$$

- Sequence of tests, maps $T^n : \mathcal{X}^n \to [0,1]$.
- Define errors of two kinds,

$$\alpha_n(T^n) = \mathbb{E}_{P^{\times n}}[1 - T^n(X^n)] \text{ and } \beta_n(T^n) = \mathbb{E}_{Q^{\times n}}[T^n(X^n)].$$

Critical rate

- The goal is to understand the asymptotic tradeoff between α_n and β_n for optimal test sequences.

Stein's lemma

Let T^n be a sequence with $\alpha_n \leq \varepsilon$, for $\varepsilon \in (0,1)$. Then

$$\beta_n \geq \exp\big(-nD(P\|Q) + o(n)\big)$$

and there is a sequence that achieves this.

- This gives operational significance to the relative entropy:

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}\,.$$

- $D(P\|Q)$ is a critical rate: if β_n vanishes faster than $\exp(-nD(P\|Q))$ then α_n must converge to 1.

Small deviations

- Strassen (1962) showed a refinement for small deviations from the critical rate.

Second order refinement

Let T^n be a sequence with $\beta_n \leq \exp(-nD(P\|Q) - \sqrt{n}r)$ for some $r \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \alpha_n \geq \Phi\left(\frac{r}{\sqrt{V(P\|Q)}}\right)$$

and there is a sequence that achieves this.

- Φ is the cumulative standard normal distribution function.
- The relative entropy variance characterizes the second order:

$$V(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\left(\log\frac{P(x)}{Q(x)} - D(P\|Q)\right)^2 .$$

Large deviations

- For rates below the relative entropy we find the error exponent (attributed to Hoeffding).

Error exponent

Let T^n be a sequence with $\beta_n \leq \exp(-nR)$ for $R \geq 0$. Then

$$\lim_{n\to\infty} -\frac{1}{n}\log \alpha_n \leq \sup_{s\in(0,1)} \left\{ \frac{1-s}{s}\big(D_s(P\|Q) - R\big) \right\}$$

and there is a sequence that achieves this.

- Here the Rényi divergence is given by (Rényi, 1961)

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\left(\sum_{x\in\mathcal{X}} P(x)^\alpha Q(x)^{1-\alpha}\right)$$

- This result is only meaningful for $R \leq D_1(P\|Q) = D(P\|Q)$.

Composite hypothesis testing

- In our work we look at a general framework of HT problems:

HT with composite alternative hypothesis

- Sequence of random variables $X^n = (X_1, X_2, \ldots, X_n)$ with X_i taking values in $\mathcal{X}$.
- A distributions $P \in \mathcal{P}(\mathcal{X})$ and a sequence of sets $\{\mathcal{Q}_n\}_{n\in\mathbb{N}}$ with $\mathcal{Q}_n \subset \mathcal{P}(\mathcal{X}^n)$.

$$\text{null hypothesis: } X^n \sim P^{\times n},$$
$$\text{alternative hypothesis: } X^n \sim Q^n, \text{ for } Q^n \in \mathcal{Q}_n.$$

- Error is now $\beta_n(T) = \max_{Q^n \in \mathcal{Q}_n} \mathbb{E}_{Q^n}[T^n(X^n)]$.
- The $\mathcal{Q}_n$ characterize the composite hypothesis.
- We show that under certain conditions on $\mathcal{Q}_n$ variations of the above results still hold.

Axioms for $\mathcal{Q}_n$

- Define $D_\alpha(P\|\mathcal{Q}) := \inf_{Q\in\mathcal{Q}} D_\alpha(P\|Q)$.

Axiom 1: convexity

The base set $\mathcal{Q} = \mathcal{Q}_1$ is convex. Moreover, $\arg\min_{Q\in\mathcal{Q}} D_s(P\|Q)$ lies in the relative interior of $\mathcal{Q}$ for all s (and is thus unique).

Axiom 2: independent identical distributions (i.i.d.)

We have $Q^{\times n} \in \mathcal{Q}_n$ for every $Q \in \mathcal{Q}$.

- From Axiom 2 follows that $D_s(P^{\times n}\|\mathcal{Q}_n) \leq nD_s(P\|\mathcal{Q})$.

Axiom 3: superadditivity

For all $s \geq 0$ we have $D_s(P^{\times n}\|\mathcal{Q}_n) \geq nD_s(P\|\mathcal{Q})$.

- Hence if Axioms 2&3 hold we have equality, or additivity.

- A distribution $Q^n \in \mathcal{P}(\mathcal{X}^n)$ is permutation invariant (p.i.) if

$$\underbrace{Q^n(x_1, x_2, \ldots, x_n)}_{Q^n(x^n)} = \underbrace{Q^n(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})}_{Q^n(\pi(x^n))}$$

for all $\pi \in S_n$ and $x^n \in \mathcal{X}^n$.
- The set $\mathcal{Q}_n^{\text{p.i.}}$ comprises all p.i. elements of $\mathcal{Q}_n$.

Axiom 4a: universal distribution

There exists a sequence of distributions $U^n \in \mathcal{Q}_n^{\text{p.i.}}$ and a polynomial $v(n)$ such that, for all $Q^n \in \mathcal{Q}_n^{\text{p.i.}}$,

$$Q^n(x^n) \leq v(n) U^n(x^n), \qquad \forall x^n \in \mathcal{X}^n \,.$$

- The map $Q^n(\cdot) \mapsto \frac{1}{n!} \sum_\pi Q^n(\pi(\cdot))$ is called symmetrization.

Axiom 4b: symmetrization

The set $\mathcal{Q}_n$ is closed under symmetrization.

An important consequence

- The importance of the universal distribution lies here:

Lemma: universal test

If Axioms 2–4 hold, then

$$\lim_{n\to\infty} \frac{1}{n} D_s(P^{\times n} \| U^n) = D_s(P \| \mathcal{Q}) .$$

Proof of '$\geq$': Implied by additivity. □

Proof of '$\leq$': For every $Q \in \mathcal{Q}$ we find that $Q^{\times n} \in \mathcal{Q}_n^{\text{p.i.}}$. Hence,

$$\begin{aligned} D_s(P^{\times n} \| U^n) &\leq D_s(P^{\times n} \| Q^{\times n}) + \log v(n) \\ &= n D_s(P \| Q) + O(\log n) . \end{aligned}$$

Inequality follows by taking limit and supremum over $Q \in \mathcal{Q}$. □

Main result: large deviations

- Define optimal constrained error as

$$\hat{\alpha}_n(\mu) := \min_{T^n}\{\alpha_n(T^n) : \beta_n(T^n) \leq \mu\}.$$

Theorem: error exponent

Assume Axioms 1–4 hold. For any $R \leq D(P\|\mathcal{Q})$,

$$\lim_{n\to\infty} -\frac{1}{n}\log \hat{\alpha}_n\left(\exp\left(-nR\right)\right) = \sup_{s\in(0,1)} \left\{\frac{1-s}{s}\big(D_s(P\|\mathcal{Q}) - R\big)\right\}.$$

Proof of achievability:

- We use a Neyman-Pearson tests between $P^{\times n}$ and the universal distribution U^n.

$$T_n(x^n) = \begin{cases} 1 & \text{if } P^{\times n}(x^n) \geq \lambda_n U^n(x^n) \\ 0 & \text{else} \end{cases}.$$

Proof of achievability (continued):

- For the error α_n we find

$$\begin{aligned}\alpha_n(T_n) &= P^{\times n}\big[P^{\times n}(X^n) < \lambda_n U^n(X^n)\big]\\ &\le \lambda_n^{1-s}\exp\big((s-1)D_s(P^{\times n}\|U^n)\big).\end{aligned}$$

- For the error β_n we find

$$\begin{aligned}\beta_n(T_n) &= \max_{Q^n\in\mathcal{Q}_n} Q^n\big[P^{\times n}(X^n) \ge \lambda_n U^n(X^n)\big]\\ &= \max_{Q^n\in\mathcal{Q}_n^{\text{p.i.}}} Q^n\big[P^{\times n}(X^n) \ge \lambda_n U^n(X^n)\big]\\ &\le v(n)\, U^n\big[P^{\times n}(X^n) \ge \lambda_n U^n(X^n)\big]\\ &\le v(n)\, \lambda_n^{-s}\exp\big((s-1)D_s(P^{\times n}\|U^n)\big).\end{aligned}$$

- We chose λ_n such that the above is bounded by $\exp(-nR)$. The corresponding $\alpha(T^n)$ is an upper bound on $\hat{\alpha}_n$. We find

$$-\log\hat{\alpha}_n\left(\exp\left(-nR\right)\right) \ge \frac{1-s}{s}\big(D_s(P^{\times n}\|U^n) - nR - \log v(n)\big).$$

- And we see that the rhs. converges to the expected quantity.
- We optimize over $s\in(0,1)$. □

Main result: second order

- Let $Q^* = \arg\min_{Q\in\mathcal{Q}} D(P\|Q)$, define $V(P\|\mathcal{Q}) = V(P\|Q^*)$.

Theorem: second order

Assume Axioms 1–4 hold. For any $r \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \hat{\alpha}_n \left(\exp\left(-nD(P\|\mathcal{Q}) - \sqrt{n}r\right)\right) = \Phi\left(\frac{r}{\sqrt{V(P\|\mathcal{Q})}}\right).$$

Proof of achievability:

- We use the same test.

$$T_n(x^n) = \begin{cases} 1 & \text{if } P^{\times n}(x^n) \geq \lambda_n U^n(x^n) \\ 0 & \text{else} \end{cases}.$$

- For $s = 1$ the errors are bounded as

$$\beta_n(T^n) \leq v(n)\lambda_n^{-1} \quad \text{and} \quad \alpha_n = P^{\times n}\left[P^{\times n}(X^n) < \lambda_n U^n(X^n)\right].$$

Proof of achievability (continued):

- We set $\lambda_n = v(n)\ \exp(nD(P\|\mathcal{Q}) + \sqrt{n}r)$ and find

$$\alpha(T^n) = P^{\times n}[Y_n(X^n) < r] \qquad \text{with}$$

$$Y_n = \frac{1}{\sqrt{n}} \left(\log P^{\times n}(X^n) - \log U^n(X^n) - nD(P\|\mathcal{Q}) - \log v(n) \right).$$

- The cumulant generating function of the sequence Y_n converges to a quadratic function:

$$\begin{aligned} \log M_Y(t) &= \lim_{n\to\infty} \log \mathbb{E}[\exp(tY_n)] \\ &= \lim_{n\to\infty} \left\{ \frac{t}{\sqrt{n}} \left(D_{1+\frac{t}{\sqrt{n}}}(P^{\times n}\|U^n) - nD(P\|\mathcal{Q}) \right) \right\} \\ &= \frac{t^2}{2} V(P\|\mathcal{Q}). \end{aligned}$$

- By Lévi's theorem, Y_n converges in probability to a Gaussian distribution with zero mean and variance $V(P\|\mathcal{Q})$. □

Example: testing against Markov distributions

HT against Markov distribution

- Sequences of random variables (X^n, Y^n, Z^n).
- A distribution $P_{XYZ} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Y})$.

$$\text{null hypothesis: } (X^n, Y^n, Z^n) \sim P_{XYZ}^{\times n}\,,$$
$$\text{alternative hypothesis: } X^n \leftrightarrow Y^n \leftrightarrow Z^n,\ (X^n, Y^n) \sim P_{XY}^{\times n}.$$

- The alternate hypothesis has fixed i.i.d. marginal on (X^n, Y^n), but arbitrarily correlated with Z^n.

$$\mathcal{Q}_n = \left\{ P_{XY}^{\times n} \times Q_{Z^n|Y^n} : Q_{Z^n|Y^n} \in \mathcal{P}(\mathcal{Z}^n|\mathcal{Y}^n) \right\}$$

- Other variants are discussed in the paper.

Checking axioms: α-conditional mutual information

- Minimizing the relative entropy yields the conditional mutual information (CMI):

$$\min_{Q_{XYZ}\in\mathcal{Q}} D(P_{XYZ}\|Q_{XYZ}) = \min_{Q_{Z|Y}\in\mathcal{P}(Z|Y)} D(P_{XYZ}\|P_{XY}\times Q_{Z|Y})$$
$$= D(P_{XYZ}\|P_{XY}\times P_{Z|Y}) = I(X:Z|Y),$$

- Minimizing the Rényi divergence yields a Rényi or α-CMI:

$$\min_{Q_{XYZ}\in\mathcal{Q}} D_\alpha(P_{XYZ}\|Q_{XYZ}) = D_\alpha(P_{XYZ}\|P_{XY}\times Q^{*,\alpha}_{Z|Y}) = I_\alpha(X:Z|Y)\,.$$

where the optimal distribution is given by

$$Q^{*,\alpha}_{Z|Y}(z|y) = \frac{P_{Z|Y=y}(z)\left(\sum_x P^{\alpha}_{X|Z=z,Y=y}(x)P^{1-\alpha}_{X|Y=y}(x)\right)^{\frac{1}{\alpha}}}{\sum_z P_{Z|Y=y}(z)\left(\sum_x P^{\alpha}_{X|Z=z,Y=y}(x)P^{1-\alpha}_{X|Y=y}(x)\right)^{\frac{1}{\alpha}}}$$

and the α-CMI thus evaluates to $I_\alpha(X:Z|Y) =$

$$\frac{1}{\alpha-1}\log\left(\sum_y P_Y(y)\left(\sum_z P_{Z|Y=y}(z)\left(\sum_x P_{X|Y=y,Z=z}(x)^{\alpha}P_{X|Y=y}(x)^{1-\alpha}\right)^{\frac{1}{\alpha}}\right)^{\alpha}\right).$$

Checking axioms: Universal Markov distribution

- Axiom 1 satisfied: The set $\mathcal{Q} = \mathcal{Q}_1$ is convex, the optimizers $Q^{*,\alpha}_{X|Y}$ lie in its relative interior.
- Axiom 2,4b satisfied: The sets $\mathcal{Q}_n$ contain product distributions and are closed under permutations.
- Axiom 3 satisfied: Additivity implied by structure of $Q^{*,\alpha}_{Z|Y}$, i.e.

$$Q^{*,\alpha}_{Z^n|Y^n} = \left(Q^{*,\alpha}_{Z|Y}\right)^{\times n}$$

- Axiom 4a satisfied: There exists a sequence of permutation covariant universal channels $U^n_{Z^n|Y^n}$.

Proof for trivial Y^n: Let $\mathcal{T}_n$ be the set of n-types.

$$U^n_{Z^n}(z^n) = \frac{1}{|\mathcal{T}_n|} \sum_{\lambda \in \mathcal{T}_n} \frac{1\{x^n \text{ is of type } \lambda\}}{\sum_{y^n} 1\{y^n \text{ is of type } \lambda\}}$$

For any p.i. distribution $P_{Z^n} \leq |\mathcal{T}_n| U^n_{Z^n}$ and $|\mathcal{T}_n| = \mathrm{poly}(n)$. □

Connection to channel coding

HT against Markov distribution

null hypothesis: $(X^n, Y^n, Z^n) \sim P_{XYZ}^{\times n}$,

alternative hypothesis: $X^n \leftrightarrow Y^n \leftrightarrow Z^n$, $(X^n, Y^n) \sim P_{XY}^{\times n}$.

- The error exponent/reliability function is given by

$$\sup_{s \in (0,1)} \left\{ \frac{1-s}{s} \big(I_s(X:Z|Y) - R \big) \right\}, \quad R \leq I(X:Z|Y).$$

- For trivial Y this is simply the Gallager function:

$$\begin{aligned} I_s(X:Z) &= \min_{Q_Z \in \mathcal{P}(Z)} D_s(P_{XY} \| P_X \times Q_Z) \\ &= \frac{s}{1-s} \log \sum_z \left(\sum_x P_X(x) P_{Z|X=x}(z)^s \right)^{1/s} = E_0\left(\frac{1-s}{s}, P_X, P_{Z|X} \right). \end{aligned}$$

- We may rewrite the exponent as: $\sup_{\rho \geq 0} E_0(\rho, P_X, P_{Z|X}) - \rho R$.

- This is not entirely expected in light of the Polyanskiy et al. (2010) and Vasquez-Vilar et al. (2016).
- The latter show that the average error for a codebook P_X with $P_X(x) \in \{0, \frac{1}{M}\}$ of size M satisfies

$$\bar{\varepsilon}(P_X) = \hat{\alpha}\left(\frac{1}{M}\right)$$

for the HT problem

HT against crappy channel

null hypothesis: $(X, Y) \sim P_X \times W_{Y|X}$,

alternative hypothesis: $X \sim P_X$, independent of Y.

- The meta converse bounds the average error for any codebook:

$$\bar{\varepsilon} \geq \min_{P_X \in \mathcal{P}(\mathcal{X})} \hat{\alpha}\left(\frac{1}{M}\right)$$

Summary and Outlook

- In the paper we analyze error exponents, strong converse exponents and second order asymptotics for HT problems where the composite alternative hypothesis satisfies slightly weaker axioms.
- We show how HT against Markov distributions yields an operational interpretation for Rényi CMI.
- The relation between the channel coding single-shot bounds and our asymptotics remain unclear.
  - Can we derive the sphere packing and random coding bounds in the composite hypothesis testing picture?
- Does composite hypothesis testing against Markov distribution have similar relations to single-shot network coding problems?