On the Rényi Divergence, the Joint Range of Relative Entropies, and a Channel Coding Theorem

Igal Sason

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

2015 IEEE International Symposium on Information Theory

Hong Kong
June 14–19, 2015.

Total Variation (TV) Distance

Let P, Q be probability measures defined on the measurable space $(\mathcal{A}, \mathscr{F})$.

$$|P - Q| = 2 \sup_{\mathcal{F} \in \mathscr{F}} |P(\mathcal{F}) - Q(\mathcal{F})| = |P - Q|_1.$$

Total Variation (TV) Distance

Let P, Q be probability measures defined on the measurable space $(\mathcal{A}, \mathscr{F})$.

$$|P - Q| = 2 \sup_{\mathcal{F} \in \mathscr{F}} |P(\mathcal{F}) - Q(\mathcal{F})| = |P - Q|_1.$$

The Rényi Divergence of order α

Let

- $P \ll Q$.
- $Y \sim Q$.
- $\alpha \in (0,1) \cup (1, \infty)$.

$$D_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^{\alpha} (Y) \right].$$

If $D(P\|Q) < \infty \;\Rightarrow\; D(P\|Q) = \lim_{\alpha \to 1^-} D_\alpha(P\|Q)$.

Exact Characterization of the Joint Range of the Relative Entropies

Question

Let

- $\varepsilon \in (0, 2)$ *be fixed.*
- P_1, P_2 *be arbitrary PDs s.t.* $|P_1 - P_2| \geq \varepsilon$.
- Q *is an arbitrary PD s.t.* $Q \ll P_1, P_2$.

1. *What is the achievable region of* $\big(D(Q\|P_1),\, D(Q\|P_2)\big)$ *where none of these three distributions is fixed ?*

Exact Characterization of the Joint Range of the Relative Entropies

Question

Let

- $\varepsilon \in (0, 2)$ *be fixed.*
- P_1, P_2 *be arbitrary PDs s.t.* $|P_1 - P_2| \geq \varepsilon$.
- Q *is an arbitrary PD s.t.* $Q \ll P_1, P_2$.

1. *What is the achievable region of* $\big(D(Q\|P_1),\, D(Q\|P_2)\big)$ *where none of these three distributions is fixed ?*
2. *Given an arbitrary point in this region, specify PDs* P_1, P_2, Q *that achieve this point.*

Possible Context

Methods of Types:

$$P_1^n\big(T(Q)\big) \doteq e^{-nD(Q\|P_1)}, \quad P_2^n\big(T(Q)\big) \doteq e^{-nD(Q\|P_2)}$$

$\Rightarrow$ Exponential decay rates for probabilities of rare events.

Possible Context

Methods of Types:

$$P_1^n\big(T(Q)\big) \doteq e^{-nD(Q\|P_1)}, \quad P_2^n\big(T(Q)\big) \doteq e^{-nD(Q\|P_2)}$$

$\Rightarrow$ Exponential decay rates for probabilities of rare events.

Approach for Solving the Problem

- Minimizing the Rényi divergence subject to a minimal TV distance.
- Using the solution for answering the question.

Minimization of the Rényi Divergence s.t. Minimal TV Distance

For $\alpha > 0$, let

$$g_\alpha(\varepsilon) = \inf_{P_1, P_2 \colon |P_1 - P_2| \geq \varepsilon} D_\alpha(P_1 \| P_2), \quad \forall\, \varepsilon \in [0, 2).$$

Minimization of the Rényi Divergence s.t. Minimal TV Distance

For $\alpha > 0$, let

$$g_\alpha(\varepsilon) = \inf_{P_1, P_2 \colon |P_1 - P_2| \geq \varepsilon} D_\alpha(P_1 \| P_2), \quad \forall \, \varepsilon \in [0, 2).$$

Proposition: There is no loss of generality by restricting the minimization of $g_\alpha(\varepsilon)$, for $\varepsilon \in (0, 2)$, to pairs of 2-element PDs.

Minimization of the Rényi Divergence s.t. Minimal TV Distance

For $\alpha > 0$, let

$$g_\alpha(\varepsilon) = \inf_{P_1, P_2 \colon |P_1 - P_2| \geq \varepsilon} D_\alpha(P_1 \| P_2), \quad \forall \, \varepsilon \in [0, 2).$$

Proposition: There is no loss of generality by restricting the minimization of $g_\alpha(\varepsilon)$, for $\varepsilon \in (0, 2)$, to pairs of 2-element PDs. Hence,

$$g_\alpha(\varepsilon) = \min_{p, q \in [0,1] \colon |p-q| \geq \frac{\varepsilon}{2}} d_\alpha(p \| q)$$

where

$$d_\alpha(p \| q) \triangleq \frac{\log\Bigl(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Bigr)}{\alpha - 1}.$$

The minimizing probability distributions: $P_1 = (p, 1-p)$, $P_2 = (q, 1-q)$.

An identity for the Rényi divergence

For $\alpha \in (0,1) \cup (1,\infty) \setminus \{1\}$

$$D_\alpha(P_1\|P_2) = D(Q\|P_2) + \frac{\alpha}{1-\alpha} \cdot D(Q\|P_1) + \frac{1}{\alpha-1} \cdot D(Q\|Q_\alpha)$$

where Q_α is given by

$$Q_\alpha(x) \triangleq \frac{P_1^\alpha(x)\, P_2^{1-\alpha}(x)}{\sum_u P_1^\alpha(u)\, P_2^{1-\alpha}(u)}, \quad \forall\, x \in \mathsf{Supp}(P_1).$$

An identity for the Rényi divergence

For $\alpha \in (0,1) \cup (1,\infty) \setminus \{1\}$

$$D_\alpha(P_1 \| P_2) = D(Q \| P_2) + \frac{\alpha}{1-\alpha} \cdot D(Q \| P_1) + \frac{1}{\alpha - 1} \cdot D(Q \| Q_\alpha)$$

where Q_α is given by

$$Q_\alpha(x) \triangleq \frac{P_1^\alpha(x)\, P_2^{1-\alpha}(x)}{\sum_u P_1^\alpha(u)\, P_2^{1-\alpha}(u)}, \quad \forall\, x \in \mathrm{Supp}(P_1).$$

This comes as a direct calculation, following a result by Shayevitz (ISIT '11) where for $\alpha > 1$

$$D_\alpha(P_1 \| P_2) = \max_{Q \ll P_1} \left\{ D(Q \| P_2) + \frac{\alpha}{\alpha - 1} \cdot D(Q \| P_1) \right\}$$

and the $\max$ is replaced by $\min$ for $\alpha \in (0,1)$.

An Exact Characterization of the Region

The boundary is determined by letting α increase continuously in (0,1), and drawing the straight lines in the plane of $\big(D(Q\|P_1), D(Q\|P_2)\big)$:

$$D(Q\|P_2) + \frac{\alpha}{1-\alpha} \cdot D(Q\|P_1) = g_\alpha(\varepsilon), \quad \forall\, \alpha \in (0,1).$$

Every point on the boundary is a tangent point to one of the straight lines.

An Exact Characterization of the Region (Cont.)

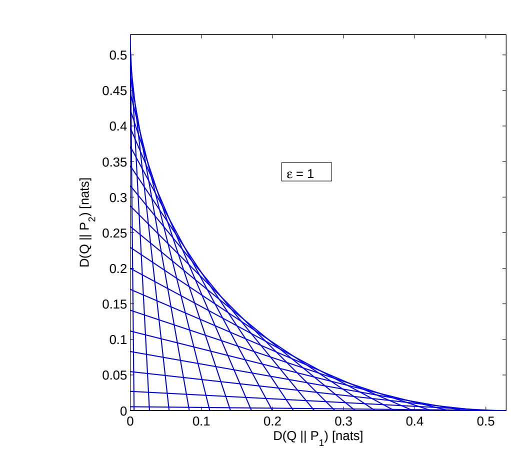

Figure: The achievable region of $(D(Q\|P_1), D(Q\|P_2))$ where $|P_1 - P_2| \geq 1$ is the upper envelope of the straight lines.

An Exact Characterization of the Region (Cont.)

The triple of 2-element PDs P_1, P_2 and Q that achieves an arbitrary point on the boundary of this region is determined as follows:

- Find the slope s of the tangent line $(s < 0)$, and determine $\alpha \in (0,1)$ such that $-\frac{\alpha}{1-\alpha} = s \ \Rightarrow \ \alpha = -\frac{s}{1-s}$.

An Exact Characterization of the Region (Cont.)

The triple of 2-element PDs P_1, P_2 and Q that achieves an arbitrary point on the boundary of this region is determined as follows:

- Find the slope s of the tangent line $(s < 0)$, and determine $\alpha \in (0, 1)$ such that $-\frac{\alpha}{1-\alpha} = s \;\Rightarrow\; \alpha = -\frac{s}{1-s}$.
- Determine the 2-element PDs $P_1 = (p, 1-p)$, $P_2 = (q, 1-q)$ such that $d_\alpha(p\|q) = g_\alpha(\varepsilon)$.

An Exact Characterization of the Region (Cont.)

The triple of 2-element PDs P_1, P_2 and Q that achieves an arbitrary point on the boundary of this region is determined as follows:

- Find the slope s of the tangent line ($s < 0$), and determine $\alpha \in (0,1)$ such that $-\frac{\alpha}{1-\alpha} = s \Rightarrow \alpha = -\frac{s}{1-s}$.
- Determine the 2-element PDs $P_1 = (p, 1-p)$, $P_2 = (q, 1-q)$ such that $d_\alpha(p\|q) = g_\alpha(\varepsilon)$.
- Calculate the 2-element PD $Q = Q_\alpha$ (as above) for the calculated α, p and q.

An Exact Characterization of the Region (Cont.)

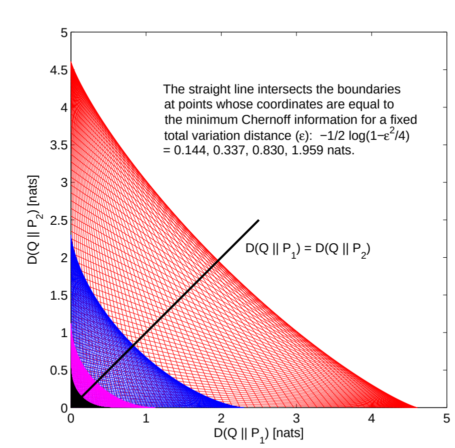

Figure: The boundary of the achievable region of $(D(Q\|P_1), D(Q\|P_2))$ where the TV distance $|P_1 - P_2|$ is at least $\varepsilon = 1.00, 1.40, 1.80, 1.98$.

Theorem: A New Upper Bound on the ML Decoding Error Probability

- Consider a binary linear block code of length N and rate $R = \frac{\log(M)}{N}$ where M designates the number of codewords.

Theorem: A New Upper Bound on the ML Decoding Error Probability

- Consider a binary linear block code of length N and rate $R = \frac{\log(M)}{N}$ where M designates the number of codewords.
- Let $S_0 = 0$ and, for $l \in \{1, \ldots, N\}$, let S_l be the number of non-zero codewords of Hamming weight l.

Theorem: A New Upper Bound on the ML Decoding Error Probability

- Consider a binary linear block code of length N and rate $R = \frac{\log(M)}{N}$ where M designates the number of codewords.
- Let $S_0 = 0$ and, for $l \in \{1, \ldots, N\}$, let S_l be the number of non-zero codewords of Hamming weight l.
- Assume that the transmission of the code takes place over a memoryless, binary-input and output-symmetric channel.

Theorem: A New Upper Bound on the ML Decoding Error Probability

- Consider a binary linear block code of length N and rate $R = \frac{\log(M)}{N}$ where M designates the number of codewords.
- Let $S_0 = 0$ and, for $l \in \{1, \ldots, N\}$, let S_l be the number of non-zero codewords of Hamming weight l.
- Assume that the transmission of the code takes place over a memoryless, binary-input and output-symmetric channel.
- Assume that the code is maximum-likelihood (ML) decoded.

Theorem: A New Upper Bound (Cont.)

The block error probability satisfies

$$P_{\mathsf{e}} = P_{\mathsf{e}|0} \leq \exp\left(-N \sup_{r \geq 1} \max_{0 \leq \rho \leq \frac{1}{r}} \left[E_0\left(\rho, \underline{q} = \left(\frac{1}{2}, \frac{1}{2}\right)\right) - \rho\left(rR + \frac{D_s(P_N \| Q_N)}{N}\right)\right]\right)$$

where

- $s \triangleq s(r) = \frac{r}{r-1}$ for $r \geq 1$ (with the convention that $s = \infty$ for $r = 1$),
- Q_N is the binomial distribution with parameter $\frac{1}{2}$ and N i.i.d. trials,
- P_N is the PMF defined by $P_N(l) = \frac{S_l}{M-1}$ for $l \in \{0, \ldots, N\}$,
- $D_s(\cdot \| \cdot)$ is the Rényi divergence of order s,
- $E_0(\rho, \underline{q})$ is the Gallager random coding error exponent.

Special Case: The Shulman-Feder Bound

Loosening the bound by taking $r = 1 \Rightarrow s = \infty$ gives

$$P_{\mathrm{e}} \leq \exp\left(-N\,E_{\mathrm{r}}\left(R + \frac{1}{N}\,\log \max_{0 \leq l \leq N} \frac{S_l}{e^{-N(\log 2 - R)}\binom{N}{l}}\right)\right)$$

which coincides with the Shulman-Feder bound.

Novelty of the Bound & Proof

- The proof of this theorem has an overlap with a bound by Shamai and Sason (2002).
- The bound is also valid for code ensembles, while referring to the average distance spectrum.
- The novelty is the use of the Rényi divergence of order $s \geq 1$, instead of the Kullback-Leibler divergence as a lower bound.
- This reveals a need for an optimization of the error exponent: If $r \geq 1$ is increased, $s = \frac{r}{r-1} \geq 1$ is decreased, and $D_s(P_N \| Q_N)$ is decreased (unless it is 0; note that P_N, Q_N do not depend on r, s).

Novelty of the Bound & Proof

- The proof of this theorem has an overlap with a bound by Shamai and Sason (2002).
- The bound is also valid for code ensembles, while referring to the average distance spectrum.
- The novelty is the use of the Rényi divergence of order $s \geq 1$, instead of the Kullback-Leibler divergence as a lower bound.
- This reveals a need for an optimization of the error exponent: If $r \geq 1$ is increased, $s = \frac{r}{r-1} \geq 1$ is decreased, and $D_s(P_N \| Q_N)$ is decreased (unless it is 0; note that P_N, Q_N do not depend on r, s).

Numerical Results

Numerical results for the binary-input AWGN channel support that the new bound provides an improvement over the Shulman-Feder bound. For high rate codes, there is an improvement over the tangential-sphere bound.

Full Paper Version

http://arxiv.org/abs/1501.03616.
Submitted to the *IEEE Trans. on Information Theory*, February 2015.