Inequalities in Statistics and Information Measures

Christos P. Kitsos and Thomas L. Toulias

Abstract This paper presents and discusses a number of inequalities in the area of two distinct mathematical branches, with not that different line of thought: Statistics and Mathematical Information, which apply different "measures" to analyze the collected data. In principle, in these two fields, inequalities appear either as bounds in different measures or when different measures are compared. We discuss both and we prove new bounds for the Kullback–Leibler relative entropy measure, when the Generalized Normal distribution is involved.

1 Introduction

Inequalities play an important role in Mathematical Sciences. Provides bounds to the existing calculations, or even to the non-existing ones: we may not know the exact closed expression of a mathematical expression, but it is often possible to know the corresponding boundaries. Typical example is the bound of the n roots of an n-th degree polynomial, say

$$P_n(x) = a_0 + a_1\, x + a_2\, x^2 + \cdots + a_n\, x^n, \quad x \in \mathbb{C}.$$

Then, for $n > 4$, it is well known that there are no closed algebraic forms describing the root values, but we can have certain bounds for them. Indeed, if

$$A := \max\left\{|a_0|, |a_1|, \ldots, |a_{n-1}|\right\} \text{ and } B := \left\{|a_n|, |a_{n-1}|, \ldots, |a_1|\right\},$$

C. P. Kitsos (✉) · T. L. Toulias
University of West Attica, Egaleo, Athens, Greece
e-mail: xkitsos@uniwa.gr; th.toulias@uniwa.gr

D. Andrica, T. M. Rassias (eds.), *Differential and Integral Inequalities*,
Springer Optimization and Its Applications 151,
https://doi.org/10.1007/978-3-030-27407-8_16

then for the k-th root x_k, $k = 1, 2, \ldots, n$, it holds that

$$r := \frac{1}{1 + \frac{B}{|a_0|}} < |x_k| < 1 + \frac{A}{|a_n|} =: R.$$

Therefore, the roots x_k, $k = 1, 2, \ldots, n$, lie within the set-difference of the circles $C(O, R)$ and $C(O, r)$, i.e. $x_k \in C(O, R) \setminus C(O, r)$, $k = 1, 2, \ldots, n$.

Moreover, if the highest-order coefficient of the polynomial P_n as above is non-negative, i.e. $a_n > 0$, and $\delta := \max\left\{|a_k|, k \in \{0, 1, \ldots, n\} : \ a_k < 0\right\}$, then—according to the Lagrange theory for the positive roots of P_n—it holds that

$$0 < x_k \leq 1 + \sqrt[p]{\delta / a_n}, \quad k \in K \subseteq \{1, 2, \ldots, n\},$$

where p declares the position of the highest-order negative coefficient of P_n.

Inequalities appear in almost all the subject fields of Mathematics. The following Sect. 2 presents some classical inequalities in Mathematics, while Sect. 3 demonstrates the importance of inequalities in Statistics. Section 4 discusses certain inequalities that appear in Probability Theory. Section 5 shows some of the most important inequalities in Information Theory, while Sect. 6 briefly introduces the generalized Normal distribution and its relation to a generalized form of the logarithm Sobolev inequality, and to information measures in general. Finally, Sect. 7 proves and discusses some inequalities derived from the study of the information divergence between two generalized forms of the multivariate Normal distribution.

2 Fundamental Inequalities in Mathematics

Some of the main, in our opinion, inequalities widely used in Mathematics are presented in the following.

- *The Cauchy–Schwarz inequality.* Let f and g be two real functions defined on the interval $[a, b]$. Then, their inner product is defined to be

$$\langle f, g \rangle := \int_a^b f(x)\, g(x)\, w(x)\, \mathrm{d}x, \quad w(x) \geq 0.$$

  The well-known Cauchy–Schwarz inequality is then formulated as

$$\langle f, g \rangle^2 \leq \langle f, f \rangle \langle g, g \rangle, \ \text{ or } \ \langle f, g \rangle \geq \|f\| \, \|g\|.$$

  When f and g assumed to be n-dimensional vectors $\mathbf{a} := (a_i), \mathbf{b} := (b_i) \in \mathbb{R}^n$ and $w \equiv 1$, their inner product is then given by the finite sum $\langle \mathbf{a}, \mathbf{b} \rangle = a_1\, b_1 +$

$a_2\, b_2 + \cdots + a_n\, b_n$. As a result, the corresponding Cauchy–Schwarz inequality can then be written as $\big|\langle \mathbf{a}, \mathbf{b}\rangle\big|/(\|\mathbf{a}\|\,\|\mathbf{b}\|) \le 1$.

- *The determinant inequality.* From Linear Algebra, it is known that the determinant of a square real matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is bounded. Indeed,

$$\frac{n}{\mathrm{tr}(\mathbf{A}^{-1})} \le (\det \mathbf{A})^{1/n} \le \tfrac{1}{n}\,\mathrm{tr}(\mathbf{A}).$$

- *The triangle inequality.* In Euclidian Plane Geometry, for every three non-collinear points A, B, and C, forming the triangle ABC, it holds that $|\overline{AC}| < |\overline{AB}| + |\overline{BC}|$, which is known as the *triangle inequality*. Considering now the Euclidian p-dimensional space, equipped with the usual Euclidian metric/norm, i.e. $\|\mathbf{a}\|^2 := a_1^2 + a_2^2 + \cdots + a_p^2$, $\mathbf{a} = (a_i) \in \mathbb{R}^p$, the triangle inequality holds, formulated as $\|\mathbf{a} + \mathbf{b}\| \le \|\mathbf{a}\| + \|\mathbf{b}\|$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. This is also one of the most widely known inequalities in Analytic/Convex Geometry as well as in the study of metric spaces.
- *The Minkowski inequality.* Triangle inequality can be considered as a special case of the Minkowski inequality $\|f + g\|_p \le \|f\|_p + \|g\|_p$, $f, g \in \mathscr{L}^p(S)$, where S is a metric space with measure μ with $f + g \in \mathscr{L}^p(S)$, and where the p-norm $\|\cdot\|_p$ is defined as $\|f\|_p^p := \int |f|^p \,\mathrm{d}\mu$; see [30] among others. The equality holds for $f := \lambda\, g$, $\lambda \in \mathbb{R}^+$, or when $g \equiv 0$. Finally, if we are considering vectors, the Minkowski inequality is reduced to $\|\mathbf{a}+\mathbf{b}\|_p \le \|\mathbf{a}\|_p + \|\mathbf{b}\|_p$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, for the non-Euclidian p-norm $\|\mathbf{a}\|_p^p := |a_1|^p + |a_2|^p + \cdots + |a_p|^p$, $\mathbf{a} = (a_i) \in \mathbb{R}^p$.
- *Factorial bounds.* Two interesting inequalities are known as the lower and upper bounds for the factorial, i.e.

$$\left(\tfrac{n}{\mathrm{e}}\right)^n \sqrt{2\pi\, n} \le n! \le \mathrm{e}\left(\tfrac{n}{\mathrm{e}}\right)^n \sqrt{n}, \quad n \in \mathbb{N}, \tag{1}$$

or, generalizing via the Gamma function,

$$\left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{2\pi\, x} \le \Gamma(x+1) \le \mathrm{e}\left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{x}, \quad x \in \mathbb{R}^+. \tag{2}$$

Recall that the lower boundary of (1) is the well-known Stirling's approximation formula, $n! \overset{\text{asym.}}{\approx} (n/\mathrm{e})^n \sqrt{2\pi\, n}$, meaning that the quantities $n!$ and $(n/\mathrm{e})^n \sqrt{2\pi\, n}$ are asymptotically convergent. Historically speaking, the Stirling's formula was first introduced by Abraham de Moivre in the form of $n! \sim (\text{const.})\,(n/\mathrm{e})^n \sqrt{n}$, and later James Stirling evaluated the constant to be $\sqrt{2\pi}$. Note that the bounds in (1) shall be used later in Sect. 5. More precise bounds introduced by Robbins in [39] were formulated as

$$\mathrm{e}^{\frac{1}{12n+1}} \left(\tfrac{n}{\mathrm{e}}\right)^n \sqrt{2\pi\, n} < n! < \mathrm{e}^{\frac{1}{12n}} \left(\tfrac{n}{\mathrm{e}}\right)^n \sqrt{2\pi\, n}, \quad n \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}. \tag{3}$$

Finally, Srinivasa Ramanujan, in his lost notebook, [36] provided some alternative bounds for the Gamma function, in the form of

$$\left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{\pi}\sqrt[6]{8x^3+4x^2+x+\tfrac{1}{100}} < \Gamma(x+1) < \left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{\pi}\sqrt[6]{8x^3+4x^2+x+\tfrac{1}{30}}, \quad x \in \mathbb{R}^+,$$

while Mortici proved in [33], some even stricter bounds for the Gamma function when $x \geq 8$, i.e.

$$\left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{\pi}\sqrt[6]{8x^3+4x^2+x+\tfrac{1}{30}-\tfrac{11}{240x}} < \Gamma(x+1) < \left(\tfrac{x}{\mathrm{e}}\right)^x \sqrt{\pi}\sqrt[6]{8x^3+4x^2+x+\tfrac{1}{30}-\tfrac{1}{24x}},$$

although the lower boundary actually holds for $x \geq 2$.

- *Rayleigh quotient.* Consider the Rayleigh quotient

$$R(\mathbf{A}) = R(\mathbf{A};\ \mathbf{x}) := \frac{\mathbf{x}^{\mathrm{H}}\mathbf{A}\,\mathbf{x}}{\mathbf{x}^{\mathrm{H}}\mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\},$$

for the complex Hermitian (or self-adjoint) matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, i.e. when $\mathbf{A} = \mathbf{A}^{\mathrm{H}}$, where $\mathbf{A}^{\mathrm{H}}$ denotes the *conjugate transpose* of matrix/vector $(a_{ij}) = \mathbf{A} \in \mathbb{C}^{m\times n}$, i.e. $\mathbf{A} = \mathbf{A}^{\mathrm{H}} := \overline{\mathbf{A}^{\mathrm{T}}} = (\overline{a_{ji}})$. For the case of a Hermitian (or real symmetric) matrix $\mathbf{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, it holds $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ (symmetricity), while $\lambda_1 = \max_{\mathbf{x}}\left\{R(\mathbf{x})\right\}$ and $\lambda_n = \min_{\mathbf{x}}\left\{R(\mathbf{x})\right\}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the n real eigenvalues of matrix $\mathbf{A}$.
- *Error bounds.* In all the approximation problems, there are bounds for the existing errors. Practically speaking, for the exact solution $y(x_k)$ at point x_k, the total truncation error ϵ_k is then given by $\epsilon_k = y_k - y(x_k)$, where y_k is the exact value (corresponding to x_k) which would be resulting from an algorithm. We usually calculate some value, say y_k^*, which approximates the exact y_k value, and thus the corresponding rounding error ϵ_k^* is $\epsilon_k^* = y_k^* - y_k$. Therefore, the total error r_k is given by $|r_k| \leq |\epsilon_k| + |\epsilon_k^*|$. Both the forms of truncation error and the propagation error need particular inequalities; see [13, 17].
- *Error control.* When the simultaneous equations $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ are asked to be solved, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\det\mathbf{A} \neq \mathbf{0}$, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^{n\times 1}$, errors may occur in both left- and right-hand side. These equations can then be written as $(\mathbf{A} + \delta\,\mathbf{A})\,(\mathbf{x} + \delta\,\mathbf{x}) = \mathbf{b} + \delta\,\mathbf{b}$. Froberg in [13] calculated the relative error $\epsilon_{\mathbf{x}}$ of the solution $\mathbf{x}$ and proved that it is bounded, i.e.

$$\epsilon_{\mathbf{x}} \leq \frac{c}{1 - c\epsilon_{mathbf A}}(\epsilon_{\mathbf{A}} + \epsilon_{\mathbf{b}}),$$

where $c := \mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ is the conditional number of matrix $\mathbf{A}$, and the corresponding relative errors of $\mathbf{A}$, $\mathbf{x}$, and $\mathbf{b}$ are given, respectively, by $\epsilon_{\mathbf{A}} := \|\delta\,\mathbf{A}\|/\|\mathbf{A}\|$, $\epsilon_{\mathbf{x}} := \|\delta\,\mathbf{x}\|/\|\mathbf{x}\|$ and $\epsilon_{\mathbf{b}} := \|\delta\,\mathbf{b}\|/\|\mathbf{b}\|$. For a number of evaluated bounds in Numerical Analysis, see [17] among others.

- *Stochastic approximation*. For the numerical solution of an equation, there are numerous methods in literature, with the most popular being the Newton–Raphson (and its various forms) and, alternatively, the bisection method (in cases where the differentiation is fairly complicated or not available). The stochastic approximation method, [17], introduced by Robbins and Monro in [40] provides a statistical iterative approach for the solution of $\mathbf{M}(\mathbf{x}) = \boldsymbol{\theta}$, and evaluates maximum or minimum of a function, since the problem cannot adopt the line of thought in [37]. If we assume that an experiment is performed with response y at point x, i.e. $y = Y(x)$, and probability $H(y|x) := \Pr(Y(x) \leq y)$ with expected value of random variable (r.v.) X (which measures x) of the form $\mathrm{E}(X) = \int_{\mathbb{R}} y \, \mathrm{d}H(y|x)$, it is then asked to solve the equation $\mathbf{M}(\mathbf{x}) = \boldsymbol{\theta}$. Under a certain number of restrictions (i.e. inequalities), the sequence $x_{n+1} = x_n + a_n(b - y_n)$ converges to x^*, where x^* is a solution of $\mathbf{M}(\mathbf{x}^*) = \boldsymbol{\theta}$, with a_n being an arbitrary sequence of real numbers. Kitsos in [22] applied the method for non-linear models. But why to adopt a Newton–Raphson framework in a statistical point estimation problem, under certain restrictions, and not the bisection method. The answer is that: The bisection approach leads to a (minimax) Decision Theory reasoning, and not to the classical statistical way of thinking; see Theorem 4 in Appendix 2. Stochastic Approximation is a particular method concerning statistical point estimation. Other methods were also developed; see, for example, [32, 55] for methods related to epidemiological problems.

3 Main Inequalities in Statistics

As far as the Statistics is concerned, the inequalities are strongly related to the development of the field. In the following, we present and discuss some widely used inequalities.

- *The Markov inequality*. Let X be a non-negative random variable (r.v.) with finite mean μ. Then, for every non-negative c, it holds

$$\Pr(X \geq c) \leq \frac{\mu}{c}.$$

The extra knowledge of variance results the following:

- *The Chebyshev's inequality*. Let X be an r.v. with given both finite mean μ and finite variance σ^2. Then, for every non-negative c, it holds that

$$\Pr(|X - \mu| \geq c) \leq \left(\frac{\sigma}{c}\right)^2.$$

The well-known Jensen's inequality relates the influence of a convex function when acting on the expected value operator. In particular:

- *The Jensen's inequality*. Let g be a convex function on a convex subset $\Omega \subseteq \mathbb{R}^k$, and suppose that $\Pr(X \in \Omega) = 1$. If the expected value $\mathrm{E}(X)$ of an r.v. X is finite, then $g\big(\mathrm{E}(X)\big) \leq \mathrm{E}\big(g(X)\big)$.

  The Cauchy–Schwartz inequality, mentioned in Sect. 2, is transferred in Statistics as:
- *The Statistical form of the Cauchy–Schwartz inequality*. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two random vectors of the same dimension such that $\mathrm{E}\big(\|\mathbf{X}_i\|^2\big)$, $i = 1, 2$, are finite. Then,

$$\mathrm{E}\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\right) \leq \sqrt{\mathrm{E}\left(\|\mathbf{X}_1\|^2\right)\,\mathrm{E}\left(\|\mathbf{X}_2\|^2\right)}.$$

  The Cauchy–Schwarz inequality provides food for thought on how Mathematics and Statistics communicate. In the following paragraph we discuss the sense of distance from a probabilistic point of view.
- *Distance in Probability Theory*. Let $(\Omega, \mathscr{A}, P)$ be a probability space consisting of the sample space Ω, the σ-algebra of "events" of Ω, and the probability measure P that maps each event to the real interval $[0, 1]$, i.e. $\mathscr{A} \ni A \overset{P}{\mapsto} P(A) \in [0, 1]$. Recall that $\mathscr{A} = \bigcup_{i\in\mathbb{N}} A_i$ with $A_i \bigcap A_j = \varnothing$, $i \neq j$, and $\sum_{i\in\mathbb{N}} P(A_i) = 1$. The *(probability) distance* D between two probability measures P and Q (of the same probability space) is denoted with $D(P, Q)$ and is defined as $D(P, Q) := \sup\big\{\big|P(A) - Q(A)\big|\big\}_{A\in\mathscr{A}}$. Note that the mapping D that assigns a real non-negative number to every pair of probability measures of Ω is—indeed—a distance metric. Furthermore, it is easy to see that $D(P, Q) \in [0, 1]$ for every P and Q, and the following holds.

Proposition 1 *The "exponentiated" distance D^* of a given bounded distance $0 \leq D \leq 1$, i.e. $D^*(P, Q) := \mathrm{e}^{D(P,Q)} - 1$, is also a distance metric.*

See Appendix 1 for the proof, where the exponential inequality $\mathrm{e}^x \geq (1 + x/n)^n$, $x \in \mathbb{R}$, $n \in \mathbb{N}$, was applied. We assume now that for every probability measure P of Ω, i.e. $P \in \mathscr{P}(\Omega)$, there is a σ-finite measure μ such that $P < \mu$ with $P \ll \mu$, i.e. P is absolutely continuous with respect to μ (assuming that $\mathscr{P}$ is countable, μ always exists since μ can be considered as $\mu := \sum_i 2^{-i} P_i$). Then, from the Radon–Nikodym theorem, there exists an integrable function $f : \mathscr{A} \to \mathbb{R}$ such that $P(A) = \int_A f \,\mathrm{d}\mu$, and thus $f := \mathrm{d}P/\mathrm{d}\mu$. Therefore, $D(P, Q) = \int_A |f - g| \,\mathrm{d}\mu$ with $f := \mathrm{d}P/\mathrm{d}\mu$ and $g := \mathrm{d}Q/\mathrm{d}\mu$. It holds, also, that $H^2(P, Q) < D(P, Q)$, $P, Q \in \mathscr{P}(\Omega)$, where H denotes the *Hellinger distance* defined by $H(P, Q)^2 := \int \big(\sqrt{f} - \sqrt{g}\big)^2 \,\mathrm{d}\mu = 2\big[1 - A(P, Q)\big]$, with $A(P, Q) := \int \sqrt{f\,g} \,\mathrm{d}\mu$ being the *affinity* between probability measures P and Q. This is true, since $H(P, Q)^2 < \int \big(\sqrt{f} - \sqrt{g}\big)\big(\sqrt{f} + \sqrt{g}\big) \,\mathrm{d}\mu \leq \int |f - g| \,\mathrm{d}\mu \leq D(P, Q)$ for every $P, Q \in \mathscr{P}(\Omega)$. For a study of the Hellinger distance between two generalized normal distributions, see [25].

- *Hypothesis testing for a mean*. In principle, if $\bar{\mathbf{x}} \in \mathbb{R}^{n\times 1}$ is the mean sample of n observations from the multivariate Normal distribution with mean vector

$\boldsymbol{\mu} \in \mathbb{R}^{n\times 1}$ and variance–covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$, the known region

$$n\,(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,(\overline{\mathbf{x}} - \boldsymbol{\mu}_0) \geq \chi^2_{p,\alpha},$$

is a critical region at the confidence level α for testing the hypothesis $\mathrm{H}: \boldsymbol{\mu} = \boldsymbol{\mu}_0$. As far as the confidence intervals are concerned in a Biostatistics level, there are different approaches for the Odds Ratio; see [32, 55]. As we have already mentioned, Statistical Inference is based on point estimation (see [50] for example) as well as on interval estimation. Note that the interval estimation by itself introduces the use of inequalities. The Likelihood method is still valid when the Maximum Likelihood Estimation (MLE), say $\hat{\boldsymbol{\theta}}$, of the unknown parameter vector $\boldsymbol{\theta} = (\theta_i) \in \Theta \subseteq \mathbb{R}^p$, with Θ being the parameter space, is subject to certain restrictions, say $h(\hat{\boldsymbol{\theta}}) = 0$. The well-known Lagrangian method is then applied, i.e.

$$\frac{\partial}{\partial \theta_i}\big[\ell(\boldsymbol{\theta}) - \lambda\, h(\boldsymbol{\theta})\big] = 0,$$

with $\ell(\boldsymbol{\theta})$ being the log-Likelihood function with regard to $\boldsymbol{\theta}$, and $\lambda \in \mathbb{R}$ the Lagrange multiplier. In such a case, still the estimate $\hat{\boldsymbol{\theta}}$ follows the (multivariate) Normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{\theta}$ and the asymptotic variance–covariance matrix $\boldsymbol{\Sigma} = n\,\mathbf{I}^{-1}(\boldsymbol{\theta})$, i.e. $\hat{\boldsymbol{\theta}} \sim \mathcal{N}\big(\boldsymbol{\theta}, n\,\mathbf{I}^{-1}(\boldsymbol{\theta})\big)$, where $\mathbf{I} \in \mathbb{R}^{p\times p}$ denotes the Fisher's information matrix; see the early work of Silvey in [44] among others. Moreover, Anderson in [1] discussed a number of confidence intervals concerning Multivariate Statistics, Ferguson in [9] considered a Decision Theory point of view, while Fortuin et al. in [11] focused on a particular inequality problem.

Example 1 Let us consider the vector of n observations $\mathbf{X} = (x_1, x_2, \ldots, x_k)$ which follows the k-th degree multinomial distribution, i.e.

$$p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\,x_2!\,\ldots\,x_k!}\theta_1^{x_1}\,\theta_2^{x_2}\,\ldots\,\theta_k^{x_k}, \quad \text{with} \quad \sum_{i=1}^{k} x_i = n \quad \text{and} \quad \sum_{i=1}^{k}\theta_i = 1,$$

while θ_i, $i = 1, 2, \ldots, k$, denote the involved parameters. Following, therefore, the typical procedure for the evaluation of the log-Likelihood under the restriction $h(\boldsymbol{\theta}) := \big(\sum \theta_i\big) - 1 = 0$, we can evaluate the expected value, variance, and covariance as

$$\mathrm{E}(x_i) = n\,\theta_i, \quad \mathrm{Var}(x_i) = n\,\theta_1\,(1-\theta_i), \quad \text{and} \quad \mathrm{Cov}(x_i, x_j) = -n\,\theta_i\,\theta_j, \quad i \neq j,$$

and hence, the inverse of the Fisher's information matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})$ is the variance–covariance matrix with elements $\big(I^{-1}\big)_{ii}(\boldsymbol{\theta}) = n^{-1}\,\theta_i(1-\theta_i)$, $i = 1, 2, \ldots, k$, and $\big(I^{-1}\big)_{ij}(\boldsymbol{\theta}) = n^{-1}\,\theta_i\,\theta_j$, $i \neq j = 1, 2, \ldots, k$.

Let $\mathbf{c}$ be now an appropriate constant vector for an approximate $(1-\alpha)\cdot 100\%$. The confidence interval for $\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}}$ is defined to be the real interval $\mathrm{CI}\left(\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}}\right) := \left(\mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}} - K_{\alpha/2}\left[\mathbf{c}^{\mathrm{T}}\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{c}\right],\ \mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\theta}} + K_{\alpha/2}\left[\mathbf{c}^{\mathrm{T}}\mathbf{I}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{c}\right]\right)$ with $\hat{\boldsymbol{\theta}}$ being an estimate of $\boldsymbol{\theta}$, and $K_{\alpha/2}$ the appropriate value for either standard Normal or t-distribution. Recalling the previous Example 1, notice that, although we assumed a multinomial distribution, the common marginal distribution of two components, say x_p and x_q, is a trinomial one with $x_p + x_q \leq n$, $1 \leq p, q \leq k$, $p \neq q$, while the probability distribution of $x_p + x_q = \xi$, $\xi = 0, 1, \ldots, n$, is binomial, since it is the probability distribution of x_i, $i = 1, 2, \ldots, k$, with different parameters; see also [32] for a special case in epidemiology. Notice also that the components of the corresponding Fisher's information matrix, as in Example 1, are non-linear functions of the unknown parameter vector $\boldsymbol{\theta}$. This creates a real problem regarding the calculations.

- *Sequential Probability Ratio Test (SPRT)*. The pioneering work of Wald in [52] was based on changing the probability ratio test; see also [53]. The fundamental difference is that now there are three regions testing two simple hypothesis $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$, there is a "continuation region" and the sample size is not fixed anymore but a random variable, say n, such that $\Pr(n < \infty \mid \boldsymbol{\theta}) = 1$. Moreover, the expected value $\mathrm{E}(n;\ \boldsymbol{\theta})$ exists and certain bounds for this were derived; see [14] for details, while when the average sample size is less than the appropriate sample size in a random sample see [54]. Usually, we denote the Operating Character (OC) function as $Q(\boldsymbol{\theta})$ and the power function as $R(\boldsymbol{\theta})$ $\left(:= 1 - Q(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\right)$. For given confidence levels, say α and β, for the above defined test, it is required that $Q(\boldsymbol{\theta}_0) \geq 1-\alpha$ and $Q(\boldsymbol{\theta}_1) \leq \beta$. Then, the logarithm of the probability ratio test at stage n is defined as

$$Z_n := \ln \frac{f_n(\mathbf{x}_n;\ \boldsymbol{\theta}_1)}{f_n(\mathbf{x}_n;\ \boldsymbol{\theta}_0)}, \quad n \geq 1, \quad \mathbf{x}_n = (x_1, x_2, \ldots, x_n).$$

Based on the SPRT, when two given numbers act as stopping bounds (B, A) with $-\infty < B < A < +\infty$, these numbers are defined through the decision rule:

1. Accept H_0 if $Z_n \leq B$,
2. Reject H_0 if $Z_n \geq A$, and
3. Continue by examining $\mathbf{x}_{n+1}$, i.e. $B < Z_{n+1} < A$.

The inequality $B < Z_{n+1} < A$ is known as the *critical inequality* and the test is denoted by $S(B, A)$. Following Ghosh in [14, Th. 3.2], the following is true.

Theorem 1 *The risk errors $\alpha(\boldsymbol{\theta}_0)$ and $\beta(\boldsymbol{\theta}_1)$ associated with the SPRT $S(B, A)$ for $\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, with $B < A$ being any choice of stopping bounds, then the following inequalities hold:*

$$\ln \frac{\beta(\boldsymbol{\theta}_1)}{1-\alpha(\boldsymbol{\theta}_0)} \leq \min\{0, B\}, \quad \ln \frac{1-\beta(\boldsymbol{\theta}_1)}{\alpha(\boldsymbol{\theta}_0)} \geq \max\{0, A\}.$$

However, the optimum bounds, say (B^*, A^*) have not evaluated and, therefore, the pair $(\alpha, \beta), \alpha+\beta < 1$, the optimum bounds can be approximated by $B^* \approx \ln \beta/(1-\alpha)$ and $A^* \approx \ln(1-\beta)/\alpha$.

Example 2 Let $x_1, x_2, \ldots$ be some Bernoulli variables regarding the SPRT with p being the proportion of successes, i.e. $\mathrm{H}_0 : \ p \le p_0$ vs. $\mathrm{H}_1 : \ p \ge p_1$, with $0 \le p_0 < p_1 \le 1$. For each observation x_i it is $Z_i = \ln\left\{(1-p_1)/(1-p_0)\right\} + x_i \ln\left\{p_1(1-p_0)\big/[p_0(1-p_1)]\right\}$. As $Z_n = \prod_{i=1}^n z_i$ and $X_n = \sum_{i=1}^n x_i$ then the critical inequality for $S(B^*, a^*)$ is reduced to $\mathrm{K} + \Lambda\, n < X_n < \mathrm{M} + \Lambda\, n$ where

$$\mathrm{K} := \frac{B^*}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}, \quad \Lambda := \frac{\ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}}, \quad \text{and} \quad \mathrm{M} := \frac{A^*}{\ln \frac{p_1(1-p_0)}{p_0(1-p_1)}},$$

see [14]. Moreover, the value $\mathrm{E}(n;\ p)$ is also bounded. In particular, $S \le \mathrm{E}(n;\ p) \le T$, where

$$S := \frac{Q(p)\left(\ln \frac{1-p_1}{1-p_0} + B - A\right) + A}{p \ln \frac{p_1(1-p_0)}{p_0(1-p_1)} + \ln \frac{1-p_1}{1-p_0}} \quad \text{and} \quad T := \frac{Q(p)\left(\ln \frac{p_0}{p_1} + B - A\right) + \ln \frac{p_1}{p_0} + A}{p \ln \frac{p_1(1-p_0)}{p_0(1-p_1)} + \ln \frac{1-p_1}{1-p_0}}.$$

- *Sequential design methods*. The sequential methods are the key for testing more than two hypotheses. Moreover, they are related to decision problems; see [41]. The inequalities involved to the Decision Theory, their links to the Bayesian Decision Theory and the evaluated risks are presented in a compact form by [41, Ch. 3]. The sequential way of thinking has been adopted by Kitsos in [22, 23] as well by Ford et al. in [10] with regard to optimal non-linear Design Theory. Moreover, Kitsos proved in [23] that when the initial design is D-optimal, [43], and a stochastic approximation scheme is used, then the limiting design is also D-optimal (and hence G-optimal due to the Kiefer's Equivalence Theorem). The main results of Wynn in [57, 58] rule the sequential design approach. The link between the optimal Design Theory and the moment inequalities was investigated by Torsney in [48], where Hölder's and Minkowski's inequalities were also discussed. If ξ denotes a design measure, [43], and $\mathbf{M}$ is the average-per-observation information matrix $\mathbf{M} = n^{-1}\,\mathbf{I}$, then it can be written as $\mathbf{M}(\xi) = n^{-1}\,\mathbf{I}(\xi)$ for the linear case, and $\mathbf{M}(\boldsymbol{\theta}, \xi) = n^{-1}\,\mathbf{I}(\boldsymbol{\theta}, \xi)$ for the non-linear case, where matrix $\mathbf{I}$ is the Fisher's information matrix; see [10]. In linear theory, it has been proved in [56] that $\mathbf{M}(\xi_n) \to \mathbf{M}(\xi^*)$ when $\xi_n \to \xi^*$, i.e. when a sequence of design measures converges to the optimum design, then the corresponding measures of information "follow" the scheme. That is, when we are not at the limit, inequalities are hold. This result is similar to the Dominated converge principle for a sequence of integrable functions, say u_n converging to u, provided that an integrable function w such that $|u_n| \le w$ exist, then u is also integrable and $\mathrm{E}(u_n) \to \mathrm{E}(u)$. However, this is not true for the non-linear case: there is no limiting result for $\mathbf{M}(\xi_n, \boldsymbol{\theta})$ or $\mathbf{M}(\xi_n, \boldsymbol{\theta}_n)$. Moreover, in Design Theory there is not

a similar to the Fatou's Lemma that $\mathrm{E}(\lim_{n\to\infty} u_n) \leq \lim_{n\to\infty} \mathrm{E}(u_n)$, $u_n \geq 0$. In particular, $\mathrm{E}(u) \leq \lim_{n\to\infty} \mathrm{E}(u_n)$ when $u_n \to u$.

- *Linear programming*. As far as the linear programming is concerned, the Simplex method solves linear inequalities problems, such as: evaluate $\max\{y = f(\mathbf{x})\}$, $\mathbf{x} \in \mathbb{R}^p$, under $\mathbf{A}\,\mathbf{x}^{\mathrm{T}} \leq \mathbf{b}^{\mathrm{T}}$, where $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{b} \in \mathbb{R}^p$ are known. Adding the so-called slack variables the inequalities are eventually transformed into equalities.

4 Inequalities in Probability Theory

In this section we present some essential inequalities used in Probability Theory, in order to clarify the importance of these inequalities to all the fields of Statistics.

- *Renewal Theory*. From the Renewal Theory [20], consider the elapsed number of generation, say $T(0)$, known also as a generation of equal components. Then for a finite population of constant size N, it can be proved that $\mathrm{E}\big(T(0)\big) \leq N^N$.
- *Doob's martingale*. Recall that a stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is called a *martingale* with respect to $\{Y_n\}_{n\in\mathbb{N}}$ if $\mathrm{E}\{|Y_n|\} < \infty$ and $\mathrm{E}\big(X_{n+1}(Y_0, Y_1, \ldots, Y_n)\big) = X_n$, $n \in \mathbb{N}$. In such a case, the "existing history" determines x_n in terms that, eventually, $\mathrm{E}(X_n) = \mathrm{E}(X_{n+1}|\ \ Y_0, Y_1, \ldots, Y_n) = \mathrm{E}(X_0)$ for every $n \in \mathbb{N}$. As far as the Doob's Martingale Process is concerned, the inequality is requested in its definition, as well as for the Radon–Nikodym derivatives; see [8]. Indeed, for a given r.v. X with $\mathrm{E}(|X|) < \infty$, and for an ordinary sequence of r.v.-s, say $Y_0, Y_1, \ldots, Y_n$, then from $X_n := \mathrm{E}(X \mid Y_0, Y_1, \ldots, Y_n)$, $n \in \mathbb{N}$, a martingale structure $\{X_n\}$ with respect to $\{Y_n\}$, is obtained when $\mathrm{E}(|X_n|) \leq \mathrm{E}(|X|) < \infty$ and $\mathrm{E}(X_{n+1} \mid Y_0, Y_1, \ldots, Y_n) = X_n$, known as Doob's process. Suppose, now, that U is a uniformly distributed r.v. on $[0,\ 1]$. We define $Y_n = k/2^n$, $k = k(n, U)$, unique such that $k/2^n \leq U \leq (k+1)/2^n$. Then, process $\{X_n\}$ defined as $X_n := 2^n\big[g(Y_n + 2^{-n}) - g(Y_n)\big]$ for $g|_{[0,\,1]}$ bounded forms a martingale; see [20]. Moreover, the sequence X_n is known as the Radon–Nikodym derivative of g evaluated at U.
- *Crossing inequality*. One of the well-known inequalities in Stochastic Process Theory, strongly related to Sequential Analysis, is the so-called Crossing Inequality. It counts the number of times a sub-martingale $\{X_n\}$, with respect to a sequence $\{Y_n\}$, crosses a given interval $(a,\ b) \subseteq \mathbb{R}$. That is, the number of crosses, say $N_{a,b}$, from the level below a to a level above b. In fact, $N_{a,b}$ is the number of pairs $(i,\ j)$ such that $X_i \leq a$ and $X_j \geq b$ with $a < X_k < b$, $0 \leq i < j \leq N_j$, $i < k < j$. For sub-martingales $\{X_n\}$ with given T and T' Markov times and $q \in Z$ with $0 \leq T \leq T' \leq q$, then $\mathrm{E}(X_T) \leq \mathrm{E}(X_{T'})$. The Crossing Inequality is then formulated by $\mathrm{E}(N_{a,b}) \leq \big(\mathrm{E}\big[(X_N - a)^+\big] - \mathrm{E}\big[(X_0 - a)^+\big]\big)/(b-a)$. For the *backward* martingale $\{X_n\}_{n=0,-1,-2,\ldots}$ with respect to a σ-field $\mathscr{F}_n$, $n = 0, -1, -2, \ldots$ (generated by some jointly distributed r.v.-s), the Crossing Inequality is reduced to $\mathrm{E}(N_{a,b}) \leq \mathrm{E}\big[(X_0 - a)^+\big]/(b-a)$ with the

only new restriction $N \leq i < j \leq 0$. For a given martingale $\{X_n\}$ satisfying $E\left(|X_n|^k\right) < \infty$ for every $k > 1$ and $n \in \mathbb{N}$, it can be proved that

$$\mathrm{E}\left(\max_{0\leq r\leq n}\left\{|X_r|\right\}\right) \leq \tfrac{k}{k-1}\ \mathrm{E}\left(|X_k|^k\right)^{1/k} \text{ and } \mathrm{E}\left(\max_{0\leq r\leq n}\left\{|X_r|^k\right\}\right) \leq \left(\tfrac{k}{k-1}\right)^k\ \mathrm{E}\left(|X_n|^k\right),$$

see [20] for details. When restrictions are imposed to expected value and variance, i.e. $\mathrm{E}(X_n) = 0$ and hence $\sigma^2 = \mathrm{E}(X_n^2) < \infty$ for every n, then

$$\Pr\left(\max_{0\leq r\leq n}\left\{|X_r|\right\} > k\right) \leq \tfrac{\sigma^2}{\sigma^2+k}, \quad k > 0.$$

- *Chebyshev's and Kolmogorov's inequalities*. Chebyshev's Inequality provides food for thought when an extension, known as the Kolmogorov's Inequality, is considered. For given two independent and identically distributed (i.i.e.) r.v-s $X_1, X_2, \ldots$, with mean $\mu = \mathrm{E}(X_i) = 0$ and variance $\sigma^2 = \mathrm{E}(X_i^2) < \infty$, $i = 1, 2, \ldots$, we define $S_n = X_1 + X_2 + \cdots + X_n$, $n = 1, 2, \ldots$, and $S_0 = 0$. Then, Chebyshev's Inequality is formulated by

$$\epsilon^2\ \Pr\left(|S_n| > \epsilon\right) \leq n\,\sigma^2 = \mathrm{Var}(S_n),$$

while Kolmogorov's Inequality is written as

$$\epsilon^2\ \Pr\left(\max_{k\leq n}\left\{|S_k|\right\} > \epsilon\right) \leq n\,\sigma^2 = \mathrm{Var}(S_n).$$

- *Maximal inequalities*. A number of inequalities are based on Kolmogorov's Inequality for the (sub-)martingales, and are known as the Maximal Inequalities; see [8, 20].

  1. Let $\{X_n\}$ be a martingale and $k \geq 0$. Then, $k\ \Pr\left(\max_{0\leq r\leq n}\left\{|X_r|\right\} > k\right) \leq \mathrm{E}\left(\left\{|X_n|\right\}\right)$.
  2. Let $\{X_n\}$ be a sub-martingale with $X_n \geq 0$, $n \in \mathbb{N}$, and $k \geq 0$. Then, $k\ \Pr\left(\max_{0\leq r\leq n}\{X_r\} > k\right) \leq \mathrm{E}\left(\{X_n\}\right)$.
  3. When $\{X_n\}_{n=0,-1,-2,\ldots}$ is a backward martingale with respect to a σ-field, say $\mathscr{F}_n$, $n = 0, -1, -1, \ldots$, generated by some jointly distributed r.v.-s $\{Y_n, Y_{n-1}, \ldots\}$, then $k\ \Pr\left(\max_{0\leq r\leq n}\{X_r\} > k\right) \leq \mathrm{E}\left(\{X_0\}\right)$.
  4. Let $\{X_n\}$ be a martingale. Then, $k\ \Pr\left(\min_{0\leq r\leq n}\{X_r\} < -k\right) \leq \mathrm{E}\left(\{X_n^+\} - \mathrm{E}(X_0)\right)$.
  5. Let $\{X_n\}$ be a super-martingale with $X_n \geq 0$, $n \in \mathbb{N}$. In such a case, $k\ \Pr\left(\max_{0\leq r\leq n}\{X_r\} \geq k\right) \leq \mathrm{E}(X_0)\big)$.

  The above inequalities from the Probability Theory provide evidence of how really useful inequalities can be, in terms of offering bounds, for most of the involved "sequences," such as martingales. In the next section, we present the existence of certain bounds related to information measures.

- *Distance in navigation*. Franceschetti and Meester in [12], working in similar line of thought as in [42] and [35], consider the Euclidian distance between a source point and a target, in navigation in random networks, and presented a number of interesting inequalities for the ϵ-delivery time of a decentralized algorithm. This refers to the number of steps required for the message, originating at point s to reach an ϵ-neighborhood of point t. Moreover, working on network topology, they introduced a new distance measure, the chemical distance between two points x and y (and by considering the existence of a path connecting x with y), with a number of inequalities obtained through Probability Theory: for a random grid and given points x and y, probability assigned to be 1 if $|x-y|=1$, and $1-\exp\left(-\beta/|x-y|^a\right)$ if $|x-y|>1$, $a,b>0$. Their results are related to the percolation models; see [18, 31]. Although the evolution of ideas from Shannon's work in [42] to Navigation in Random Networks is important, it has attracted the interest of Engineers rather than Mathematicians, as the former pay more attention to the information flow in random networks; see [45]. We present here an important—in our opinion—inequality related to the Phase Transition: There is an interest to express positive correlations between increasing events, say A and B, so that $\Pr(A\cap B)\geq \Pr(A)\ \Pr(B)$; see [18, 31]. Then, for increasing events $A_1, A_2, \ldots, A_n$, all having the same probability, it holds that

$$1-\left[1-\Pr\left(\bigcup_{i=1}^{n}A_i\right)\right]^{1/n}\leq \Pr(A_1).$$

Indeed, due to $\Pr(A\cap B)\geq \Pr(A)\ \Pr(B)$ and some set-theoretic algebra,

$$1-\Pr\left(\bigcup_{i=1}^{n}A_i\right)=\Pr\left(\bigcap_{i=1}^{n}A_i\right)\geq\prod_{i=1}^{n}\Pr\left(A_i^c\right)=\left[\Pr\left(A_i^c\right)\right]^n=\left[1-\Pr(A_1)\right]^n,$$

since we assumed that $\Pr(A_i)=\Pr(A_j)$, $i\neq j=1,2,\ldots,n$.

5 Information Measures and Inequalities

In the following we shall try to investigate certain bounds concerning generalized entropy type information measures from the Information Theory.

New entropy type information measures were introduced in [24], generalizing the known Fisher's entropy type information measure; see also [5, 26–29, 49]. The introduced new entropy type measure of information $J_\alpha(X)$ is a function of the density f of the p-variate random variable r.v. X defined as, [24],

$$J_\alpha(X):=\mathrm{E}\left(\left\|\nabla\log f(X)\right\|^\alpha\right)=\int_{\mathbb{R}^p}f(x)\left\|\nabla\log f(x)\right\|^\alpha\,\mathrm{d}x,\quad \alpha>1, \tag{4}$$

where $\|\cdot\|$ is the usual two-norm of $\mathscr{L}^2(\mathbb{R}^p)$. Notice that $J_2 = J$, with J being the known Fisher's entropy type information measure.

In his pioneering work [42], Shannon introduced the notion of Entropy in an Information Theory context giving a new perspective to the study of Information Systems, Signal Processing and Cryptography among other fields of application. *Shannon entropy*, or *differential entropy*, denoted by $H(X)$, measures the average uncertainty of an r.v. X and is given by

$$H(X) := -\,\mathrm{E}\left(\log f(X)\right) = -\int_{\mathbb{R}^p} f(x)\log f(x)\,\mathrm{d}x, \tag{5}$$

with f being the probability density function (p.d.f.) of r.v. X; see [6, 42]. In Information Theory, it is the minimum number of bits required, on the average, to describe the value x of the r.v. X. In Cryptography, entropy gives the ultimately achievable error-free compression in terms of the average codeword length symbol per source; see [21] among others.

For the Shannon entropy $H(X)$ of any multivariate r.v. X with zero mean vector and covariance matrix $\mathbf{\Sigma}$, an upper bound exists,

$$H(X) \le \tfrac{1}{2}\log\left\{(2\pi\,\mathrm{e})^p|\det\mathbf{\Sigma}|\right\}, \tag{6}$$

where the equality holds if and only if X is a normally distributed r.v., i.e. $X \sim \mathscr{N}(\mathbf{0}, \mathbf{\Sigma})$; see [6]. Note that the Normal distribution is usually adopted as the description variable for noise, and acts additively to the input variable when an input–output discrete time channel is formed. The known entropy power, dented by $N(X)$, and defined through the Shannon entropy $H(X)$, has been extended to

$$N_\alpha(X) := \nu_\alpha \exp\left\{\tfrac{\alpha}{p}\mathrm{H}(X)\right\}, \tag{7}$$

where

$$\nu_\alpha := \left(\tfrac{\alpha-1}{\mathrm{e}}\right)\pi^{-\alpha/2}\left[\frac{\Gamma\left(\frac{p}{2}+1\right)}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\frac{\alpha}{p}}, \quad \alpha > 1, \tag{8}$$

see [24] for details. Notice that $\nu_2 = (2\pi e)^{-1}$ and hence $N_2 = N$. It can be proved that [24],

$$J_\alpha(X)\,N_\alpha(X) \ge p, \quad \alpha > 1, \tag{9}$$

which extends the well-known Information Inequality, i.e. $J(X)\,N(X) \ge p$, obtained from (9) by setting $\alpha := 2$.

The so-called Cramér–Rao Inequality, [6, Th. 11.10.1], is generalized due to the introduced information measures, [24], and is given by

$$\sqrt{\tfrac{2\pi\,\mathrm{e}}{p}\,\mathrm{Var}(X)}\left[\tfrac{\nu_\alpha}{p}\,J_\alpha(X)\right]^{1/\alpha} \geq 1, \quad \alpha > 1. \tag{10}$$

When $\alpha := 2$ we have $\mathrm{Var}(X)\,J_2(X) \geq p^2$, which is the known Cramér–Rao inequality, $\mathrm{Var}(X)\,J(X) \geq 1$ for the univariate case. The lower boundary B_α for the introduced generalized information $J_\alpha(X)$ is then

$$\tfrac{p}{\nu_\alpha}\left[\tfrac{2\pi\,\mathrm{e}}{p}\,\mathrm{Var}(X)\right]^{-\alpha/2} =: B_\alpha \leq J_\alpha(X), \quad \alpha > 1. \tag{11}$$

Finally, the classical Entropy Inequality,

$$\mathrm{Var}(X) \geq p\,N(X) = \tfrac{p}{2\pi\,\mathrm{e}}\exp\left\{\tfrac{2}{p}\,H(X)\right\}, \tag{12}$$

can be extended, adopting the extended entropy power as in (7), to the general form

$$\mathrm{Var}(X) \geq \tfrac{p}{2\pi\,\mathrm{e}}\nu_\alpha^{-2/\alpha}\,N_\alpha^{2/\alpha}(X), \quad \alpha > 1. \tag{13}$$

Under the "normal" parameter value $\alpha := 2$, inequality (13) is reduced to (12).

The Blachman–Stam Inequality [2, 3, 47] is generalized through the generalized J_α measure. Indeed: For given two independent r.v.-s X and Y of the same dimension, it holds

$$J_\alpha\left(\lambda^{1/\alpha}X + (1-\lambda)^{1/\alpha}Y\right) \leq \lambda J_\alpha(X) + (1-\lambda)J_\alpha(Y), \quad \lambda \in (0,1),$$

where the equality holds for X and Y normally distributed r.v.-s with the same covariance matrix; see [26] for the proof. For parameter value $\alpha := 2$ we are reduced to the well-known Blachman–Stam Inequality, since $J_2 = J$.

Let now $X_1, X_2, \ldots, X_n$ be some n independent and identically distributed (i.i.d.) univariate random variables with mean 0 and variance σ^2, having density function $f(x)$ satisfying Poincaré conditions with finite restricted Poincaré constant c_P. If $\phi(x)$ denotes the corresponding probability density of $\mathcal{N}(0,\sigma^2)$, then the *Fisher's information distance* (or *standardized information*) of some univariate r.v. X (with mean 0 and variance σ^2) is defined to be

$$J_\phi(X) := \sigma^2\,\mathrm{E}\left[\tfrac{\mathrm{d}}{\mathrm{d}x}\log f(X) - \tfrac{\mathrm{d}}{\mathrm{d}x}\log\phi(X)\right]^2 = \sigma^2\,J(X) - 1,$$

with J being the known Fisher's (entropy type) information. Notice that $J_\phi(\lambda\,X) = J_\phi(X)$, so J_ϕ is scale invariant and, moreover, provides a measure of distance of "how far $f(x)$ is from normality," i.e. from $\phi(x)$. Then, for the sum $Y_n := \left(\sqrt{n}\sigma\right)^{-1}\sum_{i=1}^n X_i$, it can be proved that for every n

$$J(Y_n) = \frac{2c_P}{2c_P + (n-1)\sigma^2} J(X_1).$$

Moreover, if $\phi(x)$ represents the probability density of the standard Normal distribution, then it holds that

$$\sup_{x\in\mathbb{R}} \left\{ \left| f(x) - \phi(x) \right| \right\} \leq \left(1 + \sqrt{\sigma}/\pi\right) \sqrt{J(X)}, \quad \int_{\mathbb{R}} \left| f(x) - \phi(x) \right| \mathrm{d}x \leq 2H(f,\phi) \leq \sqrt{2J(X)},$$

with $H^2(f,\phi) := \int \left| \sqrt{f(x)} - \sqrt{\phi(x)} \right|^2 \mathrm{d}x$ being the *Hellinger distance* between densities f and ϕ; see [19] for details.

6 The Generalized Normal (GN) Distribution

The Logarithmic Sobolev Inequalities (LSI) attempt to estimate the lower-order derivatives of a given function in terms of higher-order derivatives. The well-known LSI was introduced in 1938 and translated in English 1963 as appeared in [46]; see also [16, 26] for details. The introductory and well-known Sobolev Inequality (SI) is of the form

$$\left(\int_{\mathbb{R}^p} |f(x)|^{\frac{2p}{p-2}} \mathrm{d}x \right)^{\frac{p-2}{2p}} \leq c_{\mathrm{S}} \left(\int_{\mathbb{R}^p} |\nabla f(x)|^2 \mathrm{d}x \right)^{\frac{1}{2}}, \tag{14}$$

or, using the two-norm notation, $\|f\|_q \leq c_{\mathrm{S}} \|\nabla f\|_2$, with the constant $c_{\mathrm{S}} > 0$ is known as the *Sobolev constant*.

Kitsos and Tavoularis [24] introduced and studied an exponential-power generalized form of the multivariate Normal distribution, denoted as $\mathcal{N}_\gamma(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, called the γ-order Generalized Normal (γ-GN) distribution; see also [27, 28] for further reading. The derivation of this three-parameter extended Normal distribution came up an extremal of a generalized Euclidian LSI introduced by Del Pino et al. in [7], which can be written as

$$\int_{\mathbb{R}^p} |u|^\gamma \log |u| \, \mathrm{d}\mathbf{x} \leq \frac{p}{\gamma^2} \log \left\{ K_\gamma \int_{\mathbb{R}^p} |\nabla u|^\gamma \, \mathrm{d}\mathbf{x} \right\}, \tag{15}$$

where $u = u(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^p$, belongs to the Sobolev space $H^{1/2}(\mathbb{R}^p)$ with $\|u\|_\gamma = \int_{\mathbb{R}^p} |g(\mathbf{x})|^\gamma \, \mathrm{d}\mathbf{x} = 1$. The optimal constant K_γ is being equal to

$$K_\gamma := \frac{\gamma}{p} \left(\frac{\gamma - 1}{\mathrm{e}} \right)^{\gamma-1} \pi^{-\gamma/2} \left[\frac{\Gamma(\frac{p}{2} + 1)}{\Gamma\left(p\frac{\gamma-1}{\gamma} + 1 \right)} \right]^{\gamma/p}. \tag{16}$$

The equality in (15) holds, [24], when u is considered to be the p.d.f. of an r.v. X following γ-GN distribution as defined below.

Definition 1 The p-variate random variable X follows the *γ-order generalized Normal* (γ-GN) distribution, i.e. $X \sim \mathcal{N}_\gamma(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with location parameter vector $\boldsymbol{\mu} \in \mathbb{R}^p$, shape parameter $\gamma \in \mathbb{R} \setminus [0, 1]$, and positive definite scale parameter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, when the density function f_X of X is of the form

$$f_X(\mathbf{x}) = f_X(\mathbf{x};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}, \gamma, p) := C(\boldsymbol{\Sigma}) \exp\left\{-\tfrac{\gamma-1}{\gamma} Q(\mathbf{x})^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad \mathbf{x} \in \mathbb{R}^p, \tag{17}$$

where Q is the p-quadratic form $Q(\mathbf{x}) = Q(\mathbf{x};\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) := (\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}$, $\mathbf{x} \in \mathbb{R}^p$, while the normalizing factor C is defined as

$$C(\boldsymbol{\Sigma}) = C(\boldsymbol{\Sigma};\ \gamma, p) := \frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}\,\Gamma\left(p\frac{\gamma-1}{\gamma}+1\right)\sqrt{|\boldsymbol{\Sigma}|}} \left(\tfrac{\gamma-1}{\gamma}\right)^{p\frac{\gamma-1}{\gamma}}, \tag{18}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant $\det \boldsymbol{\Sigma}$ of the scale matric $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$.

From the p.d.f. f_X as above, notice that the location vector of X is essentially the mean vector of X, i.e. $\boldsymbol{\mu} = \boldsymbol{\mu}_X := \mathrm{E}(X)$. Moreover, for the shape parameter value $\gamma = 2$, $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is reduced to the well-known multivariate normal distribution, where $\boldsymbol{\Sigma}$ is now the covariance of X, i.e. $\mathrm{Cov}\, X = \boldsymbol{\Sigma}$. Recall that

$$\mathrm{Cov}(X) = \frac{\Gamma\left((p+2)\frac{\gamma-1}{\gamma}\right)}{p\,\Gamma^3\left(p\frac{\gamma-1}{\gamma}\right)} \left(\tfrac{\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}} \boldsymbol{\Sigma}, \tag{19}$$

for the positive definite scale matrix $\boldsymbol{\Sigma}$; see [28].

Note that there are several other exponential-power generalizations of the usual Normal distribution, see [4, 15, 34], and [59] among others. Those generalizations are technically obtained and, thus, they have no specific physical interpretation. On the contrary, the γ-GN distribution has a strong information-theoretic background. Indeed, the most significant fact about the γ-GN family is that—at least for the spherically contoured case—acts to the generalized Information Inequality, the same way as the usual Normal distribution acts (i.e. providing equality) to the usual Information Inequality. In fact, the generalized form of the Information Inequality in (9) is reduced to equality for every spherically contoured γ-order normally distributed r.v., as it holds that $J_\alpha(X)\, N_\alpha(X) = p$ for $X \sim \mathcal{N}_\alpha\left(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_p\right)$; see [24, Cor. 3.2] for details. Moreover, the equality in the generalized Cramér–Rao Inequality as in (10) is achieved for r.v. X following the γ-GN distribution as above, i.e. it behaves the same way the usual Normal distribution does on the usual Cramér–Rao inequality. Indeed, using the fact that $J_\alpha(X)\, N_\alpha(X) = p$ holds for $X \sim \mathcal{N}_\alpha\left(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_p\right)$, as well as the extended Entropy Inequality as in (13), the equality of (10) can then be deduced, for the spherically contoured case; see also [26].

The family of multivariate γ-GN distributions, i.e. the family of the elliptically contoured γ-order generalized Normals, provides a smooth bridging between some important multivariate (and elliptically countered) distributions. Indeed:

1. *Case* $\gamma := 0$. For the limiting case when the shape parameter $\gamma \to 0^-$, the degenerate Dirac distribution $\mathscr{D}(\boldsymbol{\mu})$ with pole at point $\boldsymbol{\mu} \in \mathbb{R}^p$ is derived for dimensions $p := 1, 2$, while for $p \geq 3$ the corresponding p.d.f. "vanishes," i.e. $f_X \equiv 0$ for $X \sim \mathscr{N}_0(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. *Case* $\gamma := 1$. For the limiting case when $\gamma \to 1^+$, the elliptically contoured Uniform distribution $\mathscr{U}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained, which is defined over the p-ellipsoid $Q(\mathbf{x}) \leq 1, \mathbf{x} \in \mathbb{R}^p$.
3. *Case* $\gamma := 2$. For the "normality" case of $\gamma := 2$ the usual p-variate Normal distribution $\mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is obtained.
4. *Case* $\gamma := \pm\infty$. For the limiting case when $\gamma \to \pm\infty$ the elliptically contoured Laplace distribution $\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is derived.

See [28] for details. Therefore, one of the merits of the γ-GN family is that it can provide "heavy-" or "light-tailed" distributions as the change of shape parameter γ influences the "amount" of probability at the tails.

7 Information Divergencies

The informational divergence between two r.v.-s is usually calculated through the Kullback–Leibler (KL) divergence, which is acting as an "discrimination" measure of information. Recall that the *KL divergence* (also known as *relative entropy*), usually denoted by $D_{\mathrm{KL}}(X\|Y)$, of an r.v. X over an r.v. Y (of the same dimension), measures the amount of information "gained" when r.v. Y is replaced by X (say in an I/O system), and is defined by, [6],

$$D_{\mathrm{KL}}(X\|Y) := \int f_X \log \frac{f_X}{f_Y}, \tag{20}$$

where f_X and f_Y denote the corresponding density functions of r.v.-s X and Y.

In this section, we shall investigate the KL divergence measure of the multivariate γ-order normally distributed $X \sim \mathscr{N}_\gamma(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ over the multivariate t_ν-distributed $Y \sim t_\nu(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$; see [51] for the univariate case. Recall the p.d.f. f_Y of the multivariate (and scaled) t_ν-distributed r.v. Y with $\nu \geq 1$ degrees of freedom, mean vector $\boldsymbol{\mu}_2 \in \mathbb{R}^p$, and scale matrix $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{p\times p}$, which is given by

$$f_Y(\mathbf{y}) = f_Y(\mathbf{y};\ \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \nu) := C_2 \left[1 + \frac{1}{\nu} Q_2(\mathbf{y})\right]^{-\frac{\nu+p}{2}}, \quad \mathbf{y} \in \mathbb{R}^p, \tag{21}$$

with normalizing factor

$$C_2 = C_2(\mathbf{\Sigma}_2;\ \nu,\, p) := \frac{(\pi\,\nu)^{-p/2}\ \Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{|\mathbf{\Sigma}_2|}}, \tag{22}$$

and p-quadratic form $Q_2(\mathbf{y}) := (\mathbf{y}-\boldsymbol{\mu}_2)\,\mathbf{\Sigma}_2^{-1}\,(\mathbf{y}-\boldsymbol{\mu}_2)^{\mathrm{T}}$, $\mathbf{y} \in \mathbb{R}^p$. Note that parameter ν can be also a positive real $\mathbb{R}^+ \ni \nu \geq 1$.

The following theorem provides an upper bound for the "gained" information when the t_ν-distribution is replaced by a γ-GN distribution. Note that we often rely on inequalities when it comes to the calculation of information divergencies (including KL) between certain r.v.-s, since the integrals involved cannot usually be solved in a closed form.

Theorem 2 *The KL divergence $D_{\mathrm{KL}} := D_{\mathrm{KL}}(X\|Y)$, of a multivariate spherically contoured γ-order normally distributed r.v. $X \sim \mathcal{N}_\gamma\left(\boldsymbol{\mu}, \sigma_1^2\mathbb{I}_p\right)$ over a t_ν-distributed r.v. $Y \sim t_\nu\left(\boldsymbol{\mu}, \sigma_2^2\mathbb{I}_p\right)$, of the same mean $\boldsymbol{\mu} \in \mathbb{R}^p$, has the following upper bound,*

$$D_{\mathrm{KL}} \leq \log K + p\left(\log\tfrac{\sigma_2}{\sigma_1} - \tfrac{\gamma-1}{\gamma}\right) + \tfrac{\nu+p}{2\nu}\left(\tfrac{\sigma_1}{\sigma_2}\right)^2\left(\tfrac{\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}}\frac{\Gamma\left((p+2)\,\frac{\gamma-1}{\gamma}\right)}{\Gamma\left(p\,\frac{\gamma-1}{\gamma}\right)}, \tag{23}$$

where

$$K = K(\gamma,\nu,p) := \frac{\nu^{p/2}\ \Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{\nu}{2}\right)}{2\,\Gamma\left(p\,\frac{\gamma-1}{\gamma}\right)\Gamma\left(\frac{\nu+p}{2}\right)}\left(\tfrac{\gamma-1}{\gamma}\right)^{p\frac{\gamma-1}{\gamma}-1}. \tag{24}$$

Proof From the definition of the KL divergence (20) and the probability densities f_X and f_Y, as in (17) and (21), with K, C_1, and C_2 are defined as in (24), (18), and (22), respectively, while $Q_i(\mathbf{x}) := (\mathbf{x}-\boldsymbol{\mu})\,\mathbf{\Sigma}_i^{-1}\,(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}$, $\mathbf{x} \in \mathbb{R}^p$, $i = 1, 2$, with $\mathbf{\Sigma}_1 := \sigma_1^2\,\mathbb{I}_p$, $\mathbf{\Sigma}_2 := \sigma_2^2\,\mathbb{I}_p$, it holds

$$D_{\mathrm{KL}} = C_1\left[\left(\log K + p\,\log\tfrac{\sigma_2}{\sigma_1}\right)I_1 - g\,I_2 + \tfrac{p+\nu}{2}\,I_3\right], \tag{25}$$

where

$$I_1 := \int_{\mathbb{R}^p}\exp\left\{-g\left\|\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma_1}\right\|^{1/g}\right\}\mathrm{d}\mathbf{x}$$

$$I_2 := \int_{\mathbb{R}^p}\exp\left\{-g\left\|\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma_1}\right\|^{1/g}\right\}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma_1}\right\|^{1/g}\mathrm{d}\mathbf{x},\quad\text{and}$$

$$I_3 := \int_{\mathbb{R}^p}\exp\left\{-g\left\|\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma_1}\right\|^{1/g}\right\}\log\left(1+\frac{1}{\nu}\left\|\frac{\mathbf{x}-\boldsymbol{\mu}}{\sigma_2}\right\|^2\right)\mathrm{d}\mathbf{x},$$

and $g = g(\gamma) := (\gamma - 1)/\gamma$. Applying the linear transformation $\mathbf{z} = \mathbf{z}(\mathbf{x}) := g^g\,(\mathbf{x} - \boldsymbol{\mu})/\sigma_1$, $\mathbf{x} \in \mathbb{R}^p$, with $\mathrm{d}\mathbf{x} = g^{-p\,g}\,\sigma_1^p\,\mathrm{d}\mathbf{z}$, the above three multiple integrals are then written as

$$I_1 \;=\; g^{-p\,g}\,\sigma_1^p \int_{\mathbb{R}^p} \mathrm{e}^{-\|\mathbf{z}\|^{1/g}}\,\mathrm{d}\mathbf{z}, \tag{26a}$$

$$I_2 \;=\; g^{-p\,g}\sigma_1^p \int_{\mathbb{R}^p} \|\mathbf{z}\|^{1/g}\,\mathrm{e}^{-\|\mathbf{z}\|^{1/g}}\,\mathrm{d}\mathbf{z}, \text{ and} \tag{26b}$$

$$I_3 \;=\; g^{-p\,g}\,\sigma_1^p \int_{\mathbb{R}^p} \mathrm{e}^{-\|\mathbf{z}\|^{1/g}}\,\log\left(1 + \tfrac{g^{-2g}}{\nu}\left(\tfrac{\sigma_1}{\sigma_2}\right)^2\|\mathbf{z}\|^2\right)\mathrm{d}\mathbf{z}. \tag{26c}$$

Applying then the known integrals

$$\int_{\mathbb{R}^p} \mathrm{e}^{-\|\mathbf{z}\|^\beta}\,\mathrm{d}\mathbf{z} = \frac{2\pi^{p/2}\,\Gamma\left(\frac{p}{\beta}\right)}{\beta\,\Gamma\left(\frac{p}{2}\right)} \;\text{ and }\; \int_{\mathbb{R}^p} \|\mathbf{z}\|^\beta\,\mathrm{e}^{-\|\mathbf{z}\|^\beta}\,\mathrm{d}\mathbf{z} = \tfrac{p}{\beta}\int_{\mathbb{R}^p} \mathrm{e}^{-\|\mathbf{z}\|^\beta}\,\mathrm{d}\mathbf{z}, \tag{27}$$

with $\beta \in \mathbb{R}^{*+} := \mathbb{R}^+ \setminus \{0\}$, integrals (26a) and (26b) are then calculated as

$$I_1 = g^{-p\,g}\,\sigma_1^p\,\frac{2\pi^{p/2}}{\Gamma(p/2)}\,g\,\Gamma(p\,g) \;\text{ and }\; I_2 = p\,g\,I_1, \tag{28}$$

, respectively. Thus, (25) is reduced to

$$D_{\mathrm{KL}} = C_1\,\left(\log K + p\,\log\frac{\sigma_2}{\sigma_1} - p\,g\right)\,I_1 + \tfrac{p+\nu}{2}\,C_1\,I_3.$$

Substituting I_1 from (28) and using C_1 from (18), and applying the Gamma function additive identity, the above is reduced to

$$D_{\mathrm{KL}} = \log K + p\,\left(\log\frac{\sigma_2}{\sigma_1} - g\right) + \frac{p+\nu}{4\left(\sqrt{\pi}\,\sigma_1\right)^p}\,\frac{\Gamma(p/2)}{\Gamma(p\,g)}\,g^{p\,g-1}\,I_3. \tag{29}$$

Notice that the function in the integral of (26c) is positive, and so, using the known logarithmic inequality $\log(x+1) \le x$, $x > -1$, relation (26c) implies

$$I_3 \le g^{-(p+2)\,g}\,\frac{\sigma_1^{p+2}}{\nu\,\sigma_2^2}\int_{\mathbb{R}^p} \|\mathbf{z}\|^2\,\mathrm{e}^{-\|\mathbf{z}\|^{1/g}}\,\mathrm{d}\mathbf{z}. \tag{30}$$

We calculate now the first and the third integral of the above inequality by switching to hyperspherical coordinates, while the second integral is calculated using the relation first of (27). Recall the known hyperspherical transformation

$H_p : S_p \to \mathbb{R}^p$, where $S_p := \mathbb{R}^+ \times [0, \pi)^{p-2} \times [0, 2\pi)$, in which $S_p \ni (\rho, \varphi_1, \varphi_2, \ldots, \varphi_{p-1}) \overset{H_p}{\longmapsto} (z_1, z_2, \ldots, z_p) \in \mathbb{R}^p$, is given by

$$z_1 = \rho \cos\varphi_1, \tag{31a}$$

$$z_i = \rho \sin\varphi_1 \sin\varphi_2 \cdots \sin\varphi_{i-1} \cos\varphi_i, \quad i = 2, 3, \ldots, p-1, \tag{31b}$$

$$z_p = \rho \sin\varphi_1 \sin\varphi_2 \cdots \sin\varphi_{p-2} \sin\varphi_{p-1}, \tag{31c}$$

where $\rho \in \mathbb{R}^+$, $\varphi_1, \varphi_2, \ldots, \varphi_{p-2} \in [0, \pi)$, and $\varphi_{p-1} \in [0, 2\pi)$. It holds that $\|\mathbf{z}\|^2 = z_1^2 + z_2^2 + \cdots + z_p^2 = \rho^2$, $\mathbf{z} \in \mathbb{R}^p$, while the volume element $\mathrm{d}\mathbf{z} = \mathrm{d}z_1\, \mathrm{d}z_2 \cdots \mathrm{d}z_p$ of the p-dimensional Euclidean space is given in hyperspherical coordinates as

$$\mathrm{d}\mathbf{z} = \mathrm{J}(H_p)\, \mathrm{d}\rho\, \mathrm{d}\varphi_1 \cdots \mathrm{d}\varphi_{p-1} = \rho^{p-1} \left(\prod_{k=1}^{p-2} \sin^{p-k-1}\varphi_k \right) \mathrm{d}\rho\, \mathrm{d}\varphi_1 \cdots \mathrm{d}\varphi_{p-1}, \tag{32}$$

where $\mathrm{J}(H_p)$ is the Jacobian determinant of the transformation H_p, i.e.

$$\mathrm{J}(H_p) := \left| \det \frac{\partial(z_1, z_2, \ldots, z_p)}{\partial(\rho, \varphi_1, \ldots, \varphi_{p-1})} \right| = \rho^{p-1} \sin^{p-2}\varphi_1 \sin^{p-3}\varphi_2 \cdots \sin\varphi_{p-2}, \tag{33}$$

Moreover, the volume element of the $(p-1)$-sphere is given by

$$\mathrm{d}^{p-1}V = \sin^{p-2}\varphi_1 \sin^{p-3}\varphi_2 \cdots \sin\varphi_{p-2}\, \mathrm{d}\varphi_1\, \mathrm{d}\varphi_2 \cdots \mathrm{d}\varphi_{p-1}.$$

Thus the corresponding volume is then $V_{p-1} = 2\pi^{p/2}/\Gamma(p/2)$. Therefore, the multiple integral in (30) is transformed to

$$I := \int_{\mathbb{R}^p} \|z\|^2\, \mathrm{e}^{-\|z\|^{1/g}}\, \mathrm{d}z = V_{p-1} \int_{\mathbb{R}^+} \rho^2\, \rho^{p-1}\, \mathrm{e}^{-\rho^{1/g}}\, \mathrm{d}\rho. \tag{34}$$

Applying the transformation $u = u(\rho) := \rho^{1/g}$, $\rho \in \mathbb{R}^+$, with $\mathrm{d}\rho = g u^{g-1}\, \mathrm{d}u$, the integral (34) is then calculated, via the definition of the Gamma function, as

$$I = g\, V_{p-1} \int_{\mathbb{R}^+} u^{(p+2)\,g-1}\, \mathrm{e}^{-u}\, \mathrm{d}u = g\, V_{p-1}\, \Gamma\big((p+2)\,g\big), \tag{35}$$

hence, the inequality (30) is then reduced to

$$I_3 \leq 2g^{1-(p+2)\,g}\, \frac{\pi^{p/2}\, \sigma_1^{p+2}\, \Gamma\big((p+2)\,g\big)}{\nu\, \sigma_2^2\, \Gamma(p/2)}. \tag{36}$$

Applying (36) to (29) we finally derive the upper bound of D_{KL} as in (23).

Consider now the (multivariate) Normal distribution instead of the t_ν distribution. Then, following Theorem 2, we can derive an exact form of the KL divergence of the γ-GN over the usual Normal distribution, extending the corresponding univariate result in [51]. Note that, in order to achieve this result, the inequality proved in Theorem 2 is studied in limit, showing that the upper bounds in (23) increase along with the degrees of freedom ν of the t_ν-distribution, until they reach a supremum. Hence, when ν tends to infinity we are approaching normality as well as equality for (23).

Theorem 3 *The KL divergence of a p-variate r.v.* $X \sim \mathscr{N}_\gamma\big(\boldsymbol{\mu}, \sigma_1^2 \mathbb{I}_p\big)$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\sigma > 0$, *over a p-variate normally distributed r.v.* $N \sim \mathscr{N}\big(\boldsymbol{\mu}, \sigma_2^2 \mathbb{I}_p\big)$, *is given by*

$$D_{\mathrm{KL}}(X\|N) = \log\left\{\frac{2^{p/2-1}\,\Gamma(p/2)}{\Gamma\big(p\frac{\gamma-1}{\gamma}\big)}\big(\tfrac{\gamma-1}{\gamma}\big)^{p\frac{\gamma-1}{\gamma}-1}\right\} + p\left(\log\frac{\sigma_2}{\sigma_1} - \tfrac{\gamma-1}{\gamma}\right) + \big(\tfrac{\gamma}{\gamma-1}\big)^{2\frac{\gamma-1}{\gamma}}\big(\tfrac{\sigma_1}{\sigma_2}\big)^2 \frac{\Gamma\big((p+2)\frac{\gamma-1}{\gamma}\big)}{2\,\Gamma\big(p\frac{\gamma-1}{\gamma}\big)}. \tag{37}$$

Proof Firstly, by substituting of (26c) to (29), we obtain

$$D_{\mathrm{KL}}(X\|Y_\nu) = \log K + p\left(\log\frac{\sigma_1}{\sigma_2} - g\right) + \frac{\Gamma(p/2)}{4\pi^{p/2} g\,\Gamma(pg)}\, I, \tag{38}$$

where $g := (\gamma-1)/\gamma$, $Y_\nu \sim t_\nu\big(\boldsymbol{\mu}, \sigma_2^2 \mathbb{I}_p\big)$, $\nu \in \mathbb{N}^*$, and

$$I := \int_{\mathbb{R}^p} \mathrm{e}^{-\|\mathbf{z}\|^{1/g}} \log\left\{1 + \tfrac{1}{\nu}\big(\tfrac{\sigma_2}{\sigma_1}\big)^2 g^{-2g}\|\mathbf{z}\|^2\right\}^{p+\nu} \mathrm{d}\mathbf{z}. \tag{39}$$

For the KL divergence of $X \sim \mathscr{N}_\gamma\big(\boldsymbol{\mu}, \sigma_1^2 \mathbb{I}_p\big)$ over the p-variate normally distributed r.v. $N \sim \mathscr{N}\big(\boldsymbol{\mu}, \sigma_2^2 \mathbb{I}_p\big)$, it holds that $D_{\mathrm{KL}}(X\|N) = \lim_{\nu\to\infty} D_{\mathrm{KL}}(X\|Y_\nu)$, as the scaled spherically contoured $t_\nu\big(\mu, \sigma_2^2 \mathbb{I}_p\big)$ distribution is, in limit, the normal distribution $\mathscr{N}\big(\boldsymbol{\mu}, \sigma_2^2 \mathbb{I}_p\big)$ when $\nu \to \infty$. As a result, the sequence

$$b_\nu := \frac{\nu^{p/2}\,\Gamma(\nu/2)}{\Gamma\big(\frac{\nu+p}{2}\big)}, \quad \nu, p \in \mathbb{N}^*, \tag{40}$$

converges to $2^{p/2}$ as $\nu \to \infty$, since $\lim_{\nu\to\infty} f_{Y_\nu} = f_N$, where f_{Y_ν} and f_N are the probability densities of the t_ν-distributed r.v. Y_ν and the normally distributed r.v. N, respectively. Indeed, $b_\nu \to 2^{p/2}$, as $\nu \to \infty$, due to the fact that the normalizing factor $C_2(\sigma_2^2 \mathbb{I}_p)$ of f_{Y_ν} converges to the normalizing factor $C_1(\sigma_2^2 \mathbb{I}_p)$ of f_N, i.e. (18) and (22) yield $\pi^{-p/2} \lim_{\nu\to\infty} b_\nu^{-1} = (2\pi)^{-p/2}$, or equivalently $\lim_{\nu\to\infty} b_\nu = 2^{p/2}$.

Therefore, substituting $C_2(\sigma_2^2 \mathbb{N}_p)$ from (18) into (38), and then computing the limit for $\nu \to \infty$, we derive, using the limit in (40) as well as the well-known

exponential limit $\lim_{\nu\to\infty}(1+\nu^{-1})^{\nu}=\mathrm{e}$, that

$$D_{\mathrm{KL}}(X\|N)=\log\left\{2^{(p/2)-1}\frac{\Gamma(p/2)}{\Gamma(p\,g)}g^{p\,g-1}\right\}+p\Big(\log\frac{\sigma_2}{\sigma_1}-g\Big)+\frac{\Gamma(p/2)}{4\pi^{p/2}\,g\;\Gamma(p\,g)}\,I, \tag{41}$$

where

$$I=\big(\tfrac{\sigma_1}{\sigma_2}\big)^2 g^{-2g}\int_{\mathbb{R}^p}\|\mathbf{z}\|^2\,\mathrm{e}^{-\|\mathbf{z}\|^{1/g}}\,\mathrm{d}\mathbf{z}. \tag{42}$$

Calculating the above integral (42) with the help of (27), we derive

$$I=\frac{2\pi^{p/2}}{\Gamma(p/2)}\big(\tfrac{\sigma_1}{\sigma_2}\big)^2 g^{1-2g}\,\Gamma\big((p+2)g\big).$$

By substitution in (41), we finally obtain (37) using the known Gamma function additive identity, i.e. $\Gamma(x+1)=x\;\Gamma(x)$, $x\in\mathbb{R}^{*+}$.

The following investigates the order behavior of the upper bounds in (23).

Proposition 2 *When the degrees of freedom $\nu\in\mathbb{N}^*$ rise, the upper bound value, say $B_{\gamma,\nu}$ of (23) approximate better the KL divergence D_{KL} for all parameters $\gamma\in\mathbb{R}\setminus[0,1]$. Furthermore, for the univariate and the bivariate case, the corresponding bounds $B_{\gamma,\nu}$ have a strict descending order converging to the D_{KL} measure of r.v. $X\sim\mathscr{N}_{\gamma}\big(\boldsymbol{\mu},\sigma_1^2\mathbb{I}_p\big)$ over the normally distributed r.v. $N\sim\mathscr{N}\big(\boldsymbol{\mu},\sigma_2^2\,\mathbb{I}_p\big)$ as υ rises, i.e. $B_{\gamma,1}<B_{\gamma,2}<\cdots<B_{\gamma,\infty}=D_{\mathrm{KL}}(X\|N)$ for $p=1,2$.*

Proof Consider the sequence $a_\nu:=(\nu+1)/\nu$, $nu\in\mathbb{N}^*$. Then a_ν and b_ν, as in (40), converge both to 1 as $\nu\to\infty$. Considering the bounds $B_{\gamma,\nu}$ as in (23) when $\nu\to\infty$, it holds that $B_{\gamma,\infty}$ approaches the KL divergence as in (37). Thus, the equality in (23) is obtained in limit as $\nu\to\infty$, i.e. $D_{\mathrm{KL}}(X\|N)=B_{\gamma,\infty}$, and therefore the bounds $B_{\gamma,\nu}$ approximate better the KL divergence $D_{\mathrm{KL}}(X\|Y)$ as υ rises, until $B_{\gamma,\nu}$ coincides eventually with D_{KL} of Theorem 3 for all parameter γ values.

Especially for the bivariate case of $p:=2$, the sequence b_ν is constant, i.e. $b_\nu=2$, $\nu\in\mathbb{N}^*$, while for univariate case of $p:=1$, sequence b_ν is descending with $b_\nu\geq\lim_{\nu\to\infty}b_\nu=\sqrt{2}$. Indeed,

$$\frac{b_{2\nu+1}}{b_{2\nu}}=\tfrac{1}{\nu}\sqrt{\tfrac{2\nu+1}{2\nu}}\,\frac{\Gamma^2\left(\nu+\frac{1}{2}\right)}{\Gamma^2(\nu)},\qquad \nu\in\mathbb{N}^*.$$

By applying the known result of Gamma function,

$$\Gamma\left(k+\tfrac{1}{2}\right)=\tfrac{(2k-1)!!}{2^k}\sqrt{\pi}=\tfrac{(2k)!}{2^{2k}\,k!}\sqrt{\pi},\quad k\in\mathbb{N}, \tag{43}$$

we obtain

$$\frac{b_{2\nu+1}}{b_{2\nu}} = \pi\,\nu\,\sqrt{\frac{2\nu+1}{2\nu}}\left[\frac{(2\nu)!}{2^{2\nu}\,(\nu!)^2}\right]^2. \tag{44}$$

Finally, utilizing the known bounds for the factorial in (1), the ratio in (44) is less than 1, as

$$\frac{b_{2\nu+1}}{b_{2\nu}} \leq \frac{\mathrm{e}^2}{4\pi}\sqrt{\frac{2\nu+1}{2\nu}} \leq \frac{\mathrm{e}^2}{4\pi}\sqrt{\frac{3}{2}} \approx 0.72015 < 1.$$

Therefore, for dimensions $p = 1$ and $p = 2$, and from the form of bounds in Theorem 23, we derive that $B_{\gamma,1} < B_{\gamma,2} < \cdots < B_{\gamma,\infty}$. That is, as t_ν-distribution approaches the Normal distribution (as $\nu \to \infty$), the bounds $B_{\gamma,\nu}$ have a strictly descending order converging to $B_{\gamma,\infty}$, i.e. to $D_{\mathrm{KL}}(X\|N)$.

8 Discussion

Inequalities cover all the Mathematical disciplines, either as bounds to different quantities or measures—with typical example being the error control, as described in Sect. 2, or confidence intervals in Sect. 3—or as an attempt to compare different measures, like the notion of distance in Probability Theory, the SPRT method in Sect. 3, the various forms of triangle inequality given in Sect. 2, or in Information Theory as discussed in Sects. 5, 6 and 7. There are cases were the inequalities are involved either in definition, as in SPRT, or imposed as restriction to the developed theory, as in Stochastic Approximation. In Statistics, inequalities are often related with the interval estimation for the estimated parameters, usually through the Maximum Likelihood methodology. Sequences under imposed assumptions create different approaches in Statistics, with the main ones being the Sequential approach and the Stochastic processes.

A number of inequalities were presented in this paper. For example, consider the maximal inequalities in Sect. 4, or the Crossing Inequality that measures the times we can exceed the imposed bounds in a stochastic process; in the SPRT case, if this happens once, the method stops. Similar inequalities can also be considered under different lines of thought, with typical example being the Cauchy–Schwarz inequality in Sect. 2, which can be also be transferred and used in Statistics as shown in Sect. 3.

The inequalities in Information theory are more "mathematically oriented" and well-known bounds have been extended, with typical examples being the Information Inequality, the Cramér–Rao Inequality, or the Blachman–Stam Inequality. The upper bound of the Kullback–Leibler divergence, as proved in Sect. 7, is essential, we believe in the sense that offers a way of approximating "how far" can be the family of the generalized Normal distributions from the multivariate Student's t-distribution, since the involved integrals cannot be computed in a closed form.

Moreover, Proposition 2 gives us an idea of how those bounds behave in relation to the degrees of freedom of the considered t-distribution.

This paper can also be considered as an attempt to increase the existed inequality problems, collected by Rassias in [38].

Appendix 1

Proof of Proposition 1 It is easy to see that D^* satisfies the positive-definiteness and symmetricity conditions, and therefore—in order to prove that D^* is indeed a proper distance metric—the triangle inequality (or subadditivity) must be fulfilled. For this purpose, three arbitrary probability measures $P, Q, R \in \mathscr{P}(\Omega)$ are considered. Applying the exponential inequality $\mathrm{e}^x \geq (1 + x/n)^n$, $x \in \mathbb{R}$, with $n := 3$, to the definition of D^*, we get

$$D^*(P, Q)+D^*(Q, R) = \mathrm{e}^{D(P,Q)} + \mathrm{e}^{D(Q,R)} - 2 \geq \left[1+\tfrac{1}{3}D(P, Q)\right]^3 + \left[1+\tfrac{1}{3}D(Q, R)\right]^3 - 2,$$

and using the simplified notations $a := D(P, Q)$, $b := d(Q, R)$ and $c := d(P, R)$,

$$\begin{aligned} D^*(P, Q) + D^*(Q, R) &\geq \tfrac{1}{27}\left(a^3 + b^3\right) + \tfrac{1}{3}\left(a^2 + b^2\right) + a + b \\ &= \tfrac{1}{27}(a + b)^3 - \tfrac{1}{9}a\,b\,(a + b) + \tfrac{1}{3}(a + b)^2 - \tfrac{2}{3}a\,b + a + b \\ &\geq \tfrac{1}{27}(a + b)^3 - \tfrac{1}{36}(a + b)^3 + \tfrac{1}{3}(a + b)^2 - \tfrac{1}{6}(a + b)^2 + a + b \\ &\geq \tfrac{1}{3}c^3 + \tfrac{1}{6}c^2 + c, \end{aligned} \tag{45}$$

where the triangle inequality of metric D was used as well as the inequality $\sqrt{ab} \leq \frac{1}{2}(a+b)$, $a, b \in \mathbb{R}^+$. By expressing D in terms of D^*, through the definition of D^*, relation (45) yields

$$D^*(P, Q)+D^*(Q, R) \geq \tfrac{1}{3}\log^3\left(1+D^*(P, R)\right)+\tfrac{1}{6}\log^2\left(1+D^*(P, R)\right)+\log\left(1+D^*(P, R)\right). \tag{46}$$

Consider now the function $f(x) := \frac{1}{3}\log^3(1+x)+\frac{1}{6}\log^2(1+x)+\log(1+x)-x$, $x \in \mathbb{R}^+$. Assuming that $f' \leq 0$, i.e. $\log^2(1 + x) + \frac{1}{3}\log(1 + x) - x \leq 0$, the logarithm identity $\log x \geq (x-1)/x$, $x \in \mathbb{R}^{*+} := \mathbb{R}^+ \setminus \{0\}$ gives $4x^2 - 2x - 3 \leq 0$, which holds for $x \geq x_0 := \frac{1}{4}\left(2+\sqrt{28}\right) \approx 1.822$. Therefore, f has a global maxima at x_0, and as $x_1 = 0 = f(0)$ is one of the two roots $x_1, x_2 \in \mathbb{R}^+$ of f, the fact that $0 = x_1 \leq x_0$ means that $f(x) \geq 0$ for $x \in [0, x_2]$, where $x_2 \approx 3.5197$ (numerically computed). Therefore, the fact that metric $D \leq 1$ implies $0 \leq D^* \leq \mathrm{e} - 1 \approx 1.718 < x_2$, resulting (from the above discussion) that $f\left(D^*(P, Q)\right) \geq 0$ which is equivalent, through (46), to the requested triangle inequality $D^*(P, Q) + D^*(Q, R) \geq D^*(P, Q)$.

Appendix 2

Some introductory definitions from the Statistical Decision Theory are needed.

Definition 2 (Decision Problem and Rules) A general *decision problem* is defined to be a triplet (Θ, D, ℓ) and a random variable X, known as *data*, following the probability distribution $F(x \mid \theta), \theta \in \Theta$, with Θ being a parameter space. Moreover, θ is called as the *state of nature* with $\ell = \ell(\theta, d)$ denoting the *loss function*, while d is a *decision* from the *decision space* D. A *non-randomized decision rule* is a function $q(\cdot)$ such that $X \ni x \overset{q}{\mapsto} q(x) = d \in D$, while a *randomized decision rule* $q(x)$ specifies a probability distribution according to which a member, say d, of D is to be chosen.

Definition 3 (Risk Function) The *risk function* $r_\theta(q)$ of a decision rule q, for a decision problem (Θ, D, ℓ), is defined by $r_\theta(q) := \mathrm{E}\big(\ell\big(\theta, q(X)\big)\big)$, when θ is referring to the true state of nature, as appeared in the expected (or average) loss in the definition. To assign an *order* to decision rules, we assume that $r_\theta(q_1) > r_\theta(q_2)$, for every $\theta \in \Theta$, and say that q_2 is more preferable than q_1.

Now, let $\mathfrak{F}$ be the class of monotone non-decreasing functions f with $f(x) = 0$ on $I := [0, 1]$. Let D^* be a collection of sub-intervals of I, and Q_n be the set of decision rules q. We try to estimate $q(f)$ with $f(x) = 0$ and n observations, with the final decision to be $q(f) \in d$ for a particular $d \in D^*$. Obviously, $q(f) = d \in D^*$ defines a decision rule with n observations. We also consider the set, say Q_n^*, of all procedures $q \in Q_n$ for which $q(f) \in D$. An *optimum procedure* q_n^* is imposed as a minimax decision procedure, depending on the length $L(f, q)$ of d, with $f \in \mathscr{F}$, $q \in Q_n^*$ of the form

$$\sup_{f \in \mathfrak{F}} L\big(f, q_n^*\big) = \inf_{q \in Q_n^*} \sup_{f \in \mathfrak{F}} L(f, q).$$

Theorem 4 *The bisection method is a q_n^* minimax procedure.*

Proof Consider the iterative procedure $x_k := (\alpha_{k-1} + \beta_{k-1})/2$, $k = 2, 3, \ldots$, with initial value $x_1 := 1/2$. We assume that the value α_{k-1} corresponds to the largest previously observed value of x, with $f(x) = 0$ if there is no largest value for x, and $f(x) < 0$ if x is the largest value. We assume also that the value β_{k-1} corresponds to the smallest value of x, for which $f(x) = 1$ if there is no such x, and $f(x) > 0$ if x is the largest value. Let $d = [\alpha_n, \beta_n]$ be each time interval. In such a procedure with $n \geq 1$, any procedure $q \in Q_n^*$ with $x_1 \neq 1/2$ shall provide a larger $\sup_{f \in \mathfrak{F}} L(f, q)$. By induction, if we accept that theorem holds for $\nu = 1, 2, \ldots, n-1$, we shall try to prove it for $\nu = n$. Any procedure q with $n-1$ evaluations at $(x_1, x_2, \ldots, x_{n-1})$ does best to adopt the x_ν value midway between $\alpha_{\nu-1}$ and $\beta_{\nu-1}$, both evaluated via $(x_1, x_2, \ldots, x_{n-1})$. Hence, x_ν reaches a minimax length of $(\alpha_{\nu-1} - \beta_{\nu-1})/2$. However, by taking into account the values $x_1, x_2, \ldots, x_{\nu-1}$, then we are in accordance with $q_{\nu-1}^*$ which minimizes $\beta_{\nu-1} - \alpha_{\nu-1}$.

References

1. W.T. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley, New York, 2003)
2. N.M. Blachman, The convolution inequality for entropy powers. IEEE Trans. Inf. Theory **11**, 267–271 (1965)
3. E.A. Carlen, Superadditivity of Fisher's information and logarithmic Sobolev inequalities. J. Funct. Anal. **101**, 194–211 (1991)
4. D. Coin, A method to estimate power parameter in exponential power distribution via polynomial regression. J. Stat. Comput. Simul. **83**(11), 1981–2001 (2013)
5. A. Cotsiolis, N.K. Tavoularis, On logarithmic Sobolev inequalities for higher order fractional derivatives. C. R. Acad. Sci. Paris (Ser. I) **340**, 205–208 (2005)
6. T.M. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, New York, 2006)
7. M. Del Pino, J. Dolbeault, I. Gentil, Nonlinear diffusions, hypercontractivity and the optimal $\mathscr{L}^p$–Euclidean logarithmic Sobolev inequality. J. Math. Anal. Appl. **293**(2), 375–388 (2004)
8. J.L. Doob, *Stochastic Processes* (Wiley, New York, 1953)
9. T.S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach* (Academic, New York, 1967)
10. I. Ford, C.P. Kitsos, D.M. Titterington, Recent advances in nonlinear experimental design. Technometrics **31**, 49–60 (1989)
11. C. Fortuin, C. Kasteley, J. Ginibe, Correlation inequalities on some partially ordered sets. Commun. Math. Phys. **22**, 89–103 (1971)
12. M. Franceschetti, R. Meester, *Random Networks for Communication: From Statistical Physics to Information Systems* (Cambridge University Press, Cambridge, 2007)
13. L.-E. Fröberg, *Numerical Mathematics: Theory and Computer Applications* (Benjamin-Cummings, Menlo Park, 1985)
14. K.B. Ghosh, *Sequential Tests of Statistical Hypotheses* (Addison-Wesley, Reading, 1970)
15. E. Gómez, M.A. Gómez-Villegas, J.M. Martin, A multivariate generalization of the power exponential family of distributions. Commun. Stat. Theory Methods **27**(3), 589–600 (1998)
16. L. Gross, Logarithmic Sobolev inequalities. Am. J. Math. **97**(761), 1061–1083 (1975)
17. R.W. Hamming, *Numerical Methods for Engineers and Scientists* (McGraw-Hill, New York, 1962)
18. T. Harris, A lower bound for the critical probability in a certain percolation process. Proc. Camb. Philos. Soc. **56**(1), 13–20 (1960)
19. O. Johnson, A. Barron, Fisher information inequalities and the central limit theorem. Prob. Theory Relat. Fields **129**, 391–409 (2004)
20. S. Karlin, H.M. Taylor, *A First Course in Stochastic Processes* (Academic, New York, 1975)
21. J. Katz, A.J. Menezes, S.A. Vanstone, P.C. van Oorschot, *Handbook of Applied Cryptography* (CRC Press, Boca Raton, 1996)
22. C.P. Kitsos, Design and inference for nonlinear problems. PhD Thesis, University of Glasgow, 1986
23. C.P. Kitsos, Fully sequential procedures in nonlinear design problems. Comput. Stat. Data Anal. **8**, 13–19 (1989)
24. C.P. Kitsos, N.K. Tavoularis, Logarithmic Sobolev inequalities for information measures. IEEE Trans. Inf. Theory **55**(6), 2554–2561 (2009)
25. C.P. Kitsos, T.L. Toulias, Hellinger distance between generalized normal distributions. Br. J. Math. Comput. Sci. **21**(2), 1–16 (2017)
26. C.P. Kitsos, T.L. Toulias, Inequalities for the Fisher's information measures, in *Handbook of Functional Equations: Functional Inequalities*, ed. by Th.M. Rassias (Springer, New York, 2014), pp. 281–313
27. C.P. Kitsos, T.L. Toulias, New information measures for the generalized normal distribution. Information **1**, 13–27 (2010)

28. C.P. Kitsos, C.P. Toulias, P.C. Trandafir, On the multivariate γ–ordered normal distribution. Far East J. Theor. Stat. **38**(1), 49–73 (2012)
29. C.P. Kitsos, V.G. Vassiliadis, T.L. Toulias, MLE for the γ-order generalized normal distribution. Discuss. Math. Probab. Stat. **34**, 143–158 (2014)
30. A.W. Marshall, I. Olkin, Reversal of the Lyapunov, Hölder and Minkowski inequalities and other extensions of the Kantorovich inequality. J. Math. Anal. Appl. **8**, 503–514 (1964)
31. R. Meester, R. Roy, Uniqueness of unbounded and vacant components in Boolean models. Adv. Appl. Probab. **4**, 933–951 (1994)
32. O.S. Miettinen, Estimability and estimation in case-referent studies. Am. J. Epidemiol. **103**, 226–235 (1976)
33. C. Mortici, Improved asymptotic formulas for the gamma function. Comput. Math. Appl. **61**, 3364–3369 (2011)
34. S. Nadarajah, A generalized normal distribution. J. Appl. Stat. **32**(7), 685–694 (2005)
35. H. Nyquist, Certain factors affecting telegraph speed. Trans. Am. Inst. Electr. Eng. **XLIII**, 412–422 (1924)
36. S. Ramanujan, G.E. Andrews, *The Lost Notebook and Other Unpublished Papers* (Springer, Berlin, 1988)
37. Th.M. Rassias, On the stability of minimum points. Mathematica **45**, 93–104 (2003)
38. Th.M. Rassias, *Functional Equations, Inequalities and Applications* (Kluwer Academic, Dordrecht, 2003)
39. H. Robbins, A remark on Stirling's formula. Am. Math. Mon. **62**(1), 26–29 (1955)
40. H. Robbins, S. Monro, A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
41. M.J. Schervish, *Theory of Statistics* (Springer, New York, 1995)
42. C.E. Shannon, A mathematical theory of communication. Bell Syst. Technol. J. **27**, 379–423, 623–656 (1948)
43. S.D. Silvey, *Optimal Design – An Introduction to the Theory for Parameter Estimation* (Springer Netherlands, Dordrecht, 1980)
44. S.D. Silvey, The Lagrangian multiplier set. Ann. Math. Stat. **30**, 399–407 (1959)
45. S. Smirnov, W. Werner, Critical exponents for two dimensional percolation. Math. Res. Lett. **8**, 724–744 (2001)
46. S. Sobolev, On a theorem of functional analysis. AMS Transl. Ser. 2 (English translation) **34**, 39–68 (1963)
47. A.J. Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon. Inform. Control **2**, 255–269 (1959)
48. B. Torsney, Moment inequalities via optimal design theory. Linear Algebra Appl. **82**, 237–253 (1986)
49. T.L. Toulias, C.P. Kitsos, Generalizations of entropy and information measures, in *Computation, Cryptography and Network Security*, ed. by N.J. Darras, M.Th. Rassias (Springer, Cham, 2015), pp. 495–526
50. T.L. Toulias, C.P. Kitsos, Estimation aspects of the Michaelis-Menten model. REVSTAT Stat. J. **14**(2), 101–118 (2016)
51. T.L. Toulias, C.P. Kitsos, Kullback-Leibler divergence of the γ-ordered normal over t-distribution. Br. J. Math. Comput. Sci. **2**(4), 198–212 (2012)
52. A. Wald, Sequential tests of statistical hypotheses. Ann. Math. Stat. **16**, 117–186 (1945)
53. A. Wald, *Sequential Analysis* (Wiley, New York, 1947)
54. A. Wald, J. Wolfowitz, Optimum character of the sequential probability ratio test. Ann. Math. Stat. **19**, 326–339 (1948)
55. B. Woolf, On estimating the relation between blood group and disease. Am. Hum. Genet. **19**, 251–253 (1955)
56. C.F.J. Wu, H.P. Wynn, The convergence of general step-length algorithms for regular optimum design criteria. Ann. Stat. **6**, 1273–1285 (1978)

57. P.H. Wynn, Results in the theory and construction of D-optimal experimental design. J. R. Stat. Soc. B **34**(2), 133–147 (1972)
58. P.H. Wynn, The sequential generation of the D-optimal experimental designs. Ann. Math. Stat. **41**, 1655–1664 (1970)
59. S. Yu, A. Zhang, H. Li, A review of estimating the shape parameter of the generalized Gaussian distribution. J. Comput. Inf. Syst. **8**(21), 9055–9064 (2012)