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

Marco Tomamichel\textsuperscript{1}, Masahito Hayashi\textsuperscript{2}

\textsuperscript{1}School of Physics, The University of Sydney
\textsuperscript{2}Graduate School of Mathematics, Nagoya University, and Centre for Quantum Technologies, National University of Singapore

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 $P, Q \in \mathcal{P}(\mathcal{X})$.

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

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

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

- The goal is to understand the asymptotic tradeoff between $\alpha_n$ and $\beta_n$ for optimal test sequences.

### Stein’s lemma

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

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

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 $\beta_n$ vanishes faster than $\exp(-nD(P\|Q))$ then $\alpha_n$ 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.

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

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

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

and there is a sequence that achieves this.

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

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

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

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

<table>
<thead>
<tr>
<th>HT with composite alternative hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Sequence of random variables ( X^n = (X_1, X_2, \ldots, X_n) ) with ( X_i ) taking values in ( \mathcal{X} ).</td>
</tr>
<tr>
<td>• A distributions ( P \in \mathcal{P}(\mathcal{X}) ) and a sequence of sets ( {Q_n}_{n \in \mathbb{N}} ) with ( Q_n \subset \mathcal{P}(\mathcal{X}^n) ).</td>
</tr>
</tbody>
</table>

null hypothesis: \( X^n \sim P \times^n \),
alternative hypothesis: \( X^n \sim Q^n \), for \( Q^n \in Q_n \).

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

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

Axiom 1: convexity

The base set $Q = Q_1$ is convex. Moreover, $\arg \min_{Q \in Q} D_s(P\|Q)$ lies in the relative interior of $Q$ for all $s$ (and is thus unique).

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

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

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

Axiom 3: superadditivity

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

- 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

\[
Q^n(x_1, x_2, \ldots, x_n) = Q^n(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})
\]

for all \( \pi \in S_n \) and \( x^n \in \mathcal{X}^n \).

• The set \( Q_{n}^{\text{p.i.}} \) comprises all p.i. elements of \( Q_n \).

**Axiom 4a: universal distribution**

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

\[
Q^n(x^n) \leq v(n)U^n(x^n), \quad \forall x^n \in \mathcal{X}^n.
\]

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

**Axiom 4b: symmetrization**

The set \( Q_n \) is closed under symmetrization.
An important consequence

- The importance of the universal distribution lies here:

**Lemma: universal test**

If Axioms 2–4 hold, then

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

**Proof of ‘≥’:** Implied by additivity.

**Proof of ‘≤’:** For every \( Q \in \mathcal{Q} \) we find that \( Q^{\times n} \in \mathcal{Q}_n^{p.i.} \). Hence,

\[
D_s(P^{\times n} \| U^n) \leq D_s(P^{\times n} \| Q^{\times n}) + \log v(n)
= nD_s(P \| Q) + O(\log n).
\]

Inequality follows by taking limit and supremum over \( Q \in \mathcal{Q} \).
Main result: large deviations

- Define optimal constrained error as

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

Theorem: error exponent

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

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

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 $\alpha_n$ we find
  \[
  \alpha_n(T_n) = P^{\times n} \left[ P^{\times n}(X^n) < \lambda_n U^n(X^n) \right] \leq \lambda_n^{-s} \exp \left( (s - 1) D_s(P^{\times n} \| U^n) \right).
  \]

- For the error $\beta_n$ we find
  \[
  \beta_n(T_n) = \max_{Q^n \in \mathcal{Q}_n} Q^n \left[ P^{\times n}(X^n) \geq \lambda_n U^n(X^n) \right] \leq v(n) \lambda_n^{-s} \exp \left( (s - 1) D_s(P^{\times n} \| U^n) \right).
  \]

- We chose $\lambda_n$ such that the above is bounded by $\exp(-nR)$. The corresponding $\alpha(T^n)$ is an upper bound on $\hat{\alpha}_n$. We find
  \[- \log \alpha_n \left( \exp(-nR) \right) \geq \frac{1-s}{s} \left( D_s(P^{\times n} \| U^n) - nR - \log v(n) \right).\]

- 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 $V(P \| Q) = V(P \| Q^*)$.

**Theorem: second order**

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

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

**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 \( \lambda_n = v(n) \exp(nD(P\|Q) + \sqrt{n}r) \) and find
  \[
  \alpha(T^n) = P^\times_n [Y_n(X^n) < r] \quad \text{with} \quad Y_n = \frac{1}{\sqrt{n}} \left( \log P^\times_n(X^n) - \log U^n(X^n) - nD(P\|Q) - \log v(n) \right).
  \]

- The **cumulant generating function** of the sequence \( Y_n \) converges to a quadratic function:
  \[
  \log M_Y(t) = \lim_{n \to \infty} \log E[\exp(tY_n)]
  = \lim_{n \to \infty} \left\{ \frac{t}{\sqrt{n}} \left( D_1 + \frac{t}{\sqrt{n}} (P^\times_n\|U^n) - nD(P\|Q) \right) \right\}
  = \frac{t^2}{2} V(P\|Q).
  \]

- By **Lévi’s theorem**, \( Y_n \) converges in probability to a Gaussian distribution with zero mean and variance \( V(P\|Q) \). \( \square \)
Example: testing against Markov distributions

HT against Markov distribution

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

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

\[ Q_n = \left\{ P_{XY}^{\times n} \times Q_{Z^n|Y^n} : Q_{Z^n|Y^n} \in \mathcal{P}(Z^n|Y^n) \right\} \]

- Other variants are discussed in the paper.
Checking axioms: $\alpha$-conditional mutual information

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

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

$$= D(P_{XYZ} \parallel P_{XY} \times P_{Z|Y}) = I(X : Z|Y),$$

- Minimizing the Rényi divergence yields a Rényi or $\alpha$-CMI:

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

where the optimal distribution is given by

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

and the $\alpha$-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 $Q = Q_1$ is convex, the optimizers $Q^*_{X|Y}$ lie in its relative interior.

- **Axiom 2,4b satisfied**: The sets $Q_n$ contain product distributions and are closed under permutations.

- **Axiom 3 satisfied**: Additivity implied by structure of $Q^*_{Z|Y}$, i.e.

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

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

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

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

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

Connection to channel coding

HT against Markov distribution

null hypothesis: \((X^n, Y^n, Z^n) \sim P^n_{XYZ}\),
alternative hypothesis: \(X^n \leftrightarrow Y^n \leftrightarrow Z^n,\ (X^n, Y^n) \sim P^n_{XY}\).

- The error exponent/reliability function is given by

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

- For trivial \(Y\) this is simply the Gallager function:

  \[
  I_s(X:Z) = \min_{Q_Z \in \mathcal{P}(Z)} D_s(P_{XY} \| P_X \times Q_Z)
  \]

  \[
  = \frac{s}{1 - s} \log \sum_z \left( \sum_x P_X(x) P_{Z|X=x}(z)^s \right)^{1/s} = E_0\left( \frac{1 - s}{s}, P_X, P_{Z|X} \right).
  \]

- 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?