Front. Electr. Electron. Eng. China 2010, 5(3): 237–240
DOI 10.1007/s11460-010-0100-4

EDITORIAL

Lei XU, Yanda LI

Emerging themes on information theory and Bayesian approach

Though efforts on the quantification of information started several decades earlier, the foundations of information theoretic studies were laid during the middle and late 1940's, from two perspectives that both based on probability theory. The most famous one is a systematic theory from a perspective of information transmission over a noisy channel, namely the information theory developed by Claude E. Shannon [1]. The other consists of some fundamental results obtained along the line of Fisher information from statistical inference perspective, featured by the Cramér-Rao bound and Rao's finding that the Fisher information metric is a Riemannian metric in term of differential geometry [2]. The first perspective puts attention on the relation between bits to code samples from a probability distribution P and channel capacity, while the second perspective puts attention on the discrepancy of using a parametric probability distribution Q to represent samples from the unknown P. Efforts of two perspectives intersect by the Kullback Leibler (KL) divergence (also information gain, relative entropy) [3], which is not only a non-symmetric measure of the difference between P and Q, but also a measure of the expected number of extra bits required to code samples from P when using a code based on Q.

Several decades of efforts have been made on statistical inference or recently on machine learning via minimizing the KL divergence. It not only includes the maximum likelihood estimation as a special case, but also implements other estimation principles with certain constraints imposed on P and/or Q, e.g., maximum entropy or cross-entropy minimization [4,5]. Following Ref. [2], the Fisher information metric is found to be a unique intrinsic metric and the family of α-affine is a unique one parameter family of affine connections [6]. In the 1980's, Amari and colleagues further introduced a new concept of dual-affine connections and got a new finding that the family of dual divergence (including KL divergence) defines uniquely a Riemannian metric and a family of dual affine connections [7], which brought all the results together into a theory under the name of information geometry. This theory provides a useful and strong tool to many areas of information sciences and engineering. In the first article of this issue, Amari not only provides an easy understanding tutorial on fundamentals about information geometry in a manifold of probability distributions, but also introduces several emerging topics of this theory.

Amari's article also gives extensions of information geometry to cover the manifolds of nonnegative matrices and visual signals. Nonnegativity has been shown to be a powerful principle in linear matrix decompositions, resulting in sparse component matrices for applications of feature analysis and data compression. In recent years, a lot of efforts have been made on this topic under the name of nonnegative matrix factorization (NMF). In the second article of this special issue, Oja and Yang suggests to integrate an orthonormality constraint into NMF, and shows how multiplicative updating rules are obtained to find a nonnegative and highly orthogonal matrix for approximated graph partitioning problems, outperforming those partitioning approaches without the orthogonality condition.

Moreover, Sects. 3, 4, and 5 of Amari's article introduce recent progresses on information geometry of alternative minimization, of Bayesian Ying-Yang (BYY) system, of belief propagation in networks, respectively, which provides

Received May 31, 2010

Lei XU
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
E-mail: lxu@cse.cuhk.edu.hk

Yanda LI
Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: daulyd@tsinghua.edu.cn

a new tool for analyzing these studies and developing new learning algorithms in help of dual affine/flat connections and of natural gradient searching by considering the curvatures of sub-manifolds. These studies are all about that observations of X which are generated via $q(X|Y,\Theta_{x|y})$ from an inner or hidden representation Y with a priori $q(Y|\Theta_y)$ and that we inversely encode X into Y as well as estimating the parameters $\Theta=\{\Theta_{x|y},\Theta_y\}$. Typically, we estimate Y by the Bayesian posteriori $q(Y|X,\Theta)=q(X|Y,\Theta_{x|y})q(Y|\Theta_y)/q(X|\Theta)$ and estimate Θ by maximizing the likelihood $q(X|\Theta)=\int q(X|Y,\Theta_{x|y})q(Y|\Theta_y)\mathrm{d}Y$, implemented by the EM algorithm [8]. Section 3 of Amari's article shows that the EM algorithm is a particular alternation of e-projection and m-projection, which provides not only a generic link between Bayesian approach and Fisher information based geometry theory but also a practical guide for developing learning algorithms. Furthermore, it follows that getting a maximum posterior (MAP) estimator of Y by $\max_Y q(Y|X,\Theta)$ is equivalent to minimizing the length of two part encoding $-\ln q(X|Y,\Theta_{x|y})-\ln q(Y|\Theta_y)$ from a perspective of Shannon information theory.

The above Bayesian and information theoretic based formulation has found applications in many machine learning problems and in various scientific and engineering fields, with different probabilistic models for $q(X|Y,\Theta_{x|y})$ and $q(Y|\Theta_y)$. One typical example is computer vision and pattern recognition [9]. Vision can be considered to be a decoding problem where the encoding of information is performed by the physics of the world — by light rays striking objects and being reflected to cameras or eyes. In the past decade, there have been many efforts on developing probabilistic models which are capable of capturing the richness of visual stimuli and hence are efficient at encoding them. In the 5th article of this issue, Yuille provides a comprehensive introduction on these efforts, especially efficient inference and learning algorithms, from an information theory perspective. The formulations of $q(X|Y,\Theta_{x|y})$ and $q(Y|\Theta_y)$ are featured by not only the classic Markov random field models, but also recent image models in help of stochastic grammars and hierarchical representations. Also, Sect. 5 of Amari's article in this special issue has investigated the information geometry of one of Yuille's algorithms, namely the CCCP, which indicates a potential for studying other algorithms from a similar perspective too.

It is difficult computationally to get the Bayesian posteriori $q(Y|X,\Theta)$ from probabilistic image models for $q(X|Y,\Theta_{x|y})$ and $q(Y|\Theta_y)$, since it needs to handle a summation or integral over Y to get $q(X|\Theta)$. Still, solving $\max_Y[q(X|Y,\Theta_{x|y})q(Y|\Theta_y)]$ is computational expensive too. To facilitate implementation, $q(Y|X,\Theta)$ is approximated by a simpler $p(Y|X,\Theta)$, in help of the mean field free energy or the Bethe free energy in Sect. 5 of Yuille's article. Alternatively, this approximation can also be explained from a best Ying-Yang matching perspective [10] with a Ying machine $q(X|Y,\Theta_{x|y})q(Y|\Theta_y)$ from probabilistic image models and a Yang machine $p(Y|X,\Theta)p(X)$ from image data and the above approximation. Actually, minimizing the mean field or Bethe free energy is equivalent to minimizing the divergence $\mathrm{KL}(p||q)$ between $p(Y|X,\Theta)p(X)$ and $q(X|Y,\Theta_{x|y})q(Y|\Theta_y)$. A similar scenario applies to the Helmholtz machine [11] too.

Even generally, this formulation applies to many applications with different structures for $q(X|Y,\Theta_{x|y})$ and $q(Y|\Theta_y)$, as well as $p(Y|X,\Theta)$, respectively. In Sect. 4 of his article in this special issue, Amari also provides an information geometry perspective on this type of Ying-Yang system, which gives a guide line for developing new algorithms. Actually, this Ying-Yang system merely handle the mappings $Y\to X$ and $X\to Y$, while minimizing the divergence $\mathrm{KL}(p||q)$ is only able to estimate Y and Θ for a Ying-Yang best matching, to which the initial efforts in 1995 about BYY learning [10] mainly devoted. Also, Eq. (24) in Ref. [10] started an effort on the BYY learning based criterion for selecting the cluster number k. During the period 1996–1999, it first evolved into a model selection criterion $J_2(k)$ for Gaussian mixture, factor analysis, etc., and then further reached its generic formulation $H(p||q)$ as a Ying-Yang harmony measure [12]. With Yang structure designed from Ying structure according to a variety preservation principle, both parameters Θ and model complexity $\boldsymbol{k}$ are determined by maximizing $H(p||q)$, which makes Ying and Yang reach a best agreement in a most tacit way (with a least amount of information communication) or become a best matching pair in a most compact form with a least complexity. Interestingly, minimizing $\mathrm{KL}(p||q)$ and maximizing $H(p||q)$ are two typical cases of the Radon-Nikodym derivative based harmony functional $H_\mu(P||Q)$. Also, the above Ying-Yang formulation $p(Y|X,\Theta)p(X)$ and $q(X|Y,\Theta_{x|y})q(Y|\Theta_y)$ is only a special case of the general framework $q(X|R)q(R)$ and $p(R|X)p(X)$ with a representation $R=\{Y,\Theta,\boldsymbol{k}\}$, a priori $q(\Theta)$, and a posteriori $p(\Theta|X)$ all in consideration. Details are referred to Xu's article in this special issue. Furthermore, Sect. 4 of Amari's article may be extended to this general framework, while maximizing $H(p||q)$ arises a further information geometry topic too.

Efforts on model complexity $\boldsymbol{k}$ have been made extensively for decades. The first KL divergence based study was pioneered by Akaike information criterion (AIC) in 1974 [13], which is derived from that the maximum likelihood estimator is asymptotically subject to a Gaussian distribution with the covariance matrix equal to the inverse of the Fisher information matrix divided by the sample size. Efforts that take a priori $q(\Theta)$ in consideration for model complexity $\boldsymbol{k}$ even started from the early 1960's, which were pursuit by several pioneers along this direction

(unfortunately, many of them left in recent few years). Their results are featured by different but closely related principles, namely algorithmic probability (AP) [14], Kolmogorov complexity (KC) [15], minimum message length (MML) [16], Bayesian information criterion (BIC) [17], and minimum description length (MDL) [18], as well as renewed waves since the early 1990's under the name of Bayesian, marginal Bayesian, and variational Bayesian [19,20]. The spirit is that the goodness of hypothesis (algorithm/program/model) is measured by the length of strings which encodes/computes the data or a computable object according to that hypothesis. The hypothesis H that corresponds to the shortest length is preferred, which reflects the Occam's razor, namely simple explanations of data are preferable to complex ones.

Except AP and KC, all the other studies mentioned above differ in estimating the coding length, based on different approximations of the marginal $q(X|H) = \int q(X|\Theta)q(\Theta)\mathrm{d}\Theta$, e.g., Laplace approximation or variational approximation. These studies also differ in considerations on priories of hypotheses. One prefers that $q(\Theta)$ is given by a model-dependent priori (e.g., MML), while other prefers that $q(\Theta)\mathrm{d}\Theta$ gets an equal weight for any Θ or all the distributions in a model family get equal prior weight (e.g., BIC, MDL). Among these efforts, the MDL studies by Rissanen have grown into the most systematic and deepest stream, which actually activated the waves of model selection studies and popularity of Bayesian approach in the past one or two decades. In the 3rd article of this special issue, Rissanen outlines his new results, namely a generic estimation theory that defines optimality for all sizes of data instead of only asymptotically, covering estimation of both real-valued parameters and their number, i.e., model selection. The objective is to fit 'models' as distributions to the data in order to find the regular statistical features. The performance of the fitted models is measured by the probability they assign to the data, i.e., a large probability means a good fit and a small probability a bad fit. Rissanen further tells us that there are three equivalent characterizations of optimal estimators, the first defined by estimation capacity, the second to satisfy necessary conditions for optimality for all data, and the third by the complete Minimum Description Length (MDL) principle.

The 4th article in this special issue proceeds to introduce fundamentals of BYY learning and to give a tutorial on learning algorithms for several typical learning tasks. The BYY learning provides not only a general framework that accommodates typical learning approaches from a unified perspective but also a different road towards parameter learning and model selection, that is, via maximizing $H(p||q)$ for Ying-Yang best harmony. Instead of taking a latent role in $q(X|\Theta) = \int q(X|Y,\Theta_{x|y})q(Y|\Theta_y)\mathrm{d}Y$ for a maximum likelihood estimator on $\Theta, q(Y|\Theta_y)$ in the maximization of $H(p||q)$ takes a role of an equal importance to $q(\Theta)$ for model selection on the primary part of complexity about the representation Y, which results in new features of improved model selection criteria, improved performances of automatic model selection, and a coordinated implementation of Ying based model selection and Yang based learning regularization.

In fact, the earliest efforts of encoding parameters and model complexity should be backtracked to the theory of universal inductive inference or algorithmic probability (AP) by Solomonoff [14], which was seemly proposed even earlier than Kolmogorov complexity (KC) [15] — another similar but independently proposed idea. AP is a method of assigning a probability to each hypothesis (algorithm/program) that explains a given observation, with the simplest hypothesis (the shortest program) having the highest probability and the increasingly complex hypotheses receiving increasingly small probabilities. These probabilities form a priori probability distribution for the observation, and then Bayes theorem is used to predict the most likely continuation of that observation. In the 6th article in this special issue, Schmidhuber provides an overview on the emergence of "new" AI, namely general artificial intelligence that aims at making optimally reliable statements about future events, given the past under certain broad computability assumptions. Rooted in the work of Solomonoff and Kolmogorov, recent years have brought substantial progress in the field of computable and feasible variants of optimal algorithms for prediction, search, inductive inference, decision making, and reinforcement learning in general environments.

Instead of focusing on a best encoding of past observations to make reliable statements about future events, the Shannon information theory consists of source coding theorem that on average the number of bits needed to represent the result of an uncertain event is given by its entropy, and channel coding theorem that reliable communication is possible over noisy channels provided that the rate of communication is below the channel capacity. This theory asserts that asymptotic optimality can be achieved by separating source coding and channel coding. The goal of source coding is to represent the information source in fair bits. The goal of channel coding is to enable the transmission of fair bits through the channel essentially free of error with no reference to the meaning of these fair bits. This elegant theory makes it got broad and deep applications for several decades, also including recent network communication. For a network consisting of noiseless point-to-point communication channels, routers are deployed at the intermediate nodes to switch a data packet from an input channel to an output channel without processing the data content. In this special issue, the 7th article provides an introduction to an emerging area of

network communication, called network coding. Instead of simply routing and/or replicating information within the network, coding is also employed at the intermediate nodes in order to achieve bandwidth optimality. In this article, Yeung provides a tutorial on fundamentals of network coding, with applications to various areas such as channel coding, wireless communication, computer networks, etc.

As a vast amount of data have been accumulating, bioinformatics research takes ever increasing important roles in genomic research and life science, which also provides new challenges and great opportunities for studies of information theory and Bayesian approach. In the 8th article of this special issue, Chen introduces a interdisciplinary field — bioinformatics, he points out that bioinformatics uses genomic DNA sequence analysis as its information source to discovering protein and RNA genes embedded in the sequences, elucidating the informational content of non-protein coding sequence, and deciphering the grammatical code of the genetic language contained in the DNA sequence, while at the same time inferring, ordering and releasing the information of the genetic language and its RNA and protein profiles, thereby learning the rules governing metabolism, development, differentiation and evolution. A central aim of bioinformatics is to bring to light "the complexity of the structure of genomic information and the basic rules of the genetic language" and thus ultimately leading to a better understanding of human. Actually, DNA sequence is only a genetic information storage system, in the last article of the special issue, Li further proposes an idea that an organism may be considered as an information system in nature. His paper analyzes this idea from different ways: DNA sequence satisfies the basic requirements of an information system, the controls of a man and a robot both obey the principle of cybernetics, and why a man can have ideas but a robot has no such capacity.

Lei XU
Guest Editor of the special issue

Yanda LI
Editor-in-Chief of *Frontiers of Electrical and Electronic Engineering in China*

References

1. Shannon C E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27: 379–423, 623–656
2. Rao C R. Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 1945, 37: 81–91
3. Kullback S, Leibler R A. On information and sufficiency. Annals of Mathematical Statistics, 1951, 22 (1): 79–86
4. Jaynes E T. Information theory and statistical mechanics. Physical Review, 1957, 106(4): 620–630
5. Shore J, Johnson R. Properties of cross-entropy minimization. IEEE Transactions on Information Theory, 1981, 27(4): 472–482
6. Chentsov N N. Statistical Decision Rules and Optimal Inference, Translations of Mathematical Monographs; v. 53. American Mathematical Society, 1982
7. Amari S. Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, Berlin: Springer-Verlag, 1985
8. Dempster A P, Laird N M, Rubin D B. Maximum-likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B, 1977, 39(1): 1–38
9. Yuille A L, Kersten D. Vision as Bayesian inference: Analysis by synthesis? Trends in Cognitive Sciences, 2006, 10(7): 301–308
10. Xu L. YING-YANG machines: A Bayesian-Kullback scheme for unified learning and new results on vector quantization. In: Proceedings of the International Conference on Neural Information Processing (ICONIP95). 1995, 977–988
11. Hinton G E, Dayan P, Frey B J, Neal R N. The wake-sleep algorithm for unsupervised learning neural networks. Science, 1995, 268(5214): 1158–1160
12. Xu L. Temporal BYY learning for state space approach, hidden Markov model and blind source separation. IEEE Transactions on Signal Processing, 2000, 48(7): 2132–2144
13. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6): 716–723
14. Solomonoff R J. A formal theory of inductive inference. Part I. Information and Control, 1964, 7(1): 1–22
15. Kolmogorov A N. Three approaches to the quantitative definition of information. Problems of Information Transmission, 1965, 1(1): 1–11
16. Wallace C S, Boulton D M. An information measure for classification. Computer Journal, 1968, 11(2): 185–194
17. Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2): 461–464
18. Rissanen J. Modeling by shortest data description. Automatica, 1978, 14: 465–471
19. MacKay D J C. Bayesian interpolation. Neural Computation, 1992, 4(3): 415–447
20. McGrory C A, Titterington D M. Variational approximations in Bayesian model selection for finite mixture distributions. Computational Statistics & Data Analysis, 2007, 51(11): 5352–5367