On f- and Rényi Divergences

Igal Sason (Technion, Israel)

Part of this talk relies on a joint work with
Sergio Verdú (Princeton, NJ, USA)

Seminar Talk in Communications and Information Theory
Electrical Engineering Department, Technion
December 22, 2016

Motivation for talking about f-divergences

Divergences measure the dissimilarity between probability measures, and many metrics fall under the paradigm of an f-divergence (to be defined).

Motivation for talking about f-divergences

Divergences measure the dissimilarity between probability measures, and many metrics fall under the paradigm of an f-divergence (to be defined).

$\Rightarrow$ f-divergences are useful in, e.g.,

- Ensuring convergence of prob. measures with various metrics
- Evaluation of rates of convergence
- Transportation-cost inequalities (e.g., Pinsker's inequality)
- Estimation and modeling (e.g., bounds on the minimax risk)
- Connection to mutual information, capacity, and Arimoto information
- Hypothesis testing in the Bayesian setup
- Non-asymptotic concentration of measure inequalities
- Inequalities related to strong data processing and maximal correlation
- One-shot achievability results via fidelity
- Numerous other applications

Motivation for talking about f-divergences here

Divergences measure the dissimilarity between probability measures, and many metrics fall under the paradigm of an f-divergence (to be defined).

$\Rightarrow$ f-divergences are useful in, e.g.,

- Ensuring convergence of prob. measures with various metrics
- Evaluation of rates of convergence
- Transportation-cost inequalities (e.g., Pinsker's inequality)
- Estimation and modeling (e.g., bounds on the minimax risk)
- Connection to mutual information, capacity, and Arimoto information
- Hypothesis testing in the Bayesian setup
- Non-asymptotic concentration of measure inequalities
- Inequalities related to strong data processing and maximal correlation
- One-shot achievability results via fidelity
- Numerous other applications

f-Divergence

Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let $P \ll Q$. The *f-divergence* from P to Q is given by

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{1}$$

If $P, Q \ll \mu$, $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$, then

$$D_f(P\|Q) = \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu. \tag{2}$$

f-Divergence

f-divergences were suggested independently by Ali & Silvey (1966), Csiszár (1963), and Morimoto (1963):

- S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, series B, vol. 28, no. 1, pp. 131–142, 1966.
- I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, Jan. 1963.
- T. Morimoto, "Markov processes and the H-theorem," *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, March 1963.

Examples of f-divergences

- Relative entropy

$$D(P\|Q) = D_f(P\|Q) \tag{3}$$

where

$$f(t) = t \log t, \quad t > 0. \tag{4}$$

Examples of f-divergences

- Relative entropy

$$D(P\|Q) = D_f(P\|Q) \tag{3}$$

where

$$f(t) = t\log t, \quad t > 0. \tag{4}$$

- Relative entropy

$$D(Q\|P) = D_f(P\|Q) \tag{5}$$

where

$$f(t) = -\log t, \quad t > 0. \tag{6}$$

Examples of f-divergences (cont.)

- χ^2-divergence: $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$,

$$\chi^2(P\|Q) = D_f(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q. \tag{7}$$

Examples of f-divergences (cont.)

- χ^2-divergence: $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$,

$$\chi^2(P\|Q) = D_f(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q. \tag{7}$$

- Total variation (TV) distance: Setting $f(t) = |t-1|$ results in

$$|P - Q| = D_f(P\|Q) \tag{8}$$

$$= \int \left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right| \mathrm{d}Q \tag{9}$$

$$= 2 \sup_{\mathcal{F} \in \mathscr{F}} \big(P(\mathcal{F}) - Q(\mathcal{F})\big). \tag{10}$$

Examples of f-divergences (cont.)

Hellinger divergence of order $\alpha \in (0,1) \cup (1,\infty)$:

$$\mathscr{H}_\alpha(P\|Q) = D_{f_\alpha}(P\|Q) \tag{11}$$

with $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for $t \geq 0$.

- The χ^2-divergence is the Hellinger divergence of order 2;
- Continuous extension at $\alpha = 1$ yields

$$\mathscr{H}_1(P\|Q)\,\log e = D(P\|Q). \tag{12}$$

Basic Property 1: Reflexivity Property

If $f\colon (0,\infty) \to \mathbb{R}$ is convex and $f(1) = 0$, $P \ll Q$, then

$$D_f(P\|Q) \geq 0. \tag{13}$$

If f is strictly convex at $t = 1$, then $D_f(P\|Q) = 0$ yields $P = Q$.

Basic Property 2: Uniqueness Theorem

Let f and g be convex functions on $(0, \infty)$ with $f(1) = g(1) = 0$. Then, the following conditions are equivalent:

1. $$D_f(P\|Q) = D_g(P\|Q), \quad \forall\, P, Q; \tag{14}$$

2. there exists $c \in \mathbb{R}$ such that

$$f(t) - g(t) = c(t-1), \quad \forall\, t \in \mathbb{R}. \tag{15}$$

In other words, f and g-divergences are identical if and only if f and g differ by a linear function that vanishes at 1.

Basic Property 3

An f-divergence satisfies the following three properties:

1. If P and Q are probability measures defined on $\mathcal{X}$, then $D_f(P\|Q)$ is invariant under permutations on $\mathcal{X}$.
2. It satisfies the data processing inequality: Let $\mathcal{A} = \{A_i,\, i \geq 1\}$ be any partition of $\mathcal{X}$, and let $\mathcal{P}_{\mathcal{A}} = \{P(A_i),\, i \geq 1\}$ and $\mathcal{Q}_{\mathcal{A}} = \{Q(A_i),\, i \geq 1\}$. Then

$$D_f(P\|Q) \geq D_f(P_{\mathcal{A}}\|Q_{\mathcal{A}}). \tag{16}$$

3. Convexity: For all $P_1, P_2, Q_1, Q_2 \in \mathcal{P}$ and $\lambda \in [0,1]$, $\bar{\lambda} \triangleq 1-\lambda$

$$D_f\big(\lambda P_1 + \bar{\lambda} P_2 \| \lambda Q_1 + \bar{\lambda} Q_2\big) \leq \lambda D_f(P_1\|Q_1) + \bar{\lambda} D_f(P_2\|Q_2). \tag{17}$$

Basic Property 4: Symmetry Theorem

Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let $f^*\colon (0,\infty) \to \mathbb{R}$ be the $*$-conjugate, given by

$$f^*(t) = t\, f\left(\tfrac{1}{t}\right), \quad t > 0. \tag{18}$$

Then,

- f^* is also convex, and $f^*(1) = 0$,
- if $P \ll\gg Q$, then

$$D_f(P\|Q) = D_{f^*}(Q\|P). \tag{19}$$

By definition, we take

$$f^*(0) = \lim_{t\downarrow 0} f^*(t) = \lim_{u\to\infty} \frac{f(u)}{u}. \tag{20}$$

f-divergence Inequalities

Theorem 1: Functional Domination

Let $P \ll Q$, and assume

- f and g are convex on $(0, \infty)$ with $f(1) = g(1) = 0$;
- $g(t) > 0$ for all $t \in (0, 1) \cup (1, \infty)$.

Denote the function $\kappa \colon (0, 1) \cup (1, \infty) \to \mathbb{R}$

$$\kappa(t) = \frac{f(t)}{g(t)}, \quad t \in (0, 1) \cup (1, \infty) \tag{21}$$

and

$$\bar{\kappa} = \sup_{t \in (0,1) \cup (1,\infty)} \kappa(t). \tag{22}$$

Theorem 1 (cont.)

Then,

a)

$$D_f(P\|Q) \leq \bar{\kappa}\, D_g(P\|Q). \tag{23}$$

b) If, in addition, $f'(1) = g'(1) = 0$, then

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{D_g(P\|Q)} = \bar{\kappa}. \tag{24}$$

Range of Values Theorem for f-Divergences

- Vajda showed that the range of an f-divergence is given by

$$0 \leq D_f(P\|Q) \leq f(0) + f^*(0) \tag{25}$$

  where every value in this range is attainable by a suitable pair of probability measures $P \ll Q$.

- Basu *et al.* strengthened Vajda's result, showing that

$$D_f(P\|Q) \leq \tfrac{1}{2}\big(f(0) + f^*(0)\big)\,|P - Q|. \tag{26}$$

f-Divergences (cont.)

Using the theorem on functional domination, we assert that the constant in (26) cannot be improved.

Theorem 2

If $f : (0, \infty) \to \mathbb{R}$ is convex with $f(1) = 0$, then

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{|P - Q|} = \tfrac{1}{2}\big(f(0) + f^*(0)\big) \tag{27}$$

where the supremum is over all probability measures P, Q such that $P \ll Q$ and $P \neq Q$.

Definition: Relative Information

If $P \ll Q$, the relative information provided by $a \in \mathcal{A}$ according to (P, Q) is given by

$$\imath_{P\|Q}(a) \triangleq \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a). \tag{28}$$

Definition: Relative Information

If $P \ll Q$, the relative information provided by $a \in \mathcal{A}$ according to (P, Q) is given by

$$\imath_{P\|Q}(a) \triangleq \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a). \tag{28}$$

Relative entropy

The relative entropy of P with respect to Q is

$$D(P\|Q) = \mathbb{E}\big[\imath_{P\|Q}(X)\big] \tag{29}$$

$$= \mathbb{E}\big[\imath_{P\|Q}(Y)\, \exp\big(\imath_{P\|Q}(Y)\big)\big], \tag{30}$$

where $X \sim P$ and $Y \sim Q$.

Rényi Divergence

Let $P \ll Q$. The Rényi divergence $D_\alpha(P\|Q)$ is given as follows:

- If $\alpha \in (0,1) \cup (1,\infty)$, then with $X \sim P$

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\Big(\mathbb{E}\big[\exp\big((\alpha-1)\, \imath_{P\|Q}(X)\big)\big]\Big). \tag{31}$$

- If $\alpha = 0$, then

$$D_0(P\|Q) = \max_{\mathcal{F} \in \mathscr{F}:\ P(\mathcal{F})=1} \log\left(\frac{1}{Q(\mathcal{F})}\right). \tag{32}$$

- If $\alpha = 1$, then $D_1(P\|Q) = D(P\|Q)$.
- If $\alpha = +\infty$ then with $Y \sim Q$

$$D_\infty(P\|Q) = \log\left(\operatorname{ess\,sup} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right). \tag{33}$$

Rényi Divergence (Cont.)

Rényi divergence is a one-to-one transformation of Hellinger divergence of the same order $\alpha \in (0,1) \cup (1,\infty)$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\left(1 + (\alpha - 1)\, \mathscr{H}_\alpha(P\|Q)\right) \tag{34}$$

$$\Rightarrow D_2(P\|Q) = \log\left(1 + \chi^2(P\|Q)\right). \tag{35}$$

Rényi Divergence (Cont.)

Rényi divergence is a one-to-one transformation of Hellinger divergence of the same order $\alpha \in (0,1) \cup (1,\infty)$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\left(1 + (\alpha - 1)\, \mathscr{H}_\alpha(P\|Q)\right) \tag{34}$$

$$\Rightarrow D_2(P\|Q) = \log\left(1 + \chi^2(P\|Q)\right). \tag{35}$$

Connection Between Rényi and f-divergences

$\Rightarrow$ The Rényi divergence is not an f-divergence, but it is nevertheless a one-to-one transformation of an f-divergence.

Definition of β_1 and β_2.

Given a pair of probability measures (P, Q) on the same measurable space, denote $\beta_1, \beta_2 \in [0, 1]$ by

$$\beta_1 = \exp\big(-D_\infty(P\|Q)\big), \tag{36}$$

$$\beta_2 = \exp\big(-D_\infty(Q\|P)\big) \tag{37}$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$.

Definition of β_1 and β_2.

Given a pair of probability measures (P, Q) on the same measurable space, denote $\beta_1, \beta_2 \in [0, 1]$ by

$$\beta_1 = \exp\bigl(-D_\infty(P\|Q)\bigr), \tag{36}$$

$$\beta_2 = \exp\bigl(-D_\infty(Q\|P)\bigr) \tag{37}$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$.

- if $\beta_1 > 0$, then $P \ll Q$, while $\beta_2 > 0$ implies $Q \ll P$.
- if $P \lll\ggg Q$, then with $Y \sim Q$,

$$\beta_1 = \operatorname{ess\,inf} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right)^{-1}, \tag{38}$$

$$\beta_2 = \operatorname{ess\,inf} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y)\right)^{-1}. \tag{39}$$

Since $\beta_1 = 1 \Leftrightarrow \beta_2 = 1 \Leftrightarrow P = Q$, we avoid trivialities by excluding that case.

Theorem 3: Bounded Relative Information

Let f and g satisfy the assumptions in Theorem 1, and assume that $(\beta_1, \beta_2) \in [0, 1)^2$. Then,

$$D_f(P\|Q) \leq \kappa^* \, D_g(P\|Q) \tag{40}$$

where

$$\kappa^* = \sup_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \tag{41}$$

and $\kappa(\cdot)$ is defined in Theorem 1.

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$D(P\|Q) \geq \tfrac{1}{2}\,|P-Q|^2\,\log e \tag{42}$$

and the constant is tight in the sense that

$$\inf_{P\neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \tfrac{1}{2}\log e.$$

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$D(P\|Q) \geq \tfrac{1}{2}\,|P-Q|^2\,\log e \tag{42}$$

and the constant is tight in the sense that

$$\inf_{P\neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \tfrac{1}{2}\log e.$$

An Implication of Pinsker's Inequality

Convergence in relative entropy $\Longrightarrow$ convergence in TV distance.

Question

Is there a reverse Pinsker inequality that provides an upper bound on the relative entropy as a function of the TV distance ?

Question

Is there a reverse Pinsker inequality that provides an upper bound on the relative entropy as a function of the TV distance ?

No, for every $\varepsilon > 0$ there exist P, Q s.t. $|P - Q| \leq \varepsilon$, $D(P\|Q) = \infty$.

Question

Is there a reverse Pinsker inequality that provides an upper bound on the relative entropy as a function of the TV distance ?

No, for every $\varepsilon > 0$ there exist P, Q s.t. $|P - Q| \leq \varepsilon$, $D(P\|Q) = \infty$.

However, we can obtain a reverse Pinsker inequality when the relative information is bounded (see next theorem).

Application – Theorem 4: A Reverse Pinsker inequality

If $\beta_1 \in (0,1)$ and $\beta_2 \in [0,1)$, then,

$$D(P\|Q) \leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)\,|P-Q| \tag{43}$$

where $\varphi\colon [0,\infty) \to [0,\infty)$ is given by

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \frac{t\log t}{t-1} & t \in (0,1)\cup(1,\infty) \\ \log e & t = 1. \end{cases} \tag{44}$$

Reverse Pinsker Inequality for Finite Alphabet [Csiszár & Talata, '06]

If $\mathcal{A}$ is a finite set, P and Q are probability measures defined on $\mathcal{A}$, and $Q_{\min} \triangleq \min_{x \in \mathcal{A}} Q(x) > 0$, then

$$D(P\|Q) \leq \frac{\log e}{Q_{\min}} \cdot |P - Q|^2. \tag{45}$$

Recent Applications of (45)

- I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. on IT*, vol. 52, no. 3, pp. 1007–1016, Mar. 2006.
- V. Kostina and S. Verdú, "Channels with cost constraints: strong converse and dispersion," *IEEE Trans. on IT*, vol. 61, no. 5, pp. 2415–2429, May 2015.
- M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. on IT*, vol. 59, no. 11, pp. 7041–7051, Nov. 2013.

Theorem 5: New Reverse Pinsker Inequality for Finite Alphabet

a) If $P \ll Q$

$$D(P\|Q) \leq \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}}\right). \tag{46}$$

b) Furthermore, if $Q \ll P$ and $\beta_2 \in [0,1]$ is given by

$$\beta_2 = \min_{x\in\mathcal{A}} \frac{P(x)}{Q(x)}$$

then the following tightened bound holds:

$$D(P\|Q) \leq \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}}\right) - \frac{\beta_2 \log e}{2} \cdot |P-Q|^2. \tag{47}$$

Note on Theorem 5a)

This already improves the Csiszár-Talata inequality since

$$\log\left(1+\frac{|P-Q|^2}{2Q_{\min}}\right) \leq \frac{\log e}{\mathbf{2}Q_{\min}} \cdot |P-Q|^2.$$

Proof of Theorem 5a)

The idea is to obtain upper and lower bounds on the χ^2-divergence

$$\chi^2(P,Q) \triangleq \sum_{x\in\mathcal{A}} \frac{(P(x)-Q(x))^2}{Q(x)}.$$

Proof of Theorem 5a)

The idea is to obtain upper and lower bounds on the χ^2-divergence

$$\chi^2(P,Q) \triangleq \sum_{x\in\mathcal{A}} \frac{(P(x)-Q(x))^2}{Q(x)}.$$

Bounds on the χ^2-divergence

$$\chi^2(P\|Q) \geq e^{D(P\|Q)} - 1 \quad \text{(via Jensen's inequality)}$$

$$\chi^2(P\|Q) \leq \frac{\sum_{x\in\mathcal{A}}\big(P(x)-Q(x)\big)^2}{Q_{\min}} \leq \frac{|P-Q|^2}{2Q_{\min}}.$$

Combining the bounds yields Theorem 5a).

Definition: Relative Information Spectrum

The relative information spectrum is the cumulative distribution function

$$\mathbb{F}_{P\|Q}(x) = \mathbb{P}\big[\imath_{P\|Q}(X) \leq x\big], \tag{48}$$

with $X \sim P$.

New Integral Representations with the Relative Information Spectrum

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\left((1-\alpha)\int_0^\infty \beta^{\alpha-2}\,\mathbb{F}_{P\|Q}(\log\beta)\,\mathrm{d}\beta\right),\ \alpha\in(0,1),$$

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\left((\alpha-1)\int_0^\infty \beta^{\alpha-2}\big(1-\mathbb{F}_{P\|Q}(\log\beta)\big)\,\mathrm{d}\beta\right),\ \alpha>1$$

$$D(P\|Q) = \int_1^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{\beta}\,\mathrm{d}\beta - \int_0^1 \frac{\mathbb{F}_{P\|Q}(\log\beta)}{\beta}\,\mathrm{d}\beta,$$

$$\chi^2(P\|Q) = \int_0^\infty \big(1-\mathbb{F}_{P\|Q}(\log\beta)\big)\,\mathrm{d}\beta - 1$$

$$|P-Q| = 2\int_1^\infty \frac{1-\mathbb{F}_{P\|Q}(\log\beta)}{\beta^2}\,\mathrm{d}\beta$$

Information-Theoretic Applications of the Rényi divergence

- Channel coding error exponents.
- Generalized cutoff rates for hypothesis testing.
- Multiple source adaptation.
- Generalized guessing moments.
- Two-sensor composite hypothesis testing.
- Bounds for joint source-channel coding.
- Strong data processing theorems for DMCs.
- Strong converse theorems for networks.
- IT applications of the logarithmic probability comparison bound.

A Coding Theorem with the Rényi Divergence

Motivation for Part II of the Talk

- Performance analysis of linear codes under ML decoding is of interest for the study of the potential performance of these codes under optimal decoding.
- It is also of interest for the evaluation of the degradation in performance that is incurred by the use of sub-optimal and practical decoding algorithms.
- Similarly to the Shulman-Feder bound and related studies, the upper bound in the following theorem quantifies the degradation in the performance of block codes under ML decoding in terms of the deviation of their distance spectra from the binomial distribution.
- The latter 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.

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

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

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 the Bound & Proof

- The proof of this theorem has an overlap with Appendix A in the paper by Shamai and Sason (2002).
- The novelty here is in working with the Rényi divergence of order $s \geq 1$, instead of the relative entropy as a lower bound, reveals a need for an optimization of the error exponent:
  1. 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).
  2. 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 it includes the codewords whose hamming weights correspond to the portion of the distance spectrum which 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.)

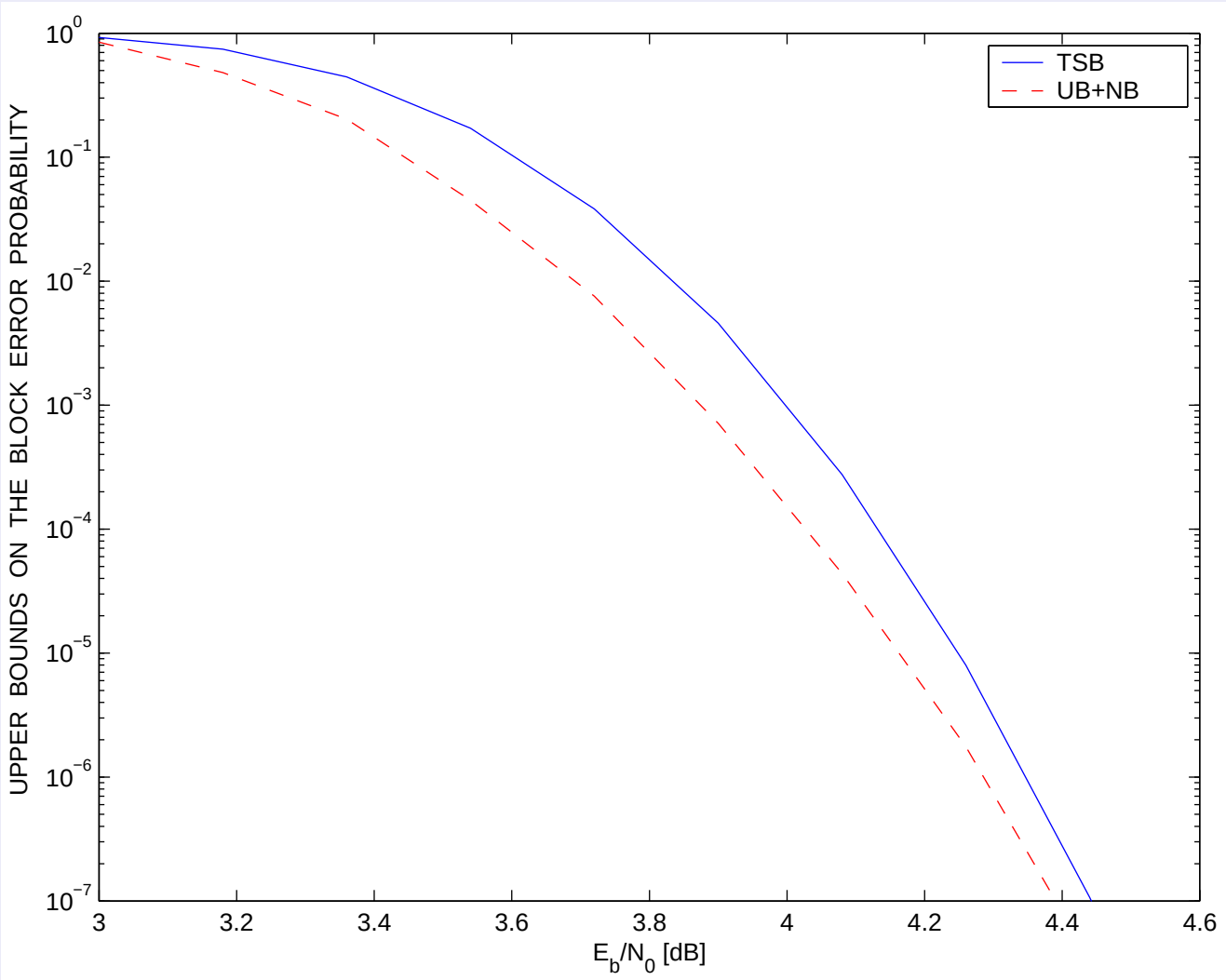

Figure: Comparison between upper bounds on the block error probability.

Summary (Part I)

- f-divergences have been presented.
- Some approaches to derive f-divergence inequalities have been introduced.
- Reverse Pinsker inequalities have been introduced.
- Expressions of f-divergence in terms of the relative information spectrum have been presented.

Summary (Part II)

- A new bound on the ML decoding error probability has been derived, involving the Rényi divergence.
- It reproduces the random coding bound (Gallager, 1965), and the Shulman-Feder bound for binary linear block codes (or ensembles).
- This bound has in general an additional parameter that is subject to optimization (the order of the Rényi divergence).
- Applicable to code ensembles, and its superiority has been exemplified for an ensemble of turbo-block codes.

Journal Papers

I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. on Information Theory*, vol. 62, no. 11, pp. 5973–6006, November 2016.

I. Sason, "On the Rényi divergence, joint range of relative entropies, and a channel coding theorem," *IEEE Trans. on Information Theory*, vol. 62, no. 1, pp. 23–34, January 2016.