An Upper Bound on the ML Decoding Error Probability with the Rényi Divergence

Igal Sason

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

Jerusalem, Israel
April 26 - May 1, 2015

2015 IEEE Information Theory Workshop
(ITW 2015).

The Rényi Divergence

- Let P and Q be two probability mass functions defined on a set $\mathcal{X}$.
- Let $\alpha \in (0,1) \cup (1,\infty)$.

The Rényi divergence of order α is given by

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

The Rényi Divergence

- Let P and Q be two probability mass functions defined on a set $\mathcal{X}$.
- Let $\alpha \in (0,1) \cup (1,\infty)$.

The Rényi divergence of order α is given by

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

Extreme cases:

- If $\alpha = 0$ then $D_0(P||Q) = -\log Q(\text{Support}(P))$,
- If $\alpha = +\infty$ then $D_\infty(P||Q) = \log \left(\text{ess sup}\, \frac{P}{Q} \right)$,
- If $\alpha = 1$, it is defined to be $D(P||Q) = \sum P(x) \log \frac{P(x)}{Q(x)}$.

The Rényi Divergence

- Let P and Q be two probability mass functions defined on a set $\mathcal{X}$.
- Let $\alpha \in (0,1) \cup (1,\infty)$.

The Rényi divergence of order α is given by

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

Extreme cases:

- If $\alpha = 0$ then $D_0(P||Q) = -\log Q(\text{Support}(P))$,
- If $\alpha = +\infty$ then $D_\infty(P||Q) = \log \left(\text{ess sup} \frac{P}{Q} \right)$,
- If $\alpha = 1$, it is defined to be $D(P||Q) = \sum P(x) \log \frac{P(x)}{Q(x)}$.

If $D(P||Q) < \infty$, L'Hôpital's rule $\Rightarrow$ $D(P||Q) = \lim_{\alpha \to 1^-} D_\alpha(P||Q)$.

Some Basic Properties of the Rényi Divergences

1. Non-negativity: $D_\alpha(P||Q) \geq 0$ with equality if and only if $P = Q$.
2. Monotonicity: $D_\alpha(P||Q)$ is monotonic increasing in the parameter α.
3. Convexity properties of $D_\alpha(P||Q)$:
   - jointly convex in (P, Q) for $\alpha \in [0, 1]$,
   - convex in Q for $\alpha \in [0, \infty]$, but not in P for $\alpha > 1$.
   - jointly quasi-convex in (P, Q) for $\alpha \in [0, \infty]$.
4. The Rényi divergence satisfies the data processing inequality (DPI).

Some Basic Properties of the Rényi Divergences

1. Non-negativity: $D_\alpha(P||Q) \geq 0$ with equality if and only if $P = Q$.
2. Monotonicity: $D_\alpha(P||Q)$ is monotonic increasing in the parameter α.
3. Convexity properties of $D_\alpha(P||Q)$:
    - jointly convex in (P, Q) for $\alpha \in [0, 1]$,
    - convex in Q for $\alpha \in [0, \infty]$, but not in P for $\alpha > 1$.
    - jointly quasi-convex in (P, Q) for $\alpha \in [0, \infty]$.
4. The Rényi divergence satisfies the data processing inequality (DPI).

Paper

T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. on Information Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

Information-Theoretic Applications of the Rényi divergence

- Channel coding error exponents (Gallager '65, Arimoto '73, Polyanskiy & Verdu '10).
- Generalized cutoff rates for hypothesis testing (Csiszár '95, Alajaji et al. '04).
- Multiple source adaptation (Mansour et al., '09).
- Generalized guessing moments (van Erven & Harremoes, '10).
- Two-sensor composite hypothesis testing (Shayevitz, '11).
- Strong data processing theorems for discrete memoryless channels (Raginsky, '13).
- Strong converse theorems for classes of networks (Fong and Tan, arXiv '14).
- IT applications of the logarithmic probability comparison bound (Atar and Merhav, arXiv '15).

Motivation

- Performance analysis of linear codes under ML decoding is of interest for the study of their potential performance under optimal decoding.
- Also of interest for evaluating the degradation in performance that is incurred by sub-optimal & practical decoding algorithms.
- The new bound quantifies the degradation in performance of ML decoded block codes in terms of the deviation of their distance spectra from the binomial distribution (same as Shulman-Feder bound).
- Binomial distribution characterizes the average distance spectrum of the ensemble of fully random binary block codes, achieving the capacity of any memoryless binary-input 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

$$\begin{aligned} P_{\mathsf{e}} &= P_{\mathsf{e}|0} \\ &\leq \exp\left(-N\,E_{\mathsf{r}}\left(R + \frac{D_\infty(P_N\|Q_N)}{N}\right)\right) \\ &= \exp\left(-N\,E_{\mathsf{r}}\left(R + \frac{1}{N}\,\log \max_{0\leq l\leq N} \frac{P_N(l)}{Q_N(l)}\right)\right) \\ &= \exp\left(-N\,E_{\mathsf{r}}\left(R + \frac{1}{N}\,\log \max_{0\leq l\leq N} \frac{S_l}{e^{-N(\log 2-R)}\binom{N}{l}}\right)\right) \end{aligned}$$

which coincides with the Shulman-Feder bound.

Related Papers on Variations of the Gallager Bounds

1. S. Shamai and I. Sason, "Variations on the Gallager bounds, connections, and applications," *IEEE Trans. on Information Theory*, vol. 48, no. 12, pp. 3029–3051, December 2002.
2. I. Sason and S. Shamai, *Performance Analysis of Linear Codes under Maximum-Likelihood Decoding: A Tutorial, Foundations and Trends in Communications and Information Theory*, vol. 3, no. 1–2, pp. 1–222, NOW Publishers, Delft, the Netherlands, July 2006.

Novelty of this Proof

- The proof of this theorem has an overlap with Appendix A in the paper by Shamai and Sason (2002).
- Unlike the analysis there, working with the Rényi divergence of order $s \geq 1$, instead of the Kullback-Leibler divergence as a lower bound 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).
- The maximization of the error exponent in the theorem aims to find a proper balance between the two summands rR and $\frac{D_s(P_N \| Q_N)}{N}$ in the exponent of the new bound, while also optimizing $\rho' \in \left[0, \frac{1}{r}\right]$.

Applicability of the New Bound to Code Ensembles

The bound can be shown to be applicable to code ensembles of binary linear block codes:

- In the probability distribution P_N, the distance spectrum is replaced by the average distance spectrum of the ensemble.

Combination of the New Bound with an Existing Approach

- We borrow a concept of bounding by Miller and Burshtein, and propose to combine it with the new bound.
- In order to utilize the Shulman-Feder bound for binary linear block codes in a clever way, they partitioned the binary linear block code $\mathcal{C}$ into two subcodes $\mathcal{C}_1$ and $\mathcal{C}_2$ where

$$\mathcal{C}_1 \cup \mathcal{C}_2 = \mathcal{C}, \quad \mathcal{C}_1 \cap \mathcal{C}_2 = \{0\}.$$

- The subcode $\mathcal{C}_1$ contains the all-zero codeword and all the codewords of $\mathcal{C}$ whose Hamming weights l belong to a subset $\mathcal{L} \subseteq \{1, 2, ..., N\}$.
- The subcode $\mathcal{C}_2$ contains the other codewords of $\mathcal{C}$ (with Hamming weights of $l \in \mathcal{L}^{\mathrm{c}} \triangleq \{1, 2, ..., N\} \setminus \mathcal{L}$), and the all-zero codeword.

Idea in Selecting $\mathcal{C}_1$

Select $\mathcal{C}_1$ such that its distance spectrum is close to the binomial distribution:

$$P_N(l) \approx Q_N(l), \quad \forall\, l \in \mathcal{L}.$$

This selection implies that the normalized Rényi divergence $\frac{D_s(P_N \| Q_N)}{N}$ in the exponent of the new bound has a marginal effect on the conditional ML decoding error probability of the subcode $\mathcal{C}_1$.

Combination of the New Bound with an Existing Approach (Cont.)

- From the symmetry of the channel,

$$P_{\mathsf{e}} = P_{\mathsf{e}|0} \leq P_{\mathsf{e}|0}(\mathcal{C}_1) + P_{\mathsf{e}|0}(\mathcal{C}_2)$$

  where $P_{\mathsf{e}|0}(\mathcal{C}_1)$ and $P_{\mathsf{e}|0}(\mathcal{C}_2)$ are the conditional ML decoding error probabilities of $\mathcal{C}_1$ and $\mathcal{C}_2$ given that the zero codeword is transmitted.
- One can rely on different upper bounds on the conditional error probabilities $P_{\mathsf{e}|0}(\mathcal{C}_1)$ and $P_{\mathsf{e}|0}(\mathcal{C}_2)$:
  1. Bound $P_{\mathsf{e}|0}(\mathcal{C}_1)$ by invoking the new bound, due to the closeness of its distance spectrum to the binomial distribution.
  2. Rely on an alternative approach for bounding $P_{\mathsf{e}|0}(\mathcal{C}_2)$ (e.g., using the union bound w.r.t. the fixed composition codes of the subcode $\mathcal{C}_2$).

Example: Performance Bounds for an Ensemble of Turbo-Block Bodes

Consider

- An ensemble of uniformly interleaved turbo codes whose two component codes are chosen uniformly at random from the ensemble of (1072, 1000) binary systematic linear block codes.
- The overall code rate is 0.8741 bits per channel use.
- The transmission of these codes takes place over an additive white Gaussian noise (AWGN) channel.
- The codes are BPSK modulated and coherently detected.

Example: Turbo-block codes (Cont.)

The following upper bounds under ML decoding are compared:

- The tangential-sphere bound (TSB) of Herzberg and Poltyrev.
- The suggested combination of the union bound (UB) and the new bound (NB). An optimal partitioning is performed to obtain the tightest bound within this form.

Example: Turbo-block codes (Cont.)

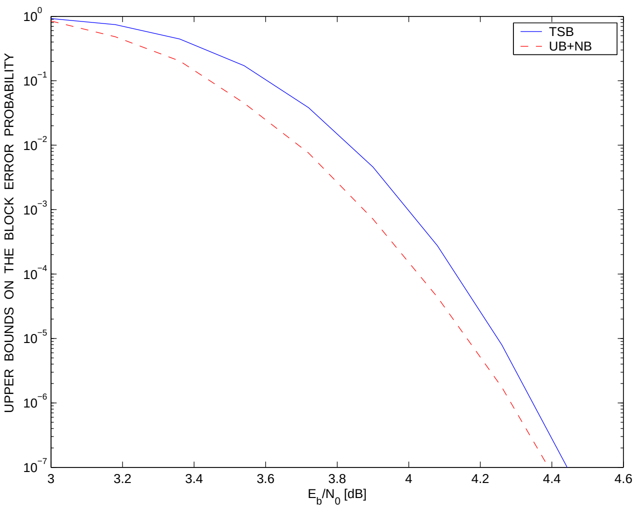

Figure: Comparison between upper bounds on the block error probability.

Summary

- A new bound on the ML decoding error probability has been derived, involving the Rényi divergence.
- The derivation of this bound relies on variations of the Gallager bounds (the Duman and Salehi bound).
- It reproduces the 1965 random coding Gallager bound, and the Shulman-Feder bound for binary linear block codes (or ensembles).
- It has an additional parameter that is subject to optimization.
- The bound has been applied to an ensemble of uniformly interleaved turbo-block codes with systematic random component codes, and its superiority has been exemplified.