entropy

MDPI

Editorial

Divergence Measures: Mathematical Foundations and Applications in Information-Theoretic and Statistical Problems

Igal Sason [1,2]

[1] Andrew & Erna Viterbi Faculty of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel; eeigal@technion.ac.il; Tel.: +972-4-8294699
[2] Faculty of Mathematics, Technion—Israel Institute of Technology, Haifa 3200003, Israel

Citation: Sason, I. Divergence Measures: Mathematical Foundations and Applications in Information-Theoretic and Statistical Problems. *Entropy* **2022**, *24*, 712. https://doi.org/10.3390/e24050712

Received: 3 May 2022
Accepted: 13 May 2022
Published: 16 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Data science, information theory, probability theory, statistical learning, statistical signal processing, and other related disciplines greatly benefit from non-negative measures of dissimilarity between pairs of probability measures. These are known as divergence measures, and exploring their mathematical foundations and diverse applications is of significant interest (see, e.g., [1–10] and references therein).

The present Special Issue, entitled *Divergence Measures: Mathematical Foundations and Applications in Information-Theoretic and Statistical Problems*, is focused on the study of the mathematical properties and applications of classical and generalized divergence measures from an information-theoretic perspective. It includes eight original contributions on the subject, which mainly deal with two key generalizations of the relative entropy: namely, the Rényi divergence and the important class of f-divergences. The Rényi divergence was introduced by Rényi as a generalization of relative entropy (relative entropy is a.k.a. the Kullback–Leibler divergence [11]), and it found numerous applications in information theory, statistics, and other related fields [12,13]. The notion of an f-divergence, which was independently introduced by Ali-Silvey [14], Csiszár [15–17], and Morimoto [18], is a useful generalization of some well-known divergence measures, retaining some of their major properties, including data-processing inequalities. It should be noted that, although the Rényi divergence of an arbitrary order is not an f-divergence, it is a one-to-one transformation of a subclass of f-divergences, so it inherits some of the key properties of f-divergences. We next describe the eight contributions in this Special Issue, and their relation to the literature.

Relative entropy is a well-known asymmetric and unbounded divergence measure [11], whereas the Jensen-Shannon divergence [19,20] (a.k.a. the capacitory discrimination [21]) is a bounded symmetrization of relative entropy, which does not require the pair of probability measures to have matching supports. It has the pleasing property that its square root is a distance metric, and it also belongs to the class of f-divergences. The latter implies, in particular, that the Jensen–Shannon divergence satisfies data-processing inequalities. The first paper in this Special Issue [22], authored by Nielsen, studies generalizations of the Jensen–Shannon divergence and the Jensen–Shannon centroid. The work in [22] further suggests an iterative algorithm for the numerical computation of the Jensen–Shannon-type centroids for a set of probability densities belonging to a mixture family in information geometry. This includes the case of calculating the Jensen–Shannon centroid of a set of categorical distributions or normalized histograms.

Many of Shannon's information measures appear naturally in the context of horse gambling, when the gambler's utility function is the expected log-wealth. The second paper [23], coauthored by Bleuler, Lapidoth, and Pfister, shows that, under a more general family of utility functions, gambling also provides a context for some of Rényi's information measures. Motivated by a horse betting problem in the setting where the gambler has side information, a new conditional Rényi divergence is introduced in [23]. It is compared with the conditional Rényi divergences by Csiszár and Sibson, and the properties of all the three are studied in depth by the authors, with an emphasis on the behavior of these conditional divergence measures under data processing. In the same way that Csiszár's and Sibson's conditional divergences lead to the respective dependence measures, so does the new conditional divergence in [23] lead to the Lapidoth–Pfister mutual information. The authors further demonstrate that their new conditional divergence measure is also related to the Arimoto–Rényi conditional entropy and to Arimoto's measure of dependence. In the second part of [23], the horse betting problem is analyzed where, instead of Kelly's expected log-wealth criterion, a more general family of power-mean utility functions is considered. The key role in the analysis is played by the Rényi divergence, and the setting where the gambler has access to side information provides an operational meaning to the Lapidoth–Pfister mutual information. Finally, a universal strategy for independent and identically distributed races is presented in [23] which, without knowing the winning probabilities or the parameter of the utility function, asymptotically maximizes the gambler's utility function.

The relative entropy [11] and the chi-squared divergence [24] are classical divergence measures which play a key role in information theory, statistical machine learning, signal processing, statistics, probability theory, and many other branches of mathematics. These divergence measures are fundamental in problems pertaining to source and channel coding, large deviations theory, tests of goodness-of-fit and independence in statistics, expectation–maximization iterative algorithms for estimating a distribution from an incomplete data, and other sorts of problems. They also belong to the generalized class of f-divergences. The third paper [25], by Nishiyama and Sason, studies integral relations between the relative entropy and chi-squared divergence, the implications of these relations, their information-theoretic applications, and some generalizations pertaining to the rich class of f-divergences. Applications that are studied in [25] include lossless compression, the method of types and large deviations, strong data-processing inequalities, bounds on contraction coefficients and maximal correlation, and the convergence rate to the stationarity of a type of discrete-time Markov chain.

The interesting interplay between inequalities and information theory has a rich history, with notable examples that include the relationship between the Brunn–Minkowski inequality and the entropy power inequality, transportation-cost inequalities and their tight connections to information theory, logarithmic Sobolev inequalities and the entropy method, inequalities for matrices obtained from the nonnegativity of relative entropy, connections between information inequalities and finite groups, combinatorics, and other fields of mathematics (see, e.g., [26–30]). The fourth paper by Reeves [31] considers applications of a two-moment inequality for the integral of fractional power of a function between zero and one. The first contribution of this paper provides an upper bound on the Rényi entropy of a random vector, expressed in terms of the two different moments. This also recovers some previous results based on maximum entropy distributions under a single moment constraint. The second contribution in [31] is a method for upper bounding mutual information in terms of certain integrals with respect to the variance of the conditional density.

Basic properties of an f-divergence are its non-negativity, convexity in the pair of probability measures, and the satisfiability of data-processing inequalities as a result of the convexity of the function f (and by the requirement that f vanishes at 1). These properties lead to f-divergence inequalities, and to information-theoretic applications (see, e.g., [4,10,32–37]). Furthermore, tightened (strong) data-processing inequalities for f-divergences have been of recent interest (see, e.g., [38–42]). The fifth paper [43], authored by Melbourne, is focused on the study of how stronger convexity properties of the function f imply improvements of classical f-divergence inequalities. It provides a systematic study of strongly convex divergences, and it quantifies how the convexity of a divergence generator f influences the behavior of the f-divergence. It proves that every (so-called) strongly convex divergence dominates the square of the total variation, which extends the classical bound provided by the chi-squared divergence. Its analysis also yields improvements of Bayes risk f-divergence inequalities, consequently achieving a sharpening of Pinsker's inequality.

Divergences between probability measures are often used in statistics and data science in order to perform inference under models of various types. The corresponding methods extend the likelihood paradigm, and suggest inference in settings of minimum distance or minimum divergence, while allowing some tradeoff between efficiency and robustness. The sixth paper [44], authored by Broniatowski, considers a subclass of f-divergences, which contains most of the classical inferential tools, and which is indexed by a single scalar parameter. This class belongs to the family of f-divergences, and is usually referred to as the power divergence class, which has been considered by Cressie and Read [7,45]. The work in [44] states that the most commonly used minimum divergence estimators are maximum-likelihood estimators for suitably generalized bootstrapped sampling schemes. It also considers optimality of associated goodness-of-fit tests under such sampling schemes.

The seventh paper by Verdú [46] is a research and tutorial paper on error exponents and α-mutual information. Similarly to [23] (the second paper in this Special Issue), it relates to Rényi's generalization of the relative entropy and mutual information. In light of the landmark paper by Shannon [47], it is well known that the analysis of the fundamental limits of noisy communication channels in the regime of vanishing error probability (by letting the blocklength of the code tend to infinity) leads to the introduction of the channel capacity as the maximal rate which enables to obtain reliable communication. The channel capacity is expressed in terms of a basic information measure: the input–output mutual information maximized over the input distribution. Furthermore, in the regime of fixed nonzero error probability, the asymptotic fundamental limit is a function of not only the channel capacity but the channel dispersion, which is expressible in terms of an information measure: the variance of the information density obtained with the capacity-achieving distribution [48]. In the regime of exponentially decreasing error probability, at fixed code rate below capacity, the analysis of the fundamental limits has gone through three distinct phases: (1) the early days of information theory and the error exponents analysis at MIT; (2) expressions for the error exponent functions by incorporating the relative entropy; and (3) the error exponent research with Rényi information measures. Thanks to Csiszár's realization of the relevance of Rényi's information measures to this problem [32], the third phase has found a way to express the error exponent functions as a function of generalized information measures, and also to solve the associated optimization problems in a systematic way. While in the absence of cost constraints, the problem reduces to finding the maximal α-mutual information, cost constraints make the problem significantly more challenging. The remained gaps in the interrelationships between three approaches, in the general case of cost-constrained encoding, motivated the present study in [46]. Furthermore, no systematic approach has been suggested so far for solving the attendant optimization problems by exploiting the specific structure of the information functions. The work by Verdú in [46] closes those gaps, while proposing a simple method to maximize the Augustin–Csiszár mutual information of order α under cost constraints [32,49], by means of the maximization of the α-mutual information subject to an exponential average constraint.

In statistical inference, the information-theoretic performance limits can often be expressed in terms of a statistical divergence measure between the underlying statistical models (see, e.g., [50] and references therein). As the data dimension grows, computing the statistics involved in decision making and the attendant performance limits (divergence measures) face complexity and stability challenges. Dimensionality reduction addresses these challenges at the expense of compromising performance because of the attendant loss of information. The eighth and last paper in the present Special Issue [51] considers linear dimensionality reduction, such that the divergence between the models is maximally preserved. Specifically, this work is focused on Gaussian models where discriminant analysis under several f-divergence measures are considered. The optimal design of the linear transformation of the data onto a lower-dimensional subspace is characterized for zero-mean Gaussian models, and numerical algorithms are employed to find the design for general Gaussian models with non-zero means.

It is our hope that the reader will find interest in the eight original contributions of this Special Issue, and that these works will stimulate further research in the study of the mathematical foundations and applications of divergence measures.

Acknowledgments: The Guest Editor is grateful to all the authors for their contributions to this Special Issue, to the anonymous peer-reviewers for their timely reports and constructive feedback.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
2. Broniatowski, M.; Stummer, W. Some universal insights on divergences for statistics, machine learning and artificial intelligence. In *Geometric Structures of Information*; Nielsen, F., Ed.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 149–211. [CrossRef]
3. Csiszár, I.; Shields, P.C. Information Theory and Statistics: A Tutorial. *Found. Trends Commun. Inf. Theory* **2004**, *1*, 417–528. [CrossRef]
4. Esposito, A.R.; Gastpar, M.; Issa, I. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Trans. Inf. Theory* **2021**, *67*, 4986–5004. [CrossRef]
5. Liese, F.; Vajda, I. *Convex Statistical Distances*; Teubner-Texte Zur Mathematik: Leipzig, Germany, 1987; Volume 95.
6. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [CrossRef]
7. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006.
8. Stummer W.; Vajda, I. On divergences of finite measures and their applicability in statistics and information theory. *Statistics* **2010**, *44*, 169–187. [CrossRef]
9. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989.
10. Zakai, M.; Ziv, J. A generalization of the rate-distortion theory and applications. In *Information Theory—New Trends and Open Problems*; Longo, G., Ed.; Springer: Berlin/Heidelberg, Germany, 1975; pp. 87–123. [CrossRef]
11. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
12. Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561. Available online: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf (accessed on 12 May 2022).
13. Van Erven, T.; Harremoës, P. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
14. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc.* **1966**, *28*, 131–142. [CrossRef]
15. Csiszár, I. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.* **1963**, *8*, 85–108.
16. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Stud. Sci. Math. Hung.* **1967**, *2*, 299–318.
17. Csiszár, I. On topological properties of f-divergences. *Stud. Sci. Math. Hung.* **1967**, *2*, 329–339.
18. Morimoto, T. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331. [CrossRef]
19. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
20. Menéndez, M.L.; Pardo, J.A.; Pardo, L.; Pardo, M.C. The Jensen–Shannon divergence. *J. Frankl. Inst.* **1997**, *334*, 307–318. [CrossRef]
21. Topsøe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609. [CrossRef]
22. Nielsen, F. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy* **2020**, *22*, 221. [CrossRef]
23. Bleuler, C.; Lapidoth, A.; Pfister, C. Conditional Rényi divergences and horse betting. *Entropy* **2020**, *22*, 316. 10.3390/e22030316. [CrossRef]
24. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
25. Nishiyama, T.; Sason, I. On relations between the relative entropy and χ^2-divergence, generalizations and applications. *Entropy* **2020**, *22*, 563. [CrossRef]
26. Dembo, A.; Cover, T.M.; Thomas, J.A. Information theoretic inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501–1518. [CrossRef]

27. Madiman, M.; Mellbourne, J.; Xeng, P. Forward and reverse entropy power inequalities in convex geometry. In *Convexity and Concentration*; Carlen, E., Madiman, M., Werner, E.M., Eds.; IMA Volumes in Mathematics and Its Applications; Springer: Berlin/Heidelberg, Germany, 2017, Volume 161, pp. 427–485. [CrossRef]
28. Marton, K. Distance-divergence inequalities. *IEEE Inf. Theory Soc. Newsl.* **2014**, *64*, 9–13.
29. Boucheron, S.; Lugosi, G.; Massart, P. *Concentration Inequalities—A Nonasymptotic Theory of Independence*; Oxford University Press: Oxford, UK, 2013.
30. Raginsky, M.; Sason, I. *Concentration of Measure Inequalities in Information Theory, Communications and Coding*, 3rd ed.; Now Publishers: Delft, The Netherlands, 2018. [CrossRef]
31. Reeves, G. A two-moment inequality with applications to Rényi entropy and mutual information. *Entropy* **2020**, *22*, 1244. [CrossRef] [PubMed]
32. Csiszár, I. A class of measures of informativity of observation channels. *Period. Mat. Hung.* **1972**, *2*, 191–213. [CrossRef]
33. Guntuboyina, A.; Saha, S.; Schiebinger, G. Sharp inequalities for f-divergences. *IEEE Trans. Inf. Theory* **2014**, *60*, 104–121. [CrossRef]
34. Harremoës, P.; Vajda, I. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory* **2011**, *57*, 3230–3235. [CrossRef]
35. Merhav, N. Data processing theorems and the second law of thermodynamics. *IEEE Trans. Inf. Theory* **2011**, *57*, 4926–4939. [CrossRef]
36. Sason, I.; Verdú, S. f-divergence inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
37. Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef]
38. Calmon, F.P.; Polyanskiy, Y.; Wu, Y. Strong data processing inequalities for input constrained additive noise channels. *IEEE Trans. Inf. Theory* **2018**, *64*, 1879–1892. [CrossRef]
39. Raginsky, M. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Trans. Inf. Theory* **2016**, *62*, 3355–3389. [CrossRef]
40. Polyanskiy, Y.; Wu, Y. Strong data processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*; Carlen, E., Madiman, M., Werner, E.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 161, pp. 211–249. [CrossRef]
41. Makur, A.; Zheng, L. Linear bounds between contraction coefficients for f-divergences. *Probl. Inf. Transm.* **2020**, *56*, 103–156. [CrossRef]
42. Sason, I. On data-processing and majorization inequalities for f-divergences with applications. *Entropy* **2019**, *21*, 1022. [CrossRef]
43. Melbourne, J. Strongly convex divergences. *Entropy* **2020**, *22*, 1327. [CrossRef] [PubMed]
44. Broniatowski, M. Minimum divergence estimators, maximum likelihood and the generalized bootstrap. *Entropy* **2021**, *23*, 185. [CrossRef] [PubMed]
45. Cressie, N.; Read, T.R.C. Multinomial Goodness-of-Fit Tests. *J. R. Stat. Soc. Ser. (Methodol.)* **1984**, *46*, 440–464. [CrossRef]
46. Verdú, S. Error Exponents and α-Mutual Information. *Entropy* **2021**, *23*, 199. [CrossRef]
47. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. 623–656. [CrossRef]
48. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
49. Augustin, U. Noisy Channels. Ph.D. Thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, 1978. Available online: http://libgen.rs/book/index.php?md5=4431004A794657A85AA356F111AA52FE (accessed on 12 May 2022).
50. Sason, I.; Verdú, S. Arimoto-Rényi conditional entropy and Bayesian M-ary hypothesis testing. *IEEE Trans. Inf. Theory* **2018**, *64*, 4–25. [CrossRef]
51. Dwivedi, A.; Wang, S.; Tajer, A. Discriminant analysis under f-divergence measures. *Entropy* **2022**, *24*, 188. [CrossRef]