ISSN 1673-3460

Volume 5 · Number 3 · September 2010

Frontiers of Electrical and Electronic Engineering in China

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Front. Electr. Electron. Eng. China 2010, 5(3): 281–328
DOI 10.1007/s11460-010-0108-9

RESEARCH ARTICLE

Lei XU

Bayesian Ying-Yang system, best harmony learning, and five action circling

Abstract Firstly proposed in 1995 and systematically developed in the past decade, Bayesian Ying-Yang learning[1] is a statistical approach for a two pathway featured intelligent system via two complementary Bayesian representations of a joint distribution on the external observation X and its inner representation R, which can be understood from a perspective of the ancient Ying-Yang philosophy. We have $q(X, R) = q(X|R)q(R)$ as Ying that is primary, with its structure designed according to tasks of the system, and $p(X, R) = p(R|X)p(X)$ as Yang that is secondary, with $p(X)$ given by samples of X while the structure of $p(R|X)$ designed from Ying according to a Ying-Yang variety preservation principle, i.e., $p(R|X)$ is designed as a functional with $q(X|R)$, $q(R)$ as its arguments. We call this pair Bayesian Ying-Yang (BYY) system. A Ying-Yang best harmony principle is proposed for learning all the unknowns in the system, in help of an implementation featured by a five action circling under the name of A5 paradigm. Interestingly, it coincides with the famous ancient WuXing theory that provides a general guide to keep the A5 circling well balanced towards a Ying-Yang best harmony. This BYY learning provides not only a general framework that accommodates typical learning approaches from a unified perspective but also a new road that leads to improved model selection criteria, Ying-Yang alternative learning with automatic model selection, as well as coordinated implementation of Ying based model selection and Yang based learning regularization.

This paper aims at an introduction of BYY learning in a twofold purpose. On one hand, we introduce fundamentals of BYY learning, including system design principles of least redundancy versus variety preservation, global learning principles of Ying-Yang harmony versus Ying-Yang matching, and local updating mechanisms of rival penalized competitive learning (RPCL) versus maximum a posteriori (MAP) competitive learning, as well as learning regularization by data smoothing and induced bias cancelation (IBC) priori. Also, we introduce basic implementing techniques, including apex approximation, primal gradient flow, Ying-Yang alternation, and Sheng-Ke-Cheng-Hui law. On the other hand, we provide a tutorial on learning algorithms for a number of typical learning tasks, including Gaussian mixture, factor analysis (FA) with independent Gaussian, binary, and non-Gaussian factors, local FA, temporal FA (TFA), hidden Markov model (HMM), hierarchical BYY, three layer networks, mixture of experts, radial basis functions (RBFs), subspace based functions (SBFs). This tutorial aims at introducing BYY learning algorithms in a comparison with typical algorithms, particularly with a benchmark of the expectation maximization (EM) algorithm for the maximum likelihood. These algorithms are summarized in a unified Ying-Yang alternation procedure with major parts in a same expression while differences simply characterized by few options in some subroutines. Additionally, a new insight is provided on the ancient Chinese philosophy of Yin-Yang and WuXing from a perspective of information science and intelligent system.

Keywords Bayesian Ying-Yang (BYY) system, Yin-Yang philosophy, best harmony, WuXing, A5 paradigm, randomized Hough transform (RHT), rival penalized competitive learning (RPCL), maximum a posteriori (MAP), semi-supervised learning, automatic model selection, Gaussian mixture, factor analysis (FA), binary FA, non-Gaussian FA, local FA, temporal FA, three layer networks, mixture of experts, radial basis function (RBF) networks, subspace based function (SBF), state

Received March 1, 2010; accepted March 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

[1] "Ying" is spelled "Yin" in Chinese Pin Yin. To keep its original harmony with Yang, we deliberately adopted the term "Ying-Yang" since 1995.

space modeling, hidden Markov model (HMM), hierarchical BYY, apex approximation, Ying-Yang alternation

1 Introduction

1.1 Two types of abilities and three inverse problems

An intelligent system, which could be an individual or a collection of natural and artificial intelligent bodies, survives in its world with needs of two types of intelligent abilities. As illustrated in Fig. 1(a), the right path denotes that Type I consists of abilities of knowing “what it is” or modeling regularities or structures among data as its knowledge about the world. The left path in Fig. 1(a) denotes Type II that consists of skills of problem solving, i.e., skills of appropriately responding upon what are currently encountering. The action can be either just perceiving (e.g., identify, recognize, etc.) or also reacting (e.g., reasoning, integrate, decide, etc.). Correspondingly, two types of intelligent abilities are obtained by Type I, Type II of learning, respectively, from pieces of uncertain evidences (or called samples).

In a general sense, the tasks of Type II can be regarded as the inverse problems of Type I. We get insights from two typical examples shown in Fig. 1(b) and Fig. 1(c). The first is a widely used example of getting an optimal classifier $p(j|x)$ from Type I knowledge and then classifying all the samples into clusters by finding the boundaries of dissimilarity [1], while Type I knowledge describes each cluster of samples by a Gaussian $G(x|\mu_j, \Sigma_j)$ with a mean vector μ_j and a covariance matrix Σ_j as well as its proportion α_j. Roughly, Type I describes how a sample is generated from one of the k Gaussian components, while Type II inversely maps each sample to one of the labels that correspond to the k Gaussian components. Shown in Fig. 1(c) is another example, Type I knowledge is that samples of x are generated through a linear system from y of uncorrelated Gaussian factors subject to an additive noise of Gaussian. The problem solving task of Type II is inversely solving y from x by maximizing the posteriori, which is achieved by an inverted linear mapping.

Both the examples are actually special cases of the first one of three levels of inverse problems [2,3]. As shown in Fig. 2(a), the observed samples come from a mapping $y \to x$, subject to certain many-to-one and probabilistic uncertainties. In the framework of probability theory, these uncertainties are described by distributions with $q(x|y)$ for a probabilistic mapping $y \to x$ and $p(y|x)$ for a probabilistic inverse map $x \to y$. The Type II task is estimating $p(y|x)$ from a given $q(x|y)$, usually in help of choosing $q(y)$ for regularizing each y on its chance to be a cause or inner representation. Far beyond the examples of Figs. 1(b) and 1(c), this task is widely encountered in real applications, e.g., geophysics, medical imaging (such as computed axial tomography, electroencephalography/event-related-potentials), remote sensing, ocean acoustic tomography, nondestructive testing, and astronomy [4], as well as controlling systems and signal processing [5,6].

Moreover, this inverse problem is a subtask of the second level of inverse problems. For the example in Fig. 1(b), the task of getting an optimal classifier $p(j|x)$ is a subtask of estimating the parameters $\Theta_j = \{\mu_j, \Sigma_j, \alpha_j\}$ if they are unknown too, which is usually referred as parameter learning [7,8]. For the example in Fig. 1(c), the task of getting $p(y|x)$ is a subtask of estimating the unknown parameters $\Theta = \{A, \mu, \Sigma, \Lambda, v\}$, which is usually referred under the name of factor analysis [9,10]. Generally as shown in Fig. 2(b), provided that $q(x|y, \theta_{x|y})$ and $q(y|\theta_y)$ come from two parametric families with their function structures pre-specified but two sets $\theta_{x|y}, \theta_y$ of unknown parameters, the task is getting

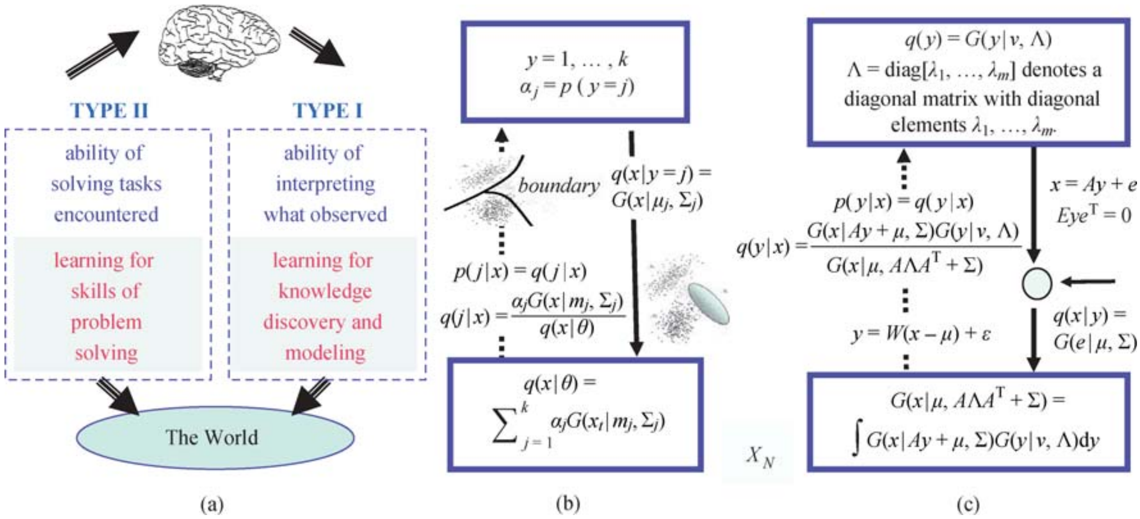

TYPE II
ability of solving tasks encountered
learning for skills of problem solving
TYPE I
ability of interpreting what observed
learning for knowledge discovery and modeling
The World
(a)
$y = 1, \ldots, k$
$\alpha_j = p(y=j)$
boundary
$q(x|y=j) = G(x|\mu_j, \Sigma_j)$
$p(j|x) = q(j|x)$
$q(j|x) = \frac{\alpha_j G(x|m_j, \Sigma_j)}{q(x|\theta)}$
$q(x|\theta) = \sum_{j=1}^{k} \alpha_j G(x_t|m_j, \Sigma_j)$
X_N
(b)
$q(y) = G(y|v, \Lambda)$
$\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_m]$ denotes a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_m$.
$x = Ay + e$
$Ey e^{\mathrm{T}} = 0$
$p(y|x) = q(y|x)$
$q(y|x) = \frac{G(x|Ay+\mu, \Sigma)G(y|v, \Lambda)}{G(x|\mu, A\Lambda A^{\mathrm{T}} + \Sigma)}$
$y = W(x-\mu) + \varepsilon$
$q(x|y) = G(e|\mu, \Sigma)$
$G(x|\mu, A\Lambda A^{\mathrm{T}} + \Sigma) = \int G(x|Ay+\mu, \Sigma)G(y|v, \Lambda)\mathrm{d}y$
(c)

Fig. 1 Two types of intelligent abilities and learning tasks, with two typical examples. (a) Two types of learning; (b) Gaussian mixture; (c) factor analysis

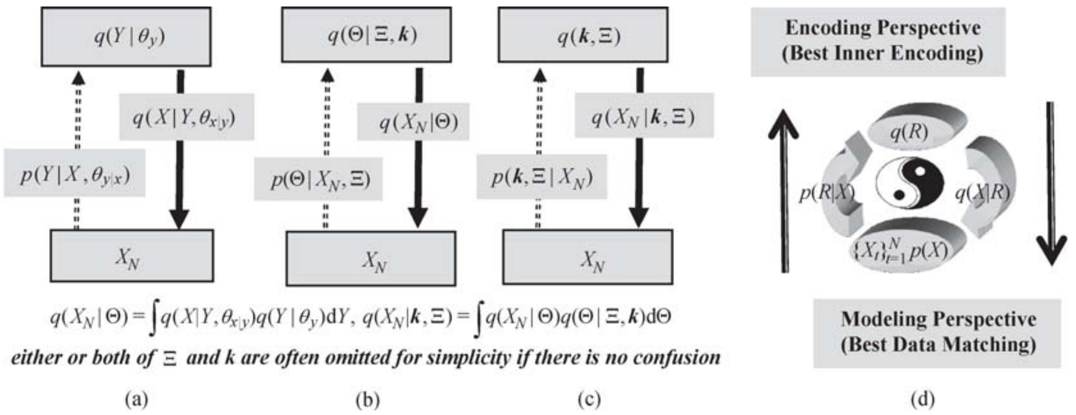

$q(Y|\theta_y)$
$q(X|Y,\theta_{x|y})$
$p(Y|X,\theta_{y|x})$
X_N
$q(\Theta|\Xi,\boldsymbol{k})$
$q(X_N|\Theta)$
$p(\Theta|X_N,\Xi)$
X_N
$q(\boldsymbol{k},\Xi)$
$q(X_N|\boldsymbol{k},\Xi)$
$p(\boldsymbol{k},\Xi|X_N)$
X_N
$q(X_N|\Theta)=\int q(X|Y,\theta_{x|y})q(Y|\theta_y)\mathrm{d}Y,\ q(X_N|\boldsymbol{k},\Xi)=\int q(X_N|\Theta)q(\Theta|\Xi,\boldsymbol{k})\mathrm{d}\Theta$
either or both of Ξ and k are often omitted for simplicity if there is no confusion
Encoding Perspective
(Best Inner Encoding)
$q(R)$
$p(R|X)$
$q(X|R)$
$\{X_t\}_{t=1}^N p(X)$
Modeling Perspective
(Best Data Matching)
(a)
(b)
(c)
(d)

Fig. 2 Three levels of inverse problems and two learning perspectives. (a) Level 1 for problem solving; (b) level 2 for parameter learning; (c) level 3 for model selection; (d) two learning perspectives

an inverse mapping $X_N \to \Theta$ from a set of samples $X_N = \{x_t\}$ under a priori $q(\Theta)$. Moreover, the mapping $X_N \to \Theta$ includes the task of getting $p(y|x)$ or $x \to y$ for every sample of x. That is, a second level of inverse problem is nested with a series of the first level of inverse problems.

Furthermore, we have another level of inverse problems. The task of estimating the parameters Θ is further a subtask of determining an appropriate number k for the example in Fig. 1(b), and a subtask of determining an appropriate dimension [11] of y for the example in Fig. 1(c). In general, we know neither $\theta_{x|y}, \theta_y$ nor the function structures of $q(x|y, \theta_{x|y})$ and $q(y|\theta_y)$. Instead, we consider a family of infinite many structures $\{S_{\boldsymbol{k}}\}$ via individual simple structures in a simple combination, e.g., each Gaussian in Fig. 1(b) and each dimension of y in Fig. 1(c). That is, each $S_{\boldsymbol{k}}$ shares a same configuration S but in a different scale denoted by a scale parameter $\boldsymbol{k}$ that consists of one integer or a set of integers, e.g., the number of Gaussians in Fig. 1(b), and the dimension of y in Fig. 1(c). The configuration S depends on the forms of simple structures and their combination, which is designed according to task-dependent knowledge.

Provided with a given configuration S, as shown in Fig. 2(c), the third level of inverse problem is getting an inverse mapping $X_N \to \boldsymbol{k}, \Xi$. When $q(\Theta|\Xi)$ is pre-specified by a priori knowledge with Ξ being given hyper-parameters, the task becomes simply $X_N \to \boldsymbol{k}$, which is usually referred as model selection [12] since we select among a set of candidate models from enumerating a series values of $\boldsymbol{k}$ and solving $X_N \to \Theta$ for every value of $\boldsymbol{k}$. On the other hand, if $\boldsymbol{k}$ is pre-specified and the structure of $q(\Theta|\Xi)$ is given but with one or more hyper-parameters Ξ unknown, the task becomes simply $X_N \to \Xi$, e.g., determining the strength in learning regularization [13,14]. In general, a third level of inverse problem $X_N \to \boldsymbol{k}, \Xi$ covers both the cases, nested with a series of the second level of inverse problems.

As a whole, all the three levels of inverse problems can be summarized into a unified expression shown in Fig. 2(d), with $y, \Theta, \boldsymbol{k}, \Xi$ all included in R. Readers are referred to Refs. [2,3] for further details on the three levels of inverse problems. These problems are tackled by various efforts, which can be summarized from the following two perspectives.

1.2 Two learning perspectives and Bayesian Ying-Yang (BYY) learning

One is named as a principle of best inner encoding, as shown in Fig. 2(d). From the bottom-up perspective, we get $p(R|X)$ for a Type II task of encoding X_N into the best inner representations. Typical studies may be summarized into two streams. One stream solves a bottom-up pathway as the inverse of a given top-down pathway. The other stream focuses on merely a bottom-up pathway that maps X_N into inner representations that become least redundant. One typical instance of the latter is the minimum mutual information (MMI) [15] among components of inner encoding, e.g., the elements of y in Fig. 1(c) become mutually independent. Specifically, its one limit case is that the maximum information (INFOR-MAX) [16] is transferred to the inner representation. Both MMI and INFOR-MAX have been widely adopted in the studies on independent component analysis (ICA), especially those via a linear mapping $y = Wx$ [17,18]. Though the second stream do not directly involve a top-down pathway, one can still conceptually regard that X_N are generated from inner representations of least redundant and thus a bottom-up pathway is sought as its inverse.

The other principle is best data matching [17,19] that is implemented from a top-down perspective shown in Fig. 2(d), which seeks a best interpretation of data X_N by Type I knowledge or a best matching of X_N by $q(X_N) = \int q(X_N|R)q(R)\mathrm{d}R$. One widely studied example is the principle of maximizing the probability or density defined by a given model $q(X_N)$. One type of instances includes the maximum likelihood (ML) learning

$\max_\Theta \ln q(X|\Theta)$ with $q(X|\Theta) = \int q(X|Y)q(Y)\mathrm{d}Y$ in help with an expectation maximization (EM) iteration [7,20–22], during which an induced bottom-up inverse $p(Y|X)$ is given by the Bayesian posteriori of $q(X|Y)$ and $q(Y)$, that is, the task of Fig. 2(a) is included as a part of job here. Also, Y could be in a much advanced representation, e.g., a vocabulary of shapes and patterns in those studies under the name of pattern theory [23], the forward-backward projections in the brain [24,25], and analysis by synthesis [26]. Moreover, efforts in recent decades have been made on the general form $q(X_N) = \int q(X_N|R)q(R)\mathrm{d}R$ under the name of the marginal likelihood or marginal Bayesian approach [20–22], including Bayesian inference criterion (BIC) [27] and minimum description length (MDL) [28,29], as well as Bayesian studies [30–32] and minimum message length (MML) [33,34].

Beyond that a bottom-up pathway is given by the Bayesian posteriori as above, efforts have also been made on estimating both the two pathways, mainly in three streams. One consists of those studies motivated from various aspects, e.g., best reconstruction based data dimension reduction [35,36] and least mean squared error reconstruction (LMSER) self-organization [19], as well as cognitive science motivated efforts such as adaptive resonance theory [37,38] and forward-inverse models for motor control [39–41]. The other two streams are within the probabilistic theoretic framework shown in Fig. 2. One is referred under the Helmholtz machine, variational Bayes, and variational approaches [42–46], targeting at an approximate implementation of Bayesian inference. To avoid tackling an intractable sum or integral for $q(X) = \int q(X|Y)q(Y)\mathrm{d}Y$, instead of exactly getting a Bayesian inverse $p(Y|X) = q(X|Y)q(Y)/q(X)$, efforts aim at estimating unknown parameters of $p(Y|X)$ in an easy computing structure such that unknowns in $q(X|Y)q(Y)$ are estimated in an approximate maximization of $q(X) = \int q(X|Y)q(Y)\mathrm{d}Y$, e.g., the Helmholtz machine is featured by a typical example of the problem shown in Fig. 2(a) with $q(x|y, \theta_{x|y})$ for a multi-layer networks and $p(y|x, \theta_{y|x})$ for a conditional independent product also by a multi-layer networks [42,43]. Moreover, further extensions proceeds to handle the problem type shown in Fig. 2(b) under the name of variational Bayes for model selection [46], by replacing Laplace approximation with variational approximation.

Firstly proposed in 1995 [47] and systematically developed over a decade [3,48,49], efforts of the third stream have been made under the name of Bayesian Ying-Yang harmony learning. From a modern science perspective that regards the famous ancient Yin-Yang philosophy as a meta theory of system sciences and intelligent systems (details are referred to Appendix B), we consider jointly a mapping $R \to X$ and an inverse $X \to R$ as shown in Fig. 2(d) via the joint distribution of X, R in two types of Bayesian decompositions:

$$\begin{aligned} \text{Ying} &: q\left(X, R\right) = q\left(X|R\right)q(R), \\ \text{Yang} &: p\left(X, R\right) = p\left(R|X\right)p(X). \end{aligned} \tag{1}$$

The decomposition of $p(X, R)$ coincides the Yang concept with a visible domain $p(X)$ for a Yang space and a forward pathway by $p(R|X)$ as a Yang pathway. Thus, $p(X, R)$ is called Yang machine. Similarly, $q(X, R)$ is called Ying machine with an invisible domain $q(R)$ for a Ying space and a backward pathway by $q(X|R)$ as a Ying pathway. Such a Ying-Yang pair is called Bayesian Ying-Yang (BYY) system. Based on a Ying-Yang variety preservation principle, Ying is primary and its structure is designed according to tasks of the system, while Yang is secondary with $p(X)$ given by samples of X as inputs to the system while $p(R|X)$ designed as a functional that varies with Ying $q(X|R)q(R)$.

All the unknowns in the system is learned by following a principle of best harmony between Yang and Ying machines, in a sense that Ying and Yang match each other in a most compact way. This principle is mathematically implemented by maximizing the following harmony functional:

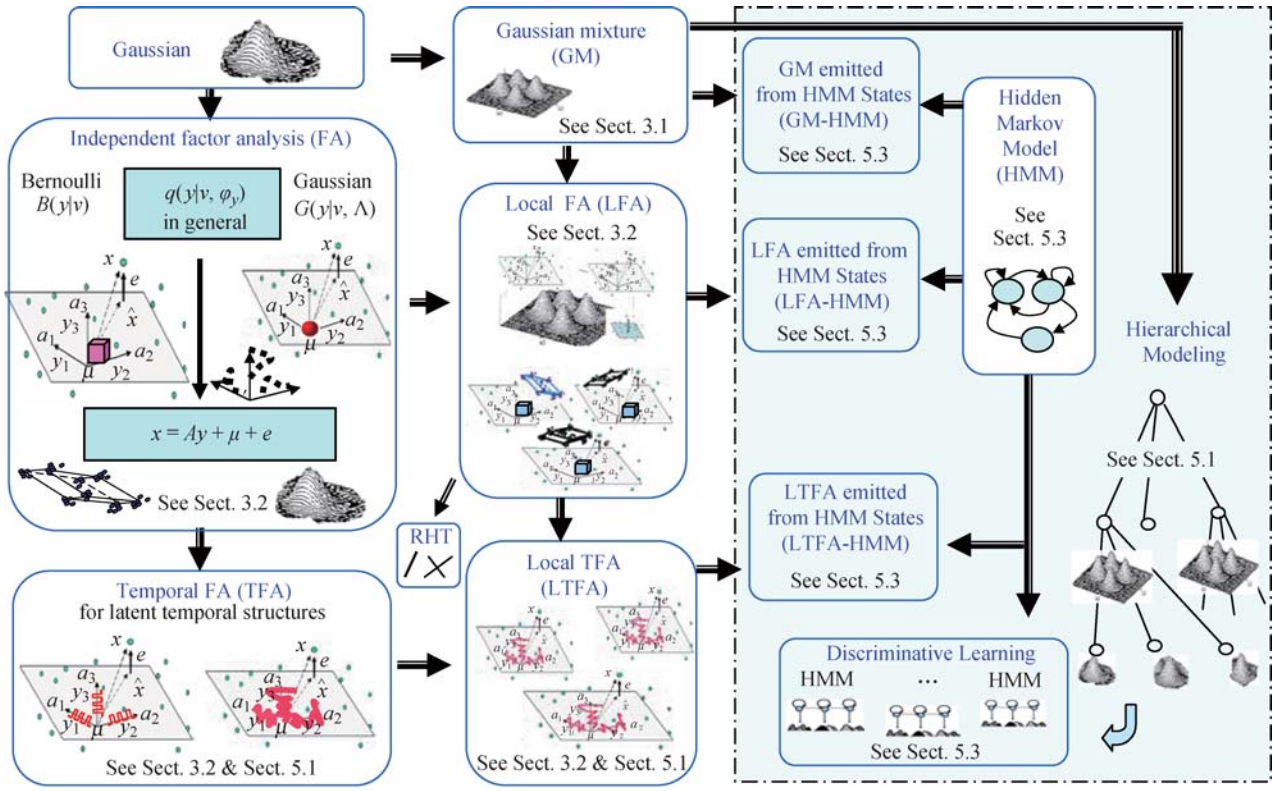
$$H\left(p\| q\right) = \int p\left(R|\, X\right)p\left(X\right)\ln\left[q\left(\,X|\, R\right)q\left(R\right)\right]\mathrm{d}R\mathrm{d}X, \tag{2}$$

such that three levels of inverse problems, especially parameter learning and model selection, are tackled jointly. The general framework not only accommodates typical learning theories and approaches under a unified perspective but also provides a new road that leads to improved model selection criteria, Ying-Yang alternating learning algorithms with automatic model selection, and a coordinated implementation of Ying based model selection and Yang based learning regularization.

1.3 Scope of this paper

This paper aims at an introduction of BYY learning in a twofold purpose. On one hand, we introduce the fundamentals of BYY learning and basic implementing techniques, and provide a roadmap to illustrate the relations to other typical learning approaches. On the other hand, we provide a tutorial on learning algorithms for a number of typical learning tasks, as illustrated in Fig. 3.

Section 2 begins with the well known small sample size challenge of statistical learning, and further introduces typical efforts towards this challenge, especially a trend from two-stage based model selection to automatic model selection and further to a coordinated implementation of making automatic model selection based on Ying machine and learning regularization based on Yang machine [2,3,50,51]. Moreover, taking random Hough


Gaussian
Gaussian mixture (GM)
See Sect. 3.1
Independent factor analysis (FA)
Bernoulli $B(y|v)$
$q(y|v, \varphi_y)$ in general
Gaussian $G(y|v, \Lambda)$
$x = Ay + \mu + e$
See Sect. 3.2
Local FA (LFA)
See Sect. 3.2
RHT
Temporal FA (TFA) for latent temporal structures
See Sect. 3.2 & Sect. 5.1
Local TFA (LTFA)
See Sect. 3.2 & Sect. 5.1
GM emitted from HMM States (GM-HMM)
See Sect. 5.3
LFA emitted from HMM States (LFA-HMM)
See Sect. 5.3
Hidden Markov Model (HMM)
See Sect. 5.3
Hierarchical Modeling
See Sect. 5.1
LTFA emitted from HMM States (LTFA-HMM)
See Sect. 5.3
Discriminative Learning
HMM ··· HMM
See Sect. 5.3

all the above studies can be extended to cover semi-supervised and supervised learning with some light modifications

Fig. 3 A roadmap of several typical learning tasks

transform (RHT) [52] as an example, a general problem solving paradigm A5 is introduced and shown to be applicable to the two tasks in Figs. 1(b) and 1(c) as well as the general tasks for three levels of inverse problems in Fig. 2. Even interestingly, it coincides with the famous ancient Chinese WuXing theory [53]. This A5 paradigm and WuXing theory will provide us a guide to improve existing algorithms and develop new algorithms.

Section 3 starts to introduce BYY harmony learning algorithms on Gaussian mixture (GM) and factor analysis (FA) with independent Gaussian, binary, and non-Gaussian factors in Fig. 2 as well as local factor analysis (LFA) that is a combined generalization of GM and FA, and these algorithms are summarized in a unified Ying-Yang alternation procedure with major parts in a same expression while differences simply characterized by few options in a subroutine, in a benchmark of the standard EM algorithm.

Section 4 introduces the fundamentals of BYY learning, including not only learning principles of Ying-Yang harmony versus Ying-Yang matching, several favorable features, and relations to rival penalized competitive learning (RPCL), but also system design principles of least redundancy and variety preservation, as well as basic implementing techniques, including apex approximation, primal gradient flow, and Ying-Yang alternation. Moreover, it is further introduced that these studies actually provide a general framework that algorithms obtained for unsupervised learning can be directly used for semi-supervised and supervised learning, e.g., learning algorithms for Gaussian mixture, factor analysis (FA) with independent Gaussian, binary, and non-Gaussian factors, and local FA can be easily used for learning three layer networks, mixture of experts, radial basis functions (RBFs), subspace based functions (SBFs), etc. In addition, we provide a roadmap in Appendix A to illustrate a systematic relation to several existing typical learning approaches.

Section 5 proceeds to hierarchical and temporal BYY harmony learning, including a hierarchical mixture of Gaussians, temporal factor analysis, hidden Markov models (HMMs), and discriminative learning of multiple HMM models with each state emitting a mixture of Gaussians or LFAs.

Finally, Sect. 6 provides a comprehensive summary and a number of future topics.

2 Towards small sample size challenge

2.1 Learning regularization, sparse learning, and model selection

As addressed in Sect. 1.1, parameter learning for determining Θ and model selection for choosing $\boldsymbol{k}$ are two major tasks for handling three types of inverse problems.

The principle of best data matching in Sect. 1.2 provides a guide to determine $\Theta_{\boldsymbol{k}}$ based on X_N of training samples. One most popular instance is $\max_{\Theta} \ln q(X_N|\Theta)$ or minimizing a fitting or empirical error $-\ln q(X_N|\Theta)$. For an illustration, shown in Fig. 4(a) is a case that $\boldsymbol{k}$ consists of one integer (otherwise, it is a multi-dimensional plot with all the integers in $\boldsymbol{k}$ enumerated in certain order). For a sample size N that is not large enough, this fitting error decreases monotonically as k grows up, until the error reaches zero at $k = k_N$, a value that relates to N but is usually much bigger than an appropriate k^*. It means this learning is under-selective on k such that extra resource of structures is wasted. Even worse, the wasted resource has actually been used to learn noises or outliers as if some regularity underlying X_N, which deteriorates the generalization performance. This is usually called over-fitting problem.

Shown in Fig. 4(b) is an insight about the reason of this problem. $\max_{\Theta} \ln q(X_N|\Theta)$ only takes in consideration the relation $\Theta_{\boldsymbol{k}} \rightarrow X_N$ while no consideration is made on the upper layer relation $\boldsymbol{k} \rightarrow \Theta$ to provide an appropriate constraint on determining $\boldsymbol{k}$. A direction to tackle this problem is adding more constraints to $\max_{\Theta} \ln q(X_N|\Theta)$ by a correcting term $\Delta(X_{N,}\boldsymbol{k})$. One way to get this term is directly adding certain constraint on $\boldsymbol{k} \rightarrow X_N$, e.g., Vapnik-Chervonenkis (VC) dimension based generalization error bound [54], cross validation (CV) based criteria [55,56], Akaike information criterion (AIC) and extensions [57–59]. The other way provides a constraint on $\boldsymbol{k} \rightarrow \Theta$ through a priori $q(\Theta|\boldsymbol{k})$ as shown in Figs. 4(c) and 4(d). There are two choices for using this $q(\Theta|\boldsymbol{k})$. One is maximizing $\ln[q(X_N|\Theta)q(\Theta|\boldsymbol{k})]$ [1], as shown in Fig. 4(c), which has been widely studied under the name of the maximum a posteriori (MAP) estimate and the name of the two part shortest coding length (e.g., MML [33,34] and the early MDL [28]). It directly affects the estimation of Θ and thus the performance is sensitive to whether an appropriate priori $q(\Theta|\boldsymbol{k})$ is available. Also it indirectly affects the selection of $\boldsymbol{k}$, highly depending on estimating Θ. The other is $\max_{\boldsymbol{k}} \ln q(X_N|\boldsymbol{k})$ as shown in Fig. 4(d), which considers a direct relation $\boldsymbol{k} \rightarrow X_N$, such as BIC [27], the normalized maximum likelihood (NML) based MDL [29], and variational Bayes [42–46]. Computationally, handling the integral over Θ involves certain approximation.

Considering a model with a big enough scale $\boldsymbol{k}$ to accommodate X_N, the effect of $\max_{\Theta} \ln[q(X_N|\Theta)q(\Theta|\boldsymbol{k})]$ can be further classified into two types, depending on choices of $q(\Theta|\boldsymbol{k})$. One is called learning regularization [13,14] that imposes certain isotropic constraint on either or both of Θ and strcture in order to effectively reduce model complexity, without necessarily discarding extra parameters. For an example, if we use a polynomial of a degree $k > 2$ to fit a set of samples from $y = x^2 + 3x + 2$, a model selection purpose desires to force all the terms $a_i x^i$, $i > 2$, to be zero, but minimizing $a_1^2 + a_2^2 + \cdots + a_k^2$ fails to treat the parameters a_i, $i > 2$, differently from a_0, a_1, and a_2. Such a regularization actually disturbs the purpose of model selection. That is, there is a trade-off between making model selection and learning regularization. The other is called sparse learning or Lasso shrinkage that prunes away extra weights by a Laplace prior $q(\Theta)$, which is usually workable for a regression or interpolation task [60–62]. To be further addressed at the end of Sect. 4.2, a generic priori $q(\Theta)$ comes from not only a priori $q(\Theta|\Xi)$ for adding certain information about Θ but also one $q(\Theta)$ that actually intends to cancel out certain bias implicitly resulted from using a model $q(X_N|\Theta)$ on a small size of samples.

A further task is to getting an appropriate $\boldsymbol{k}^*$ by tackling the inverse problem in Fig. 2(c), which involves a two stage implementation of model selection, i.e., the best is selected by a criterion $J(\boldsymbol{k}|\Theta^*)$ together with a series of parameter learning to get Θ^* on a set of candidate models obtained by enumerating $\boldsymbol{k}$. Unfortunately, not only a two stage implementation is very expensive

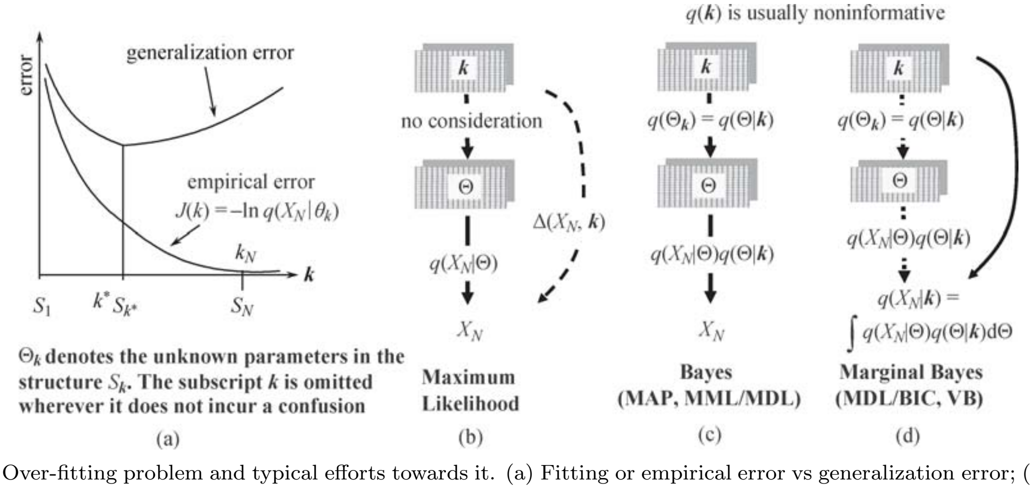


Fig. 4 Over-fitting problem and typical efforts towards it. (a) Fitting or empirical error vs generalization error; (b) adding some constraints to $\max_{\Theta} \ln q(X_N|\Theta)$ by a correcting term $\Delta(X_N, \boldsymbol{k})$; (c) Bayes; (d) Marginal Bayes

to compute, but also performances of getting Θ^* deteriorates considerably for those candidate models with large $\boldsymbol{k}$, especially when N is small and $\boldsymbol{k}$ consists of more than one integers.

One road to reduce computing cost is featured by stepwise implementation. Typical examples are those incremental algorithms that attempt to incorporate as much as possible what already learned as $\boldsymbol{k}$ increases step by step, focusing on only learning newly added parameters, e.g., the studies made on mixture of factor analysis [63]. However, it usually leads to a suboptimal performance because not only those newly added parameters but also the old parameter set actually have to be re-learned. Reversely, this suboptimal problem may be lessened in a way that $\boldsymbol{k}$ starts at a large value and decrease step by step. Taking out of Θ one or a subset of parameters, we discard this subset if the criterion indicates an improvement after updating the rest parameters or a biggest improvement by trying a number of such subsets of parameters. Precisely, this procedure is formulated as a tree searching. The initial parameter set Θ is the root of the tree. Discarding one subset leads to one immediate descendent. A depth-first searching suffers from a suboptimal performance seriously, while a breadth-first searching suffers a huge combinatorial computing cost. Usually, a trade off between the two extremes is considered.

The other road of studies is referred as automatic model selection. An early effort is RPCL [64–66] for the GM task shown in Fig. 1(b), with the number k automatically determined during learning. The key idea is that not only the winner μ_w moves a little bit to adapt the current sample x_t but also the rival (i.e., the second winner) μ_r is repelled a little bit from x_t to reduce a duplicated information allocation. As a result, an extra μ_j is driven far away from data with its corresponding $\alpha_j \to 0$ and $\mathrm{Tr}[\Sigma_j] \to 0$. In general, RPCL is applicable to any model that consists of k individual substructures, with extra substructures discarded by the rival penalized mechanism and thus model selection made automatically.

Being a quite difference nature from a usual stepwise implementation that adds or removes a subset of parameters from Θ based on whether a selection criterion indicates an improvement, automatic model selection is associated with a learning algorithm or a learning principle with the following two features.

First, there is an indicator $\psi(\theta_{\mathrm{SR}})$ on a subset θ_{SR} of scale representative (SR) parameters. Such a subset θ_{SR} actually represents a particular structural component that is effectively discarded if its corresponding $\psi(\theta_{\mathrm{SR}}) = 0$. For the GM problem in Fig. 1(b), each α_l or a set of parameters in Σ_l represents a Gaussian component, and we have either or both of $\psi(\alpha_l) = \alpha_l$ and $\psi(\Sigma_l) = \mathrm{Tr}[\Sigma_l]$. A Gaussian component is discarded if either or both of

$$\alpha_l = 0 \quad \text{and} \quad \mathrm{Tr}[\Sigma_l] = 0. \tag{3}$$

As a result, k effectively reduces to $k-1$. For the FA problem in Fig. 1(c), each eigenvalue of Λ_l represents that one dimension $y^{(l)}$ or equivalently one column of the matrix A becomes extra and thus discarded if its corresponding $\psi(\lambda_l) = \lambda_l$ gets

$$\lambda_l = 0.$$

Second, in implementation of this algorithm or this learning principle, there is an intrinsic mechanism that drives

$$\psi(\theta_{\mathrm{SR}}) \to 0, \quad \text{as } \theta_{\mathrm{SR}} \text{ tends to a specific value}, \tag{4}$$

if the corresponding structure is redundant and thus can be effectively discarded, e.g., RPCL learning drives an extra μ_j with $1/||\mu_j|| \to 0$, and its corresponding $\alpha_j \to 0, \mathrm{Tr}[\Sigma_j] \to 0$.

On a small size N of samples X_N, the maximum likelihood is not good on model selection. Though $\max_\Theta \ln q(X_N|\Theta)$ sometimes also leads to Eq. (4) for some redundant structural component, it fails to do so normally and good enough. With a priori $q(\Theta|\boldsymbol{k})$ that selectively prefers the desired value, those typical efforts on Fig. 4 usually help to enhance the tendency of Eq. (4) on redundant structural components. The above mentioned sparse learning can be regarded a special example that prunes away extra weights by using a Laplace prior for a regression or interpolation task [60–62]. The other example is pruning extra α_l in the GM problem shown in Fig. 1(b) by a Dirichlet prior [67,68]. However, those efforts highly depends on choosing an appropriate prior, which is usually a difficult task, while an inappropriate $q(\Theta)$ may deteriorate the performance of model selection seriously.

2.2 Model selection and learning regularization from a BYY learning perspective

For a task that includes the first level of inverse problem as shown in Fig. 2, the inner representation consists of not only Θ and $\boldsymbol{k}$ but also Y as the corresponding encoding of each observation. As shown in Fig. 5(a), there is a short term memory (STM) for accommodating Y with its basic representation in either or both of a number of labels and a number of real or binary vectors. The scale set $\boldsymbol{k}_Y$ of the STM domain is featured by the number of labels and the dimension of these vectors. This $\boldsymbol{k}_Y$ is a primary part of the entire scale set $\boldsymbol{k}$ in an intelligent system.

For many typical learning problems, the task of model selection is actually only the selection of this $\boldsymbol{k}_Y$ [69]. For Gaussian mixture in Fig. 1(b), the task is determining $\boldsymbol{k}_Y$ that consists of the number k of labels. For FA in Fig.

1(c), the task is determining $\boldsymbol{k}_Y$ that consists of the dimension m of y. Usually, the structure of $q\left(Y|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ is well specified by the nature of learning tasks, which provides a relation $\boldsymbol{k}_Y \rightarrow Y$.

In spite of its unique nature, those efforts in Fig. 4 handle the problems via getting $q(X_N|\Theta)$ such that the relation $\boldsymbol{k}_Y \rightarrow Y$ becomes hidden behind the marginal integral over Y. Then, the model selection task on $\boldsymbol{k}_Y$ becomes merged within the one on $\boldsymbol{k}$, which fails to observe the relation $\boldsymbol{k}_Y \rightarrow Y$, i.e., the problem is tackled in a same way as introduced in Figs. 4(b), 4(c), and 4(d). However, the relation $\boldsymbol{k}_Y \rightarrow Y$ could be in a better use for helping model selection on $\boldsymbol{k}_Y$ since this relation is actually in a position of equal importance to the relation $\boldsymbol{k} \rightarrow \Theta_{\boldsymbol{k}}$ as discussed in Figs. 4(c) and 4(d).

In contrast, BYY harmony learning gets an appropriate use of the relation $\boldsymbol{k}_Y \rightarrow Y$ via $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ for model selection on $\boldsymbol{k}_Y$. As shown in Fig. 5(b), it follows from Eq. (2) in help with mathematical derivation that $H(p||q)$ becomes $H(p||q, \boldsymbol{k}, \Xi)$ approximately, for which the details are delayed to Sect. 4. Here, we just provide an introduction on its major features. Similar to those typical efforts in Fig. 4, $H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$ provides a generic consideration for determining $\boldsymbol{k}, \Xi$ in help of an informative priori $q(\Theta^b|\boldsymbol{k}, \Xi)$ with hyperparameter Ξ via H_b and a noninformative priori or bias cancellation intended $q(\Theta^a|\boldsymbol{k})$. Moreover, it explicitly considers the relation $\boldsymbol{k}_Y \rightarrow Y$ for model selection on $\boldsymbol{k}_Y$ via $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ that takes a position of equal importance as $q(\Theta^a|\boldsymbol{k})$ within $L(X, R)$, which brings the following favorable natures to the BYY harmony learning.

Improved model selection criteria For the nature of learning tasks, the structure of $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ is usually well specified. Thus, its role on specifying the scale set $\boldsymbol{k}_Y$ is more reliable, unlike that the role of using a priori $q(\Theta^a|\boldsymbol{k})q(\Theta^b|\boldsymbol{k}, \Xi)$ for determining $\boldsymbol{k}$ may become unreliable due to a bad pre-knowledge. Therefore, it provides a model selection criterion $J(\boldsymbol{k})$ in Fig. 5(b) that improves those typical efforts in Fig. 4, especially on the selection of $\boldsymbol{k}_Y$, which has also been shown empirically [54,69,70]. Promisingly, the model selection problems of many typical learning tasks can be reformulated into selecting merely the $\boldsymbol{k}_Y$ part in a BYY system [69]. Furthermore, this criterion $J(\boldsymbol{k})$ can also improve a general selection of $\boldsymbol{k}$ by alternatively making $\Xi^* = \arg\max_\Xi H(p||q, \Xi)$ and $\{\Theta^*, h^*\} = \arg\max_{\Theta,h} H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$.

Improved automatic model selection The scale $\boldsymbol{k}_Y$ is featured by SR parameters in $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$. Each SR parameter actually represents a particular structural component. For the GM in Fig. 1(b), each α_l is such an SR parameter for representing a Gaussian component that is effectively discarded if $\alpha_l = 0$. For the FA in Fig. 1(c), each eigenvalue λ_l of Λ_l represents one dimension $y^{(l)}$ or equivalently one column of A that is discarded if $\lambda_l = 0$. These SR parameters are pushed to zeros by two types of forces. One comes from a priori $q(\Theta^a|\boldsymbol{k})q(\Theta^b|\boldsymbol{k}, \Xi)$, in a way similar to those approaches discussed in Fig. 4. Even much more importantly, the other comes from the role of $q\left(Y|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ in an equal importance to $q(\Theta^a|\boldsymbol{k})$ within $L(X, R)$. Specifically, $\max_{\Theta,h} H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$ includes maximizing

$$\int p(Y) \ln q\left(Y\,|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\,\right) \mathrm{d}Y,$$

$$\text{subject to } \; p(Y) = \int p\left(Y\,|X\,\right) p\left(X\,|X_N, h\right) \mathrm{d}X, \quad (5)$$

with respect to both $P(Y|X)$ and $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$, which will exert a force that pushes those extra SR parameters → 0, i.e., the corresponding structural components are discarded. This mechanism is illustrated by $J_{\rm o}(\boldsymbol{k})$ in

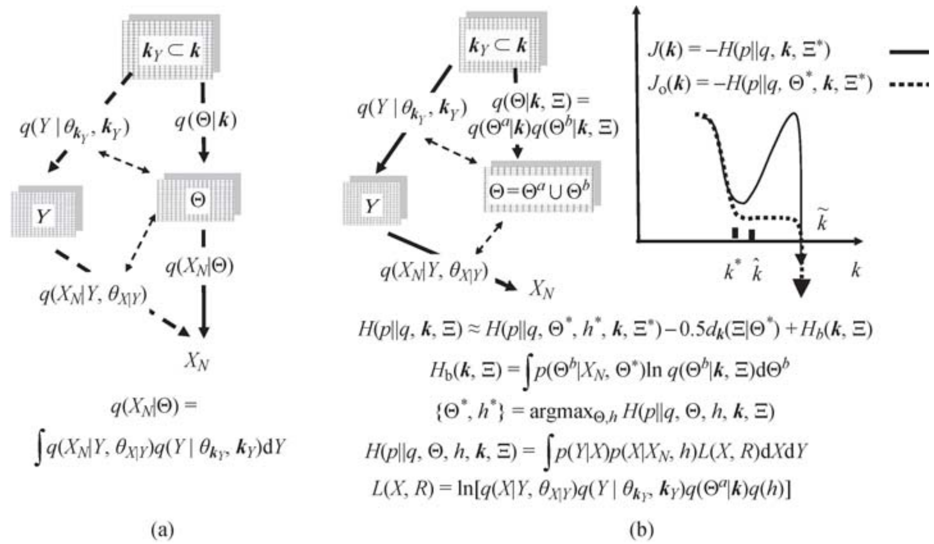

$k_Y \subset k$
$q(Y\,|\,\theta_{k_Y}, k_Y)$
$q(\Theta|k)$
Y
Θ
$q(X_N|\Theta)$
$q(X_N|Y, \theta_{X|Y})$
X_N
$q(X_N|\Theta) =$
$\int q(X_N|Y, \theta_{X|Y})q(Y\,|\,\theta_{k_Y}, k_Y)\mathrm{d}Y$
(a)
$k_Y \subset k$
$q(Y\,|\,\theta_{k_Y}, k_Y)$
$q(\Theta|k, \Xi) =$
$q(\Theta^a|k)q(\Theta^b|k, \Xi)$
Y
$\Theta = \Theta^a \cup \Theta^b$
$q(X_N|Y, \theta_{X|Y})$
X_N
$J(k) = -H(p||q, k, \Xi^*)$
$J_o(k) = -H(p||q, \Theta^*, k, \Xi^*)$
$\tilde{k}$
k^*
$\hat{k}$
k
$H(p||q, k, \Xi) \approx H(p||q, \Theta^*, h^*, k, \Xi^*) - 0.5d_k(\Xi|\Theta^*) + H_b(k, \Xi)$
$H_b(k, \Xi) = \int p(\Theta^b|X_N, \Theta^*)\ln q(\Theta^b|k, \Xi)\mathrm{d}\Theta^b$
$\{\Theta^*, h^*\} = \mathrm{argmax}_{\Theta,h} H(p||q, \Theta, h, k, \Xi)$
$H(p||q, \Theta, h, k, \Xi) = \int p(Y|X)p(X|X_N, h)L(X, R)\mathrm{d}X\mathrm{d}Y$
$L(X, R) = \ln[q(X|Y, \theta_{X|Y})q(Y\,|\,\theta_{k_Y}, k_Y)q(\Theta^a|k)q(h)]$
(b)

Fig. 5 Scale set $\boldsymbol{k}_Y$ of STM domain. (a) Relation $\boldsymbol{k}_Y \rightarrow Y$ is hidden behind the integral over Y; (b) BYY harmony learning appropriately uses relation $\boldsymbol{k}_Y \rightarrow Y$ via $q\left(Y\,\middle|\Theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\,\right)$ for selection on $\boldsymbol{k}_Y$

Fig. 5(b). For the GM in Fig. 1(b), $\alpha_l \to 0$ means its contribution to $J_{\rm o}(\boldsymbol{k})$ is 0, and a number of such parameters becoming 0 result in that $J_{\rm o}(\boldsymbol{k})$ has effectively no change on a range $[\hat{k}, \tilde{k}]$. For the FA in Fig. 1(c), $\lambda_l \to 0$ contributes to $J_{\rm o}(\boldsymbol{k})$ by $-\infty$ as shown beyond $\tilde{k}$. As long as k is initialized big enough, $\hat{k}$ can be found as an estimated upper bound of k^*. That is, an automatic model selection is incurred even without a prior $q(\Theta^a|\boldsymbol{k})q(\Theta^b|\boldsymbol{k}, \Xi)$.

At the first glance, maximizing $\int p(Y) \ln q(Y|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y){\rm d}Y$ seems also encountered in a number of typical learning approaches with an EM type two pathway implementation, such as the EM algorithm implemented maximum likelihood learning [21], information geometry based EM algorithm [22], Helmholtz Machine [42,43]. The difference is that there the maximization is made with $P(Y|X)$ fixed at what obtained by the M-step. In fact, these approaches become equivalent to $\max_\Theta q(X|\Theta) = \int q(X|Y, \theta_{X|Y}) q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right) {\rm d}Y$, where $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ has been buried behind the integral. This point can be observed clearly from the bits-back interpretation of Helmholtz free energy [71]. It is this bits-back that makes $\ln q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ fail to be taken in consideration. The details are referred to Sect. II(D) in Ref. [49] and the last section in Ref. [3].

Ying based model selection versus Yang based learning regularization The separated consideration of $\boldsymbol{k}_Y$ from the rest of $\boldsymbol{k}$ also provides a general framework that integrates the roles of regularization and model selection. Specifically, model selection is made via $q\left(Y\middle|\theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y\right)$ in Ying machine, while regularization is imposed in Yang machine via either or both of two choices. One is designing the structure of $P(Y|X)$ structure, with details delayed to Sect. 4.2. The other is data smoothing regularization. Instead of directly inputting a sample x_t or equivalently its Dirac delta density $\delta(x - x_t)$, x_t is smoothed by a Gaussian $G(x|x_t, h^2 I)$ as input. That is, we let $p(X)$ in Eq. (2) given by

$$p(X|X_N, h) = \prod_{t=1}^{N} G\left(x\left|x_t, h^2 I\right.\right) \quad \text{or}$$

$$p_h(x) = \frac{1}{N}\sum_{t=1}^{N} G\left(x\left|x_t, h^2 I\right.\right), \tag{6}$$

where h is an unknown strength to control the regularization [72–75], which is equivalent to adding a white Gaussian noise to samples with a variance h^2. Moreover, it is known that training with noise is equivalent to typical learning regularization [76]. Progressing beyond [76], not only Eq. (6) is used as an input to a BYY system, but also the difficulty of controlling this strength h, usually encountered by typical regularization approaches, has been avoided with an appropriate h via $\max_{\Theta, h} H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$.

2.3 Five basic actions and circular implementation

According to a general problem solving paradigm A5 (see Sect. 4 of Refs. [72,77]), three levels of inverse problems are implemented via a circular flow featured by five basic mechanisms or actions. We get the insights by taking Hough transform (HT) [78,79] as a starting example, as shown in Fig. 6(a), which was also the original source that motivated this A5 paradigm. The circle starts at the first action A-1 by getting data (e.g., HT picks one pixel from the image), based on which the next action A-2 leads to one or more assumptions on the inner representation. For example, HT quantizes a window of the parameter space $\theta = \{a, b\}$ of a line $y = ax + b$ into a lattice. One pixel is picked to be mapped into a line in the parameter space θ, which activates all the cells on a line as candidate assumptions. The third action A-3 allocates and updates the evidences that support the candidates under consideration, e.g., each cell of a HT window has an accumulator, and one score is added to those accumulators activated by the sample picked. Next, the action A-4 decides one or more best candidates according to the evidences accumulated (e.g., HT seeks the apex scores). Finally, the action A-5 assesses the decided ones via testing samples and then makes a final affirmation, e.g., HT tests whether a line is detected by checking the pixels fallen within a linear band illustrated at the bottom corner of Fig. 6(a).

The evolution of A5 actions is implemented sequentially with each action governed by its local rule. By identifying the feature of each action in the A5 circle, we may get a guide to improve an existing approach. In fact, it was the A5 circle analysis on Fig. 6(a) that motivated the RHT [77,78]. Instead of picking one pixel and activating candidate assumptions on a line by HT, RHT changes it's A-1 and A-2 to pick two pixels that jointly activates merely one candidate assumption, which further makes A-3 no longer limited to the windowed lattice of accumulators but available to other choices. As a result, performances have been improved significantly and computing costs have been reduced greatly. Moreover, there are many other chances to improve one or more of the five rules, which leads to other modifications too, with details referred to a recent review in Ref. [77].

Interestingly, it coincides with the famous ancient WuXing theory [53], a foundation of traditional Chinese medicine (TCM), for which readers are referred to Appendix B. It follows that the problem solving paradigm A5 in Fig. 6(a) can be regarded as a special case of the WuXing theory or a modern interpretation of this meta theory from an intelligent system perspective. One most important part is the famous Sheng-Ke-Cheng-Hui law for keeping five actions well balanced. Sheng guides how to specify the circling order. Interestingly, the order of five action circling in the A5 paradigm [77]

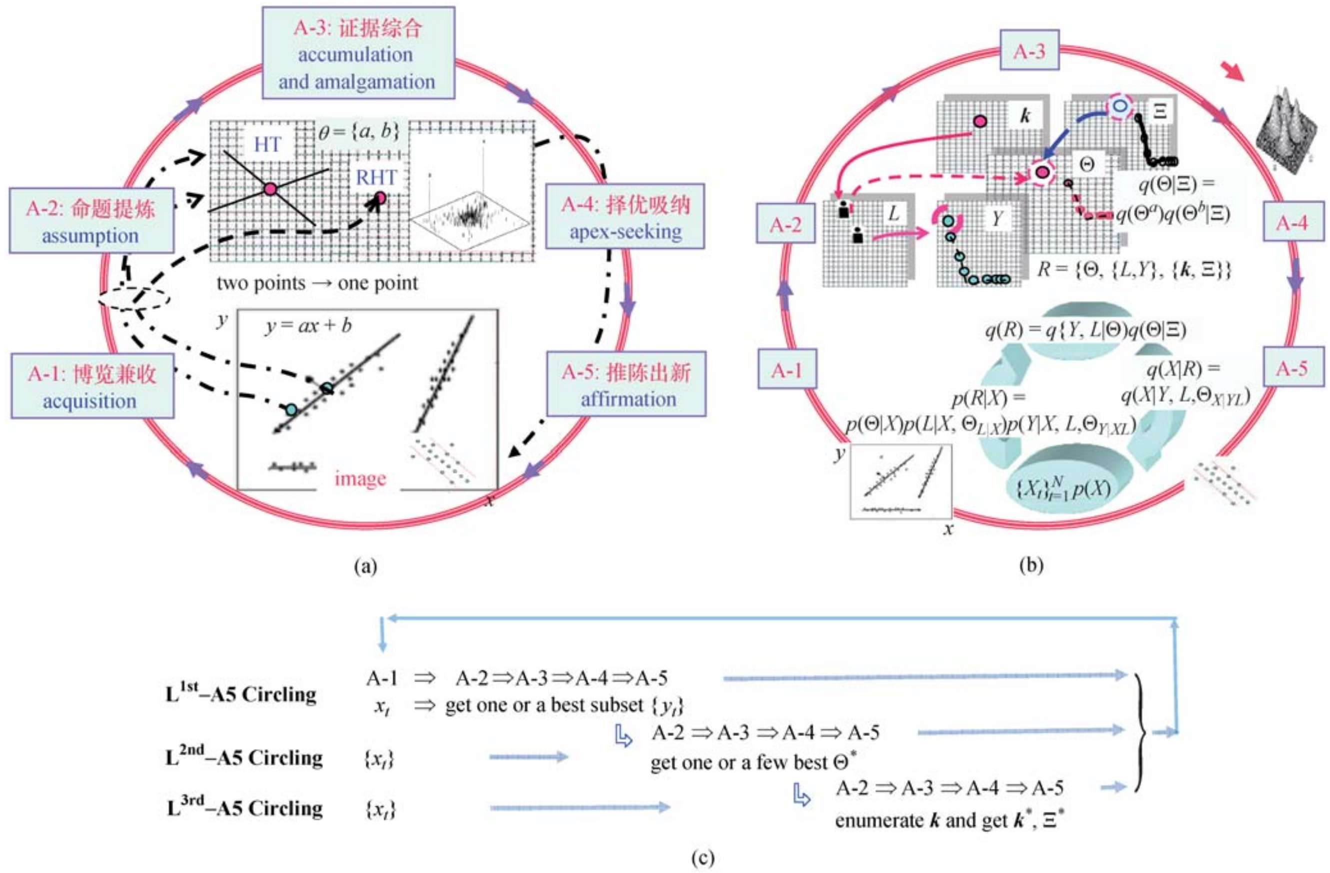

A-3: 证据综合
accumulation
and amalgamation
A-2: 命题提炼
assumption
A-4: 择优吸纳
apex-seeking
A-1: 博览兼收
acquisition
A-5: 推陈出新
affirmation
HT
$\theta = \{a, b\}$
RHT
two points → one point
y
$y = ax + b$
image
x
(a)
A-3
A-2
A-4
A-1
A-5
k
Ξ
Θ
L
Y
$q(\Theta|\Xi) = q(\Theta^a)q(\Theta^b|\Xi)$
$R = \{\Theta, \{L, Y\}, \{k, \Xi\}\}$
$q(R) = q\{Y, L|\Theta)q(\Theta|\Xi)$
$q(X|R) = q(X|Y, L, \Theta_{X|YL})$
$p(R|X) = p(\Theta|X)p(L|X, \Theta_{L|X})p(Y|X, L, \Theta_{Y|XL})$
$\{X_t\}_{t=1}^N\ p(X)$
y
x
(b)
$\text{L}^{1\text{st}}$–A5 Circling
A-1 ⇒ A-2⇒A-3⇒A-4⇒A-5
x_t ⇒ get one or a best subset $\{y_t\}$
$\text{L}^{2\text{nd}}$–A5 Circling
$\{x_t\}$
↳ A-2 ⇒A-3 ⇒A-4⇒ A-5
get one or a few best Θ^*
$\text{L}^{3\text{rd}}$–A5 Circling
$\{x_t\}$
↳ A-2 ⇒A-3 ⇒A-4⇒A-5
enumerate $\boldsymbol{k}$ and get $\boldsymbol{k}^*$, Ξ^*
(c)

Fig. 6 Five basic actions and circular implementation. (a) HT, RHT, and A5 circling; (b) BYY system and A5 circling; (c) three levels of A5 circling

was independently proposed according to the nature of intelligent problem solving, which well coincides with the one of WuXing theory. Moreover, Ke-Cheng-Hui jointly means that to reverse a unbalancing tendency of one state should be made from the one before the preceding one, towards a correct direction and with an appropriate strength. Even interestingly, to tackle a bottleneck of a huge computing load by an accumulation array of HT, one key point of RHT is using a many-to-one converging mapping to replace the one-to-many diverging mapping by HT [52,78]. This is just a case of changing A-2 to improve a bottleneck of A-4, which well coincides with the Ke-Cheng-Hui law.

The A5 circle also applies to the two tasks in Figs. 1(b) and 1(c) as well as the general tasks in Fig. 2. As shown in Fig. 6(b), the assumption space consists of three layers for three levels of inverse problems. Correspondingly, the first layer encodes the inner representation of the problem in Fig. 2(a) in help of an A5 circling (shortly $\text{L}^{1\text{st}}$-A5 circling). For the example in Fig. 1(b), the task of pattern recognition is featured by the action A-4 that gets the apex j^* of $p(j|x_t)$ or a subset $C_\kappa(x_t)$ that consists of j^* and its κ climax neighbors (see Eq. (38)) as the decision. The A-4 action is supported by the A-3 action that combines new evidence $G(x_t|\mu_j, \Sigma_j)$ with α_j for ones accumulated in past into $p(j|x_t)$. Upon receiving one or a set of samples of x_t obtained by the A-1 action, all the values of j is enumerated by A-2 as candidate assumptions. Finally, A-5 assesses whether these samples are well described by the corresponding Gaussian component that they are classified to. When k is not too large, it is tractable to handle this enumeration based computation for implementing the $\text{L}^{1\text{st}}$-A5 circling. However, when k is too large, the computing burdens of A-4 and A-3 become too loaded. In such a situation, the bottleneck of A-4 can be lessened by restricting A-2 to provide only a subset of candidate assumptions that the best one is still included, e.g., considering those with the value of $G(x_t|\mu_j, \Sigma_j)\alpha_j$ being above a pre-specified threshold.

For the example in Fig. 1(c), the task of the action A-4 is getting the apex $y^* = \arg\max_y f(y)$, $f(y) = q(x|y, \theta_{x|y})q(y|\theta_y)$. For a Gaussian $q(x|y, \theta_{x|y})$ and a Gaussian $q(y|\theta_y)$, a lumped computation of A-2, A-3, and A-4 is analytically solvable. However, when either $q(x|y, \theta_{x|y})$ or $q(y|\theta_y)$ is non-Gaussian, $\max_y f(y)$ becomes a hard nonlinear optimization, which is usually tackled by a line search $y^{\text{new}} = y^{\text{old}} + \eta g_f(y)$ along a direction $g_f(y)$ that ascends $f(y)$, e.g., the gradient direction of $f(y)$. This is another example that coincides with the Ke-Cheng-Hui law. To tackle a bottleneck for $\max_y f(y)$ at A-4, we go back A-2 to consider those candidates merely along one direction from the current $y^{\rm old}$, instead of searching the entire domain of y. In accordance to Proposition 4 in Appendix B, the L$^{\rm 1st}$-A5 circling is actually handled by a series of smaller circling within the same layer y. For each small circling, its A-2 action considers candidates merely along a direction $g_f(y)$, its A-3 action seeks the value of a scalar η, and its A-4 action compares if the difference $y^{\rm new} - y^{\rm old}$ becomes ignorable. As a result, a difficulty is solved by a series of lower level but much balanced and easy computing, while the goal of seeking a global optimum is downgraded to seeking a local optimum.

The second layer A5 circling (shortly L$^{\rm 2nd}$-A5 circling) drives parameter learning $X_N \to \Theta$, as shown in Fig. 2(b). We still take the problem of Fig. 1(b) as an example, making $\max_\Theta F(\Theta)$ in term of either the ML with $F(\Theta) = \ln q(X_N|\Theta)$ or Bayesian learning with $F(\Theta) = \ln q(X_N|\Theta)q(\Theta|\Xi)$ or BYY harmony learning with $F(\Theta) = H(p||q,\Theta,h,\boldsymbol{k},\Xi)$. The task is again featured by getting $\Theta^* = \arg\max_\Theta F(\Theta)$ at the A-4 action. However, it usually becomes not tractable to handle A-2, A-3, and A-4 by an enumeration type of computations because of the following two problems:

1) $\max_\Theta F(\Theta)$ is generally a hard nonlinear optimization. Similar to $y^{\rm new} = y^{\rm old} + \eta g_f(y)$, this problem is usually tackled by a line search $\Theta^{\rm new} = \Theta^{\rm old} + \eta g_F(\Theta)$ along a direction $g_F(\Theta)$ that ascends $F(\Theta)$.

2) $q(X_N|\Theta)$ or generally $F(\Theta)$ comes from a summation over j (e.g., in Fig. 1(b)) or an integral over y (e.g., in Fig. 1(c)), which will be very computational intensive either for a large k or for a non-Gaussian vector y. This problem is handled with the L$^{\rm 1st}$-A5 circling at either or both of its action A-2 and A-4 in help of apex approximation (see Sect. 4.2) that gets the apex j^* and y^* together with a subset of neighbors (see Eqs. (27)–(30) and Eq. (38)) to be summed or integrated.

The third layer A5 circling (shortly L$^{\rm 3rd}$-A5 circling) conducts model selection $X_N \to \boldsymbol{k},\Xi$ as shown in Fig. 2(c) or simply $X_N \to \boldsymbol{k}$ when $q(\Theta|\Xi)$ is prespecified with Ξ. The task becomes $\min_k J(\boldsymbol{k})$, $J(\boldsymbol{k}) = -\ln q(X_N|\boldsymbol{k},\Xi)$, which is usually called Bayesian approach or marginal Bayes [27,60–62,67,68], as shown in Figs. 4(c) and 4(d). Typically, an L$^{\rm 3rd}$-A5 circling starts at enumerating $\boldsymbol{k}$ for a series of values in its A-2 action. Again, not only this enumeration is computationally intensive (especially when $\boldsymbol{k}$ consists of a number of integers), but also evaluating $J(\boldsymbol{k})$ by either $J(\boldsymbol{k}) = -\ln q(X_N|\boldsymbol{k},\Xi)$ or $J(\boldsymbol{k}) = H(p||q,\boldsymbol{k},\Xi)$ in Fig. 5(b) involves an integral over Θ, which is computationally intractable. To tackle the blockage, the integral is approximated such that $J(\boldsymbol{k})$ becomes an additive approximation $J(\boldsymbol{k}) = F(\Theta^*,\boldsymbol{k}) + d_{\boldsymbol{k}}$. As a result, evaluating $J(\boldsymbol{k})$ involves an L$^{\rm 2nd}$-A5 circling for parameter learning $X_N \to \Theta^*$ per value of $\boldsymbol{k}$, which is further embedded within an L$^{\rm 3rd}$-A5 circling as one step. In other words, a two stage implementation of model selection is actually an L$^{\rm 3rd}$-A5 circling with a series of L$^{\rm 2nd}$-A5 circling for parameter learning, which is a compliment of Proposition 4 in Appendix B.

Still, a two stage implementation of model selection is too computationally intensive to be impractical. Instead, automatic model selection does not implement L$^{\rm 3rd}$-A5 circling but considers an L$^{\rm 2nd}$-A5 circling at a big initial value for $\boldsymbol{k}$, with extra parts deducted effectively by Eq. (4), as a subset of Θ is discarded during the L$^{\rm 2nd}$-A5 circling. As discussed previously about Figs. 4(c) and 4(d), the force of driving $\psi(\theta_k) \to 0$ by Eq. (4) comes from an external prior $q(\Theta)$ that is in an action within the same level L$^{\rm 2nd}$ and is modulated by Ξ from an upper level. For automatic model selection by the BYY harmony learning via $\max_{\Theta,h} H(p||q,\Theta,h,\boldsymbol{k},\Xi)$ in Fig. 5(b), an even important force comes from Eq. (5) intrinsically by a series of L$^{\rm 2nd}$-A5 circling in the lower level, in addition to the one that comes from using a priori $q(\Theta)$.

In a summary, the problem solving paradigm A5 and WuXing theory provide new insights on the HT-RHT based object detection and three levels of inverse problems, which provides a guide to develop new algorithms. For solving any problem, A-1 and A-5 are generally easy to be identified because one is associated with input while the other is associated with output. However, the actions A-2, A-3 and A-4 are usually lumped together as one optimization task that is expensive to compute, for which a guide is to allocate the task among three actions in a balanced way, e.g., as above discussed for $\max_\Theta F(\Theta)$. Also, we are guided by the Ke-Cheng-Hui law to reverse an unbalancing tendency of one state.

3 Starting from three exemplar learning algorithms

3.1 On Gaussian mixture: BYY learning algorithms versus other learning algorithms

For Gaussian mixture introduced in Fig. 1(b), we start from the standard EM algorithm for the ML learning [7,8], which is here used as a benchmark for better insights on the Bayesian Ying-Yang harmony learning and other related algorithms. Specifically, the E-step simply gets the Bayesian posteriori as follows:

$$\begin{aligned} &p_{j,t} = p\left(j|x_t,\theta^{\rm old}\right),\ p\left(j|x_t,\theta^{\rm old}\right) = q\left(j|x_t,\theta^{\rm old}\right), \\ &q\left(j|x_t,\theta\right) = \frac{q\left(x_t\left|\theta_j\right.\right)\alpha_j}{\sum_j q\left(x_t\left|\theta_j\right.\right)\alpha_j}, \end{aligned} \tag{7}$$

which is obtained with the Ying machine $G(x|\mu_j,\Sigma_j)\alpha_j$ fixed at the last updating $\theta^{\rm old}$ of the M-step. With $p_{j,t}$

fixed, the M-step renews θ^{old} into θ^{new} as follows:

$$\begin{aligned} &N_\ell = \sum_t p_{\ell,t}, \ \ \alpha_\ell^{\text{new}} = \frac{N_\ell}{N}, \\ &\mu_\ell^{\text{new}} = \frac{1}{N_\ell}\sum_t p_{\ell,t} x_t, \\ &\Sigma_\ell^{\text{new}} = \frac{1}{N_\ell}\sum_t p_{\ell,t}\delta\Sigma_{\ell,t}, \\ &\delta\Sigma_{\ell,t} = \left(x_t - \mu_\ell^{\text{old}}\right)\left(x_t - \mu_\ell^{\text{old}}\right)^{\text{T}}, \end{aligned} \tag{8}$$

which comes from maximizing the following auxiliary function $M(\theta|\theta^{\text{old}})$ via solving θ^{new} from $\nabla_{\theta_j} M\left(\theta\left|\theta^{\text{old}}\right.\right) = 0$:

$$\begin{aligned} M\left(\theta\left|\theta^{\text{old}}\right.\right) &= \sum_{t,\ell} p\left(\ell\left|x_t, \theta^{\text{old}}\right.\right)\pi_t\left(\theta_\ell\right), \\ \pi_t\left(\theta_j\right) &= \ln\left[q\left(x_t\left|\theta_j\right.\right)\alpha_j\right] \quad \text{and} \\ \nabla_{\theta_\ell} M\left(\theta\left|\theta^{\text{old}}\right.\right) &= \sum_t G_t\left(\theta_\ell\right), \ G_t\left(\theta_\ell\right) = p_{\ell,t}\nabla_{\theta_\ell}\pi_t\left(\theta_\ell\right). \end{aligned} \tag{9}$$

The EM algorithm alternates the E-step and the M-step to implement the ML learning, such that the likelihood $\ln q(X_N|\Theta)$ gradually increases to reach at least one local maximum.

In contrast, the BYY harmony learning maximizes the harmony functional by Eq. (2) that becomes here

$$\begin{aligned} H(\theta) &= \sum_{t,\ell} p\left(\ell|x_t, \theta\right) H_t\left(\theta_\ell\right), \\ H_t\left(\theta_j\right) &= \pi_t\left(\theta_j\right) + R\left(h, \theta_j\right), \\ R\left(h, \theta_j\right) &= \ln\left[q\left(h\left|X_N\right.\right) q\left(\theta_j\right)\right] - \frac{1}{2}\text{Tr}\left[h^2\Sigma_j^{-1}\right], \end{aligned} \tag{10}$$

where $p\left(\ell|x_t, \theta\right)$ is given by Eq. (38) according to the variety preservation principle (see Sect. 4.2), with its detailed derivation delayed to Sect. 4.3. This $H(\theta)$ is maximized by a Ying-Yang alternation procedure with its Ying step sharing a same expression of the M-step by Eq. (8). However, its Yang step comes from modifying the E-step by Eq. (7) in two places.

First, a key point is that $p_{j,t} = p\left(j\left|x_t\right.\right)$ is replaced by

$$\begin{aligned} &p_{j,t} = p\left(j\left|t\right.\right) + \Delta_{j,t}, \\ &p\left(j\left|t\right.\right) = p\left(j\left|x_t, \theta^{\text{old}}\right.\right), \ \Delta_{j,t} = \Delta_{j,t}\left(\theta^{\text{old}}\right), \\ &\Delta_{j,t}(\theta) = p\left(j\left|x_t, \theta\right.\right)\left[H_t\left(\theta_j\right) - \sum_j p\left(j\left|x_t, \theta\right.\right) H_t\left(\theta_j\right)\right], \end{aligned} \tag{11}$$

where $H_t(\theta_j)$ consists of $\pi_t(\theta_j)$ in Eq. (9) for describing the fitness of the jth component on the sample x_t plus a regularization term $R(h, \theta_j)$ for considering a priori $q(\theta_j)$ and data smoothing regularization by Eq. (6). Specifically, $\Delta_{j,t}$ >0 means that the jth component is better than the average of all the components in term of this regularized fitness $H_t(\theta_j)$. We thus update the jth component in Eq. (8) to enhance its contribution of x_t. If $0 > \Delta_{j,t} > -1$, i.e., the regularized fitness $H_t(\theta_j)$ by the jth component is below the average but not too far away, the contribution of x_t on updating the jth component remains a same trend as in Eq. (8) but with a reduced strength. Moreover, when $-1 > \Delta_{j,t}$, the updating on the jth component reverses the direction to become de-learning, somewhat similar to updating the rival in RPCL learning [64–66,77]. As previously discussed, RPCL learning incurs automatic model selection. More precisely, the automatic model selection nature of Eq. (11) can be better understood from Eq. (5) that includes a part of maximizing

$$\sum_{j=1}^k \alpha_j \ln \alpha_j, \quad \alpha_j = \sum_t p_{j,t} / \sum_{j,t} p_{j,t},$$

which drives $\alpha_j \to 0$ if the jth Gaussian is extra.

Second, $p\left(j\left|x_t, \theta^{\text{old}}\right.\right) = q\left(j\left|x_t, \theta^{\text{old}}\right.\right)$ in Eq. (7) is replaced by the following first equation that holds not just by a value fixed at θ^{old}, and we focus on considering merely the first κ largest $H_t(\theta_j)$ as follows:

$$\begin{aligned} p\left(j\left|x_t, \theta\right.\right) &= q\left(j\left|x_t, \theta\right.\right)\chi_{\kappa,t}\left(j\right), \\ H_t\left(\theta_j\right) &= \left[\pi_t\left(\theta_j\right) + R\left(h, \theta_j\right)\right]\chi_{\kappa,t}\left(j\right), \\ \chi_{\kappa,t}\left(j\right) &= \begin{cases} 1, i \in C_{j,t}^\kappa, \\ 0, i \notin C_{j,t}^\kappa, \end{cases} \\ C_{j,t}^\kappa &= \left\{j: \ \ \text{the} \ \ \kappa \leqslant k \text{ largests of } H_t\left(\theta_j\right)\right\}, \end{aligned} \tag{12}$$

which comes from the Ying-Yang variety preservation principle, see Eqs. (28) and (38). Accordingly, updating of Eq. (8) focuses on only the first κ apex terms, or called apex approximation. This nature may be further understood by observing the following three special cases:

1) RPCL learning [64–66] can be regarded as a simplification of the case $\kappa = 2$ with $p_{j,t} = p(j|x_t) + \Delta_{j,t}$ replaced by the following setting

$$\begin{aligned} &p_{j,t} = \delta_{jj^*} - \gamma\delta_{jj_r}, \quad 0 < \gamma \ll 1, \\ &j^* = \arg\max_i H_t\left(\theta_i\right), \quad j_r = \arg\max_{i \neq j^*} H_t\left(\theta_i\right), \end{aligned}$$

where δ_{ij} is the Kronecker delta, i.e., $\delta_{ij} = 1$ if $i = j$, otherwise $\delta_{ij} = 0$.

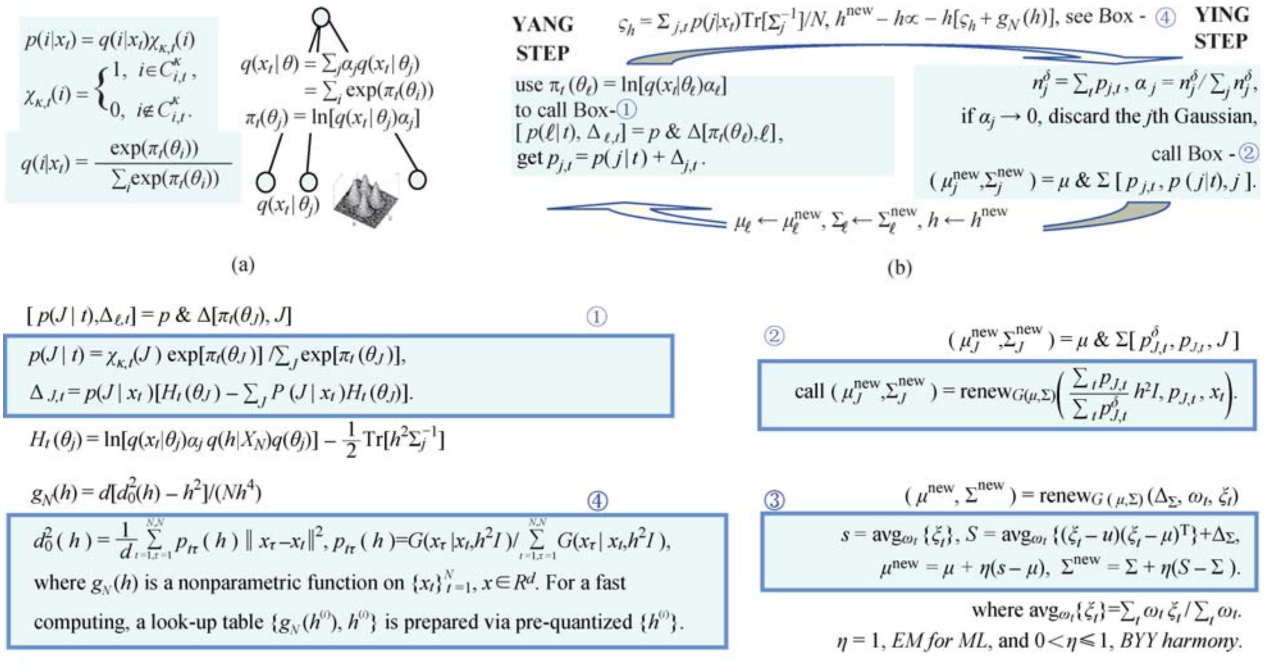
2) If $\kappa = 1$, we are simply lead to $p_{j,t} = \delta_{jj^*}, j^* = \arg\max_i H_t(\theta_i)$, which can be further simplified into Bayes winner-take-all (WTA) or MAP classifier [1] if $H_t(\theta_i)$ degenerates back to $\pi_t(\theta_j)$ in Eq. (9). The corresponding BYY harmony learning is thus called the WTA based BYY harmony learning [77]. Still, the nature $\alpha_j \to 0$ applies but becomes weak because $\Delta_{j,t} = 0$. Also, this case is previously referred as the hard-cut EM algorithm (see Sect. 3 in Ref. [80]) since it also comes from modifying the EM algorithm via hard-cutting $p_{j,t} = p(j|x_t)$ into $p_{j,t} = \delta_{jj^*}$.

3) For the cases of $1 < \kappa < k$, if we force $\Delta_{j,t} = 0$ and let $H_t(\theta_j)$ to degenerate back to $\pi_t(\theta_j)$ in Eq. (9), we are lead to one variant of the EM based ML learning by focusing on the first κ maximum posteriori ones (κ-MAP), e.g., see Table 2 and Eqs. (28) and (22) in Ref. [77], which is actually on a halfway between the EM and its hard-cut variant (i.e., $\kappa = 1$).

Due to the first equation in Eq. (12), $\nabla_{\theta_j} H(\theta)$ should consider not only $\sum_{j,t} p(j\,|x_t, \theta)\nabla_{\theta_j} H_t(\theta_j)$, i.e., a counterpart of $\nabla_{\theta_j} M\left(\theta\,\middle|\theta^{\rm old}\right)$ in Eq. (9), but also a term $\sum_{j,t} H_t(\theta_j)\nabla_{\theta_j} p(j\,|x_t, \theta)$. As a result, we get $\nabla_{\theta_\ell} H(\theta) = \sum_t G_t(\theta_\ell)$, $G_t(\theta_\ell) = p_{\ell,t}\nabla_{\theta_\ell}\pi_t(\theta_\ell) + p(\ell\,|t)\nabla_{\theta_\ell} R(h\,|\theta_\ell)$, with $p(\ell\,|t)$ and $p_{\ell,t}$ by Eq. (11).

As shown in Fig. 7, we summarize the EM algorithm and the BYY harmony learning, as well as the above special cases into a unified Ying-Yang alternation procedure. The Ying step implements Eq. (8) with its second line for detecting $\alpha_j \to 0$, while the Yang step takes options on setting $\Delta_{j,t}$ and κ for different algorithms. This procedure is featured with calling those subroutines in terms of the Boxes in Fig. 7. Also, these boxes act as the building bricks of the algorithms in the subsequent sections.

In addition, considering the second term of the above $G_t(\theta_j)$, we are lead to another difference. First, we let $\delta\Sigma_{l,t}$ in the M-step to be replaced by $\delta\Sigma_{l,t} + h^2 I$, which makes the EM algorithm by Eqs. (7) and (8) improved into a smoothed EM algorithm for Gaussian mixture, see Eq. (18) in Ref. [73]. Second, we are also lead to updating h at the top of Fig. 7(b) by calling the Box-④,


$p(i|x_t) = q(i|x_t)\chi_{\kappa,t}(i)$
$\chi_{\kappa,t}(i) = \begin{cases} 1, & i \in C^{\kappa}_{i,t}, \\ 0, & i \notin C^{\kappa}_{i,t}. \end{cases}$
$q(i|x_t) = \frac{\exp(\pi_t(\theta_i))}{\sum_i \exp(\pi_t(\theta_i))}$
$q(x_t|\theta) = \sum_j \alpha_j q(x_t|\theta_j) = \sum_i \exp(\pi_t(\theta_i))$
$\pi_t(\theta_j) = \ln[q(x_t|\theta_j)\alpha_j]$
$q(x_t|\theta_j)$
(a)
YANG STEP
$\varsigma_h = \Sigma_{j,t} p(j|x_t)\mathrm{Tr}[\Sigma_j^{-1}]/N$, $h^{\rm new} - h \propto -h[\varsigma_h + g_N(h)]$, see Box - ④
YING STEP
use $\pi_t(\theta_\ell) = \ln[q(x_t|\theta_\ell)\alpha_\ell]$
to call Box-①
$[\, p(\ell|t), \Delta_{\ell,t}] = p\ \&\ \Delta[\pi_t(\theta_\ell), \ell]$,
get $p_{j,t} = p(j|t) + \Delta_{j,t}$.
$n_j^\delta = \sum_t p_{j,t}$, $\alpha_j = n_j^\delta / \sum_j n_j^\delta$,
if $\alpha_j \to 0$, discard the jth Gaussian,
call Box - ②
$(\mu_j^{\rm new}, \Sigma_j^{\rm new}) = \mu\ \&\ \Sigma\,[\,p_{j,t}, p(j|t), j\,]$.
$\mu_\ell \leftarrow \mu_\ell^{\rm new}$, $\Sigma_\ell \leftarrow \Sigma_\ell^{\rm new}$, $h \leftarrow h^{\rm new}$
(b)
$[\,p(J\,|\,t), \Delta_{\ell,t}] = p\ \&\ \Delta[\pi_t(\theta_J), J]$
①
$p(J\,|\,t) = \chi_{\kappa,t}(J)\exp[\pi_t(\theta_J)] / \sum_J \exp[\pi_t(\theta_J)]$,
$\Delta_{J,t} = p(J\,|\,x_t)[H_t(\theta_J) - \sum_J P(J\,|\,x_t) H_t(\theta_J)]$.
$H_t(\theta_j) = \ln[q(x_t|\theta_j)\alpha_j q(h|X_N) q(\theta_j)] - \frac{1}{2}\mathrm{Tr}[h^2\Sigma_j^{-1}]$
$g_N(h) = d[d_0^2(h) - h^2]/(Nh^4)$
④
$d_0^2(h) = \frac{1}{d}\sum_{t=1,\tau=1}^{N,N} p_{t\tau}(h)\,\| x_\tau - x_t \|^2$, $p_{t\tau}(h) = G(x_\tau|x_t, h^2 I) / \sum_{t=1,\tau=1}^{N,N} G(x_\tau|x_t, h^2 I)$,
where $g_N(h)$ is a nonparametric function on $\{x_t\}_{t=1}^N$, $x \in R^d$. For a fast computing, a look-up table $\{g_N(h^{(i)}), h^{(i)}\}$ is prepared via pre-quantized $\{h^{(i)}\}$.
②
$(\mu_J^{\rm new}, \Sigma_J^{\rm new}) = \mu\ \&\ \Sigma[\,p^\delta_{J,t}, p_{J,t}, J\,]$
call $(\mu_J^{\rm new}, \Sigma_J^{\rm new}) = \mathrm{renew}_{G(\mu,\Sigma)}\left(\frac{\sum_t p_{J,t}}{\sum_t p^\delta_{J,t}} h^2 I, p_{J,t}, x_t\right)$.
③
$(\mu^{\rm new}, \Sigma^{\rm new}) = \mathrm{renew}_{G(\mu,\Sigma)}(\Delta_\Sigma, \omega_t, \xi_t)$
$s = \mathrm{avg}_{\omega_t}\{\xi_t\}$, $S = \mathrm{avg}_{\omega_t}\{(\xi_t - u)(\xi_t - \mu)^{\rm T}\} + \Delta_\Sigma$,
$\mu^{\rm new} = \mu + \eta(s - \mu)$, $\Sigma^{\rm new} = \Sigma + \eta(S - \Sigma)$.
where $\mathrm{avg}_{\omega_t}\{\xi_t\} = \sum_t \omega_t \xi_t / \sum_t \omega_t$.
$\eta = 1$, *EM for ML*, and $0 < \eta \leqslant 1$, *BYY harmony*.
(c)

Remarks (readers are suggested to skip this algorithm and details if you are not interested in programming):

a) h is updated at the top of Fig. 7(b) by Box-④, coming from $\frac{\rm d}{\rm dh}R(h, \theta_\ell)$, with $q(h) \propto \left[\frac{1}{N}\sum_{t=1,\tau=1}^{N,N} G(x_\tau|x_t, h^2 I)\right]^{-1}$ and with $q(\theta_j)$ ignored, where and hereafter in this paper, $x \sim y$ means $x = cy$ with a unspecified scalar constant c, e.g., $x^{\rm new} - x \sim y$ means $x^{\rm new} = x + cy$.

b) Typical options

- *EM* algorithm: setting $\Delta_{j,t} = 0, h = 0$, and $\kappa = k$, as weall as call Box-② that gets $\mu_\ell^{\rm new} = s_\ell$, $\Sigma_\ell^{\rm new} = \Sigma_\ell$ via Box-③ with $\eta = 1$.
- *WTA-BYY harmony* (see BYY-CL in Table 2 in Ref. [77]): setting $\kappa = 1, \Delta_{j,t} = 0$, $p(j|x_t) = p_{j,t} = \delta_{jj^*}$, $j^* = \arg\max_i H_t(\theta_i)$.
- *κ-MAP EM learning* (see BYY κ-map in Table 2 in Ref. [77]): $1 < \kappa < k$, forcing that $\Delta_{j,t} = 0$ and $H_t(\theta_j) = \pi_t(\theta_j)$.
- *RPCL learning* (see BYY-RPCL in Table 2 in Ref. [77]): $\kappa = 2$, $p_{j,t} = \delta_{jj^*} - \gamma\delta_{jj_r}$, $0 < \gamma \ll 1$, $j^* = \arg\max_i H_t(\theta_i)$, $j_r = \arg\max_{i \neq j^*} H_t(\theta_i)$.
- *BYY hamrony learning*: $\Delta_{j,t} \neq 0$ and $\kappa \leqslant k$ with and without $R(h, \theta_\ell) = \ln[q(h)q(\theta_\ell)]$, in options that $\ln q(\theta_\ell) \neq 0$ introduces a priori and $\ln q(h) \neq 0$ introduces a data smoothing regularization.

c) Additional note: since $p_{j,t}$ may be negative, calling the Box-③ may not guarantee $\Sigma_j^{\rm new}$ to keep nonnegative definite, which is handled by an appropriate $\eta > 0$ in calling the Box-③ for a linear interpolation from μ_j, Σ_j along the direction $s_j - \mu_j$ and $S_j - \Sigma_j$ that has a positive projection on the gradient directions $\nabla_{\theta_j} H(\theta)$.

Fig. 7 A unified Ying-Yang alternation procedure for typical learning algorithms on Gaussian mixture. (a) Gaussian mixture; (b) main program; (c) sub-programs

which introduces data smoothing regularization in the BYY harmony learning.

The last but not the least, in addition to automatic model selection during the BYY harmony learning, a further improvement can be obtained under the criterion $J(k)$ as shown in Fig. 5(b) by a two stage implementation, e.g., for the problem in Fig. 1(b), we have

$$\begin{aligned} J(k) = {} & 0.5\sum_{j=1}^{k}\alpha_j\left\{\ln|\Sigma_j| + h^2\mathrm{Tr}\left[\Sigma_j^{-1}\right]\right\} \\ & -\sum_{j=1}^{k}\alpha_j\ln\alpha_j + 0.5n_f\left(\Theta\right), \end{aligned} \tag{13}$$

where $n_f(\Theta)$ is the number of free parameters in Θ, e.g., $n_f(\Theta) = dk + k - 1 + 0.5d(d+1)k$ for the problem in Fig. 1(b).

3.2 FA and extensions

Next, we consider the FA in Fig. 1(c). Due to indeterminacy, the parameterization of FA has different choices. One conventional choice is that $\Lambda = I$ and A is a general $d \times m$ matrix [9,10]. One other choice is that Λ is a diagonal matrix and A is a $d \times m$ orthogonal matrix with $A^{\mathrm{T}}A = I$ [74(Sect. 3),81(Sect. 3.2)]. These two types of parameterization are equivalent in term of maximizing the likelihood $\sum_t \ln G\left(x_t\,|\mu,\Sigma_x\right)$ since it is possible to have $AA^{\mathrm{T}} + \Sigma = \Sigma_x = A\Lambda A^{\mathrm{T}} + \Sigma$. However, two types will become different in term of the BYY learning [17(Eq. (128))]. Actually, it has been empirically shown that the latter type is much better than the conventional one [82]. In the sequel, we do not differentiate the situations of Λ and A when we discuss the EM algorithm for the ML learning, while we actually consider a diagonal Λ and an orthogonal A with $A^{\mathrm{T}}A = I$ when we discuss the BYY harmony learning.

We consider the standard EM algorithm for the ML learning. The E-step gets $p(y|x) = G(y|y(x,\theta^{\mathrm{old}}),\Gamma(\theta^{\mathrm{old}}))$ as the Bayesian inverse of Ying machine with its parameters fixed at the last updating θ^{old}, that is, we get the MAP estimate $y(x,\theta^{\mathrm{old}}) = \arg\max_y[q(x|y)q(y)]$ and $\Gamma(\theta^{\mathrm{old}})$ via the following equations:

$$\begin{aligned} y\left(x,\theta\right) &= \Gamma(\theta)\left[A^{\mathrm{T}}\Sigma^{-1}\left(x_t-\mu\right)+\Lambda^{-1}\nu\right], \\ \Gamma(\theta) &= \left(A^{\mathrm{T}}\Sigma^{-1}A+\Lambda^{-1}\right)^{-1}. \end{aligned} \tag{14}$$

With $p(y|x) = G(y|y(x,\theta^{\mathrm{old}}),\Gamma(\theta^{\mathrm{old}}))$ fixed, the M-step maximizes the following auxiliary function:

$$\begin{aligned} M\left(\theta\,|\theta^{\mathrm{old}}\right) = {} & \sum_t \pi_t\left(\theta, y\left(x_t,\theta^{\mathrm{old}}\right)\right) \\ & -\frac{1}{2}\mathrm{Tr}\left[\Gamma\left(\theta^{\mathrm{old}}\right)\Pi^{y|x}\right], \\ \pi_t\left(\theta,y\right) = {} & \ln\left[G\left(x_t\,|Ay+\mu,\Sigma\right)G(y_t|\nu,\Lambda)\right], \\ \Pi^{y|x} = {} & A^{\mathrm{T}}\Sigma^{-1}A+\Lambda^{-1}, \end{aligned} \tag{15}$$

from which we get the following M-step to renew θ^{old} into θ^{new} as follows:

$$\begin{aligned} \mu^{\mathrm{new}} &= \frac{1}{N}\sum_t\left(x_t - A^{\mathrm{old}}y_t\right), \\ \Sigma^{\mathrm{new}} &= \Delta_{\Sigma}^{\mathrm{old}} + \frac{1}{N}\sum_t\left(x_t - A^{\mathrm{old}}y_t\right)\left(x_t - A^{\mathrm{old}}y_t\right)^{\mathrm{T}}, \\ \Delta_{\Sigma}^{\mathrm{old}} &= A^{\mathrm{old}}\Gamma\left(\theta^{\mathrm{old}}\right)A^{\mathrm{old\ T}}, \\ \nu^{\mathrm{new}} &= \frac{1}{N_\ell}\sum_t y_t, \\ \Lambda^{\mathrm{new}} &= \Gamma\left(\theta^{\mathrm{old}}\right) + \frac{1}{N}\sum_t\left(y_t - \nu^{\mathrm{old}}\right)\left(y_t - \nu^{\mathrm{old}}\right)^{\mathrm{T}}, \\ R_{xy} &= \frac{1}{N}\sum_t\left(x_t - A^{\mathrm{old}}y_t\right)\left(y_t - \nu^{\mathrm{old}}\right)^{\mathrm{T}}, \\ A^{\mathrm{new}} &= R_{xy}^{-1}\Lambda^{\mathrm{new}}. \end{aligned} \tag{16}$$

The above Eq. (14) and Eq. (16) are alternatively implemented as the E-step and the M-step of the EM algorithm, respectively, such that the likelihood $\ln q(X_N|\Theta)$ is maximized. It has an expression that is different from but equivalent to the conventional one [10]. We prefer this expression because its M-step shares a format that is not only same as the Ying step in Fig. 8 but also similar to the M-step by Eq. (8) of the EM algorithm for Gaussian mixture.

In contrast, the BYY harmony learning maximizes the harmony functional by Eq. (2) that becomes here:

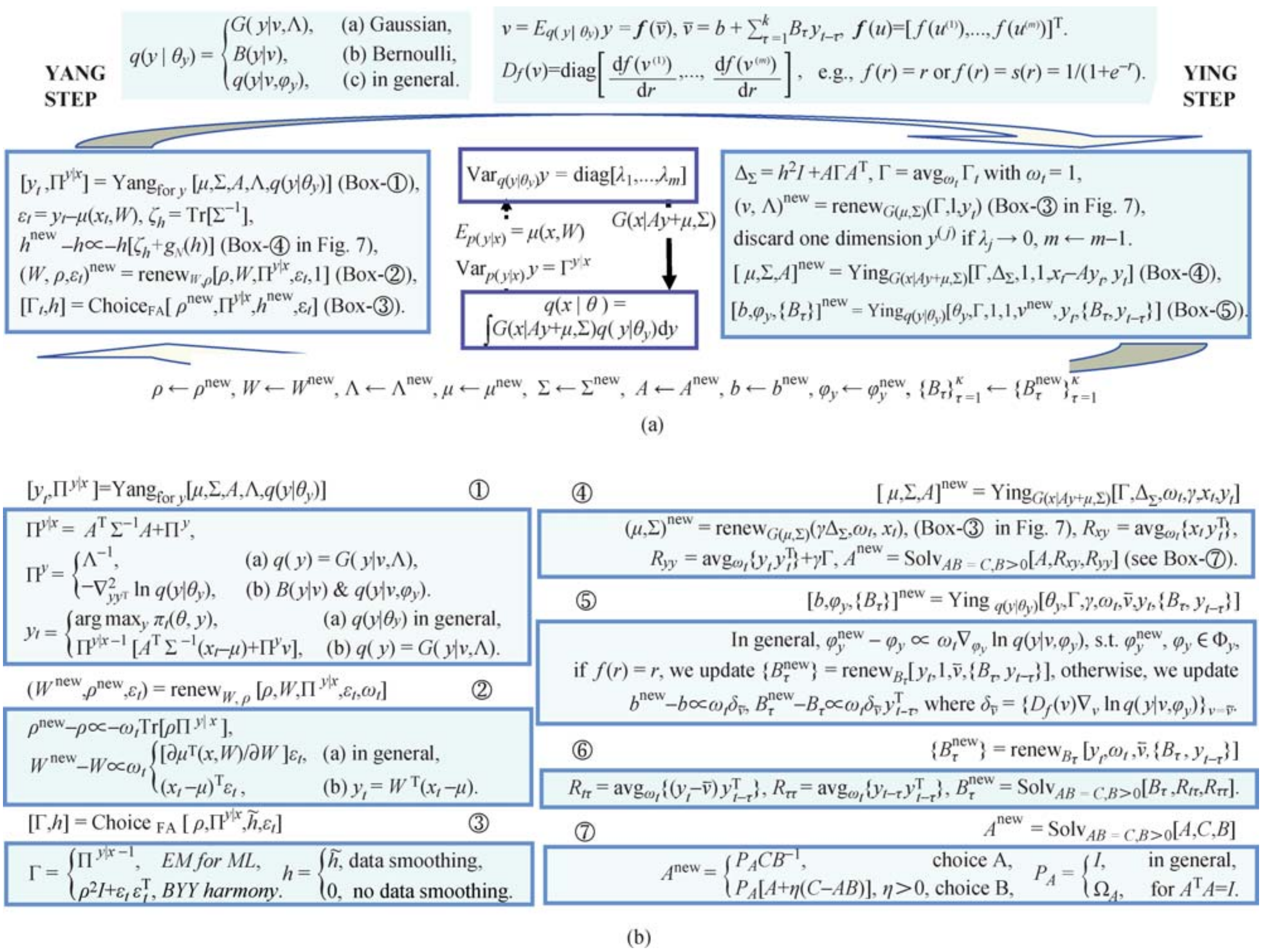
$$\begin{aligned} H(\theta) &= \sum_t[\pi_t\left(\theta,y_t\right)+R\left(h,\theta\right)], \\ R\left(h,\theta\right) &= \ln\left[q\left(h\,|X_N\right)q(\theta)\right] - \frac{1}{2}\mathrm{Tr}\left[h^2\Sigma^{-1}\right] \\ & \quad -\frac{1}{2}\mathrm{Tr}\left[\left(\Gamma^{y|x}+\varepsilon_t\varepsilon_t^{\mathrm{T}}\right)\Pi^{y|x}\right], \\ \varepsilon_t &= y_t - E_{p(y|x)}y,\ E_{p(y|x)}y = \mu\left(x,W\right), \\ \Gamma^{y|x} &= \mathrm{Var}_{p(y|x)}y = \Pi^{y|x-1}+\rho^2, \end{aligned} \tag{17}$$

and $\mathrm{Tr}\left[\left(\Gamma^{y|x}+\varepsilon_t\varepsilon_t^{\mathrm{T}}\right)\Pi^{y|x}\right] = m + \mathrm{Tr}\left[\left(\rho^2+\varepsilon_t\varepsilon_t^{\mathrm{T}}\right)\Pi^{y|x}\right]$ with the detailed derivation delayed to Sect. 4.3, where $\Gamma^{y|x} = \Pi^{y|x\,-1}+\rho^2$ in a Yang structure $p(y|x) = G(y|\mu(x,W),\Gamma^{y|x})$ is given by Eqs. (30) and (31) according to the variety preservation principle (see Sect. 4.2), where $\rho^2 = \rho\rho$ and ρ is a diagonal matrix, or even simply a scalar matrix ρI. We maximize $H(\theta)$ by a Ying-Yang alternation procedure with its Ying step sharing a same expression of the M-step by Eq. (16), while its Yang is obtained by modifying the E-step by Eq. (14) as follows.

A key difference is that $\Gamma_t = \Gamma(\theta^{\mathrm{old}})$ by Eq. (14) is replaced with

$$\Gamma_t = \rho^2+\varepsilon_t\varepsilon_t^{\mathrm{T}}.$$

This difference comes from the difference between $p(y|x) = G(y|y_t(\theta^{\mathrm{old}}),\ \Gamma(\theta^{\mathrm{old}}))$ and $p(y|x) = G(y|\mu(x,W),\ \Gamma^{y|x})$ in that $\Gamma^{y|x}$ is not fixed at $\Gamma(\theta^{\mathrm{old}})$


YANG STEP
$q(y|\theta_y)=\begin{cases}G(y|\nu,\Lambda), & \text{(a) Gaussian,}\\ B(y|\nu), & \text{(b) Bernoulli,}\\ q(y|\nu,\varphi_y), & \text{(c) in general.}\end{cases}$
$\nu=E_{q(y|\theta_y)}y=f(\bar\nu),\ \bar\nu=b+\sum_{\tau=1}^{k}B_\tau y_{t-\tau},\ f(u)=[f(u^{(1)}),\dots,f(u^{(m)})]^{\mathrm T}.$
$D_f(\nu)=\mathrm{diag}\left[\frac{\mathrm df(\nu^{(1)})}{\mathrm dr},\dots,\frac{\mathrm df(\nu^{(m)})}{\mathrm dr}\right]$, e.g., $f(r)=r$ or $f(r)=s(r)=1/(1+e^{-r})$.
YING STEP
$[y_t,\Pi^{y|x}]=\mathrm{Yang}_{\mathrm{for}\,y}[\mu,\Sigma,A,\Lambda,q(y|\theta_y)]$ (Box-①),
$\varepsilon_t=y_t-\mu(x_t,W),\ \zeta_h=\mathrm{Tr}[\Sigma^{-1}],$
$h^{\mathrm{new}}-h\propto-h[\zeta_h+g_N(h)]$ (Box-④ in Fig. 7),
$(W,\rho,\varepsilon_t)^{\mathrm{new}}=\mathrm{renew}_{W,\rho}[\rho,W,\Pi^{y|x},\varepsilon_t,1]$ (Box-②),
$[\Gamma_t,h]=\mathrm{Choice}_{\mathrm{FA}}[\rho^{\mathrm{new}},\Pi^{y|x},h^{\mathrm{new}},\varepsilon_t]$ (Box-③).
$\mathrm{Var}_{q(y|\theta_y)}y=\mathrm{diag}[\lambda_1,\dots,\lambda_m]$
$E_{p(y|x)}=\mu(x,W)$
$\mathrm{Var}_{p(y|x)}y=\Gamma^{y|x}$
$G(x|Ay+\mu,\Sigma)$
$q(x|\theta)=\int G(x|Ay+\mu,\Sigma)q(y|\theta_y)\mathrm dy$
$\Delta_\Sigma=h^2I+A\Gamma A^{\mathrm T},\ \Gamma=\mathrm{avg}_{\omega_t}\Gamma_t$ with $\omega_t=1$,
$(\nu,\Lambda)^{\mathrm{new}}=\mathrm{renew}_{G(\mu,\Sigma)}(\Gamma,1,y_t)$ (Box-③ in Fig. 7),
discard one dimension $y^{(j)}$ if $\lambda_j\to0,\ m\leftarrow m-1.$
$[\mu,\Sigma,A]^{\mathrm{new}}=\mathrm{Ying}_{G(x|Ay+\mu,\Sigma)}[\Gamma,\Delta_\Sigma,1,1,x_t-Ay_t,y_t]$ (Box-④),
$[b,\varphi_y,\{B_\tau\}]^{\mathrm{new}}=\mathrm{Ying}_{q(y|\theta_y)}[\theta_y,\Gamma,1,1,\nu^{\mathrm{new}},y_t,\{B_\tau,y_{t-\tau}\}]$ (Box-⑤).
$\rho\leftarrow\rho^{\mathrm{new}},\ W\leftarrow W^{\mathrm{new}},\ \Lambda\leftarrow\Lambda^{\mathrm{new}},\ \mu\leftarrow\mu^{\mathrm{new}},\ \Sigma\leftarrow\Sigma^{\mathrm{new}},\ A\leftarrow A^{\mathrm{new}},\ b\leftarrow b^{\mathrm{new}},\ \varphi_y\leftarrow\varphi_y^{\mathrm{new}},\ \{B_\tau\}_{\tau=1}^{\kappa}\leftarrow\{B_\tau^{\mathrm{new}}\}_{\tau=1}^{\kappa}$
(a)
$[y_t,\Pi^{y|x}]=\mathrm{Yang}_{\mathrm{for}\,y}[\mu,\Sigma,A,\Lambda,q(y|\theta_y)]$ ①
$\Pi^{y|x}=A^{\mathrm T}\Sigma^{-1}A+\Pi^y,$
$\Pi^y=\begin{cases}\Lambda^{-1}, & \text{(a) } q(y)=G(y|\nu,\Lambda),\\ -\nabla^2_{yy^{\mathrm T}}\ln q(y|\theta_y), & \text{(b) } B(y|\nu)\ \&\ q(y|\nu,\varphi_y).\end{cases}$
$y_t=\begin{cases}\arg\max_y\pi_t(\theta,y), & \text{(a) } q(y|\theta_y) \text{ in general,}\\ \Pi^{y|x-1}[A^{\mathrm T}\Sigma^{-1}(x_t-\mu)+\Pi^y\nu], & \text{(b) } q(y)=G(y|\nu,\Lambda).\end{cases}$
$(W^{\mathrm{new}},\rho^{\mathrm{new}},\varepsilon_t)=\mathrm{renew}_{W,\rho}[\rho,W,\Pi^{y|x},\varepsilon_t,\omega_t]$ ②
$\rho^{\mathrm{new}}-\rho\propto-\omega_t\mathrm{Tr}[\rho\Pi^{y|x}],$
$W^{\mathrm{new}}-W\propto\omega_t\begin{cases}[\partial\mu^{\mathrm T}(x,W)/\partial W]\varepsilon_t, & \text{(a) in general,}\\ (x_t-\mu)^{\mathrm T}\varepsilon_t, & \text{(b) } y_t=W^{\mathrm T}(x_t-\mu).\end{cases}$
$[\Gamma,h]=\mathrm{Choice}_{\mathrm{FA}}[\rho,\Pi^{y|x},\tilde h,\varepsilon_t]$ ③
$\Gamma=\begin{cases}\Pi^{y|x-1}, & \textit{EM for ML,}\\ \rho^2I+\varepsilon_t\varepsilon_t^{\mathrm T}, & \textit{BYY harmony.}\end{cases}\quad h=\begin{cases}\tilde h, & \text{data smoothing,}\\ 0, & \text{no data smoothing.}\end{cases}$
④ $[\mu,\Sigma,A]^{\mathrm{new}}=\mathrm{Ying}_{G(x|Ay+\mu,\Sigma)}[\Gamma,\Delta_\Sigma,\omega_t,\gamma,x_t,y_t]$
$(\mu,\Sigma)^{\mathrm{new}}=\mathrm{renew}_{G(\mu,\Sigma)}(\gamma\Delta_\Sigma,\omega_t,x_t)$, (Box-③ in Fig. 7), $R_{xy}=\mathrm{avg}_{\omega_t}\{x_ty_t^{\mathrm T}\},$
$R_{yy}=\mathrm{avg}_{\omega_t}\{y_ty_t^{\mathrm T}\}+\gamma\Gamma,\ A^{\mathrm{new}}=\mathrm{Solv}_{AB=C,B>0}[A,R_{xy},R_{yy}]$ (see Box-⑦).
⑤ $[b,\varphi_y,\{B_\tau\}]^{\mathrm{new}}=\mathrm{Ying}_{q(y|\theta_y)}[\theta_y,\Gamma,\gamma,\omega_t,\bar\nu,y_t,\{B_\tau,y_{t-\tau}\}]$
In general, $\varphi_y^{\mathrm{new}}-\varphi_y\propto\omega_t\nabla_{\varphi_y}\ln q(y|\nu,\varphi_y)$, s.t. $\varphi_y^{\mathrm{new}},\varphi_y\in\Phi_y,$
if $f(r)=r$, we update $\{B_\tau^{\mathrm{new}}\}=\mathrm{renew}_{B_\tau}[y_t,1,\bar\nu,\{B_\tau,y_{t-\tau}\}]$, otherwise, we update
$b^{\mathrm{new}}-b\propto\omega_t\delta_{\bar\nu},\ B_\tau^{\mathrm{new}}-B_\tau\propto\omega_t\delta_{\bar\nu}y_{t-\tau}^{\mathrm T}$, where $\delta_{\bar\nu}=\{D_f(\nu)\nabla_\nu\ln q(y|\nu,\varphi_y)\}_{\nu=\bar\nu}.$
⑥ $\{B_\tau^{\mathrm{new}}\}=\mathrm{renew}_{B_\tau}[y_t,\omega_t,\bar\nu,\{B_\tau,y_{t-\tau}\}]$
$R_{t\tau}=\mathrm{avg}_{\omega_t}\{(y_t-\bar\nu)y_{t-\tau}^{\mathrm T}\},\ R_{\tau\tau}=\mathrm{avg}_{\omega_t}\{y_{t-\tau}y_{t-\tau}^{\mathrm T}\},\ B_\tau^{\mathrm{new}}=\mathrm{Solv}_{AB=C,B>0}[B_\tau,R_{t\tau},R_{\tau\tau}].$
⑦ $A^{\mathrm{new}}=\mathrm{Solv}_{AB=C,B>0}[A,C,B]$
$A^{\mathrm{new}}=\begin{cases}P_ACB^{-1}, & \text{choice A,}\\ P_A[A+\eta(C-AB)],\ \eta>0, & \text{choice B,}\end{cases}\quad P_A=\begin{cases}I, & \text{in general,}\\ \Omega_A, & \text{for } A^{\mathrm T}A=I.\end{cases}$
(b)

Remarks (readers are suggested to skip this algorithm and details if you are not interested in programming):

a) The Ying step implements Eq. (16), shared by all the algorithms. It consists of

- the 2nd line calls Box-③ for updating ν, Λ and then the 3rd line discards extra dimension.
- the 4th line calls Box-④ for updation $G(x|Ay+\mu,\Sigma)$ in help of calling Box-⑦ that either directly solves a linear equation or iteratively makes error correcting, with a projection to satisfy the orthogonality (if any).

b) The last line of Ying step calls Box-⑤ for the extensions of $q(y|\theta_y)$ as follows:

- binary factor analysis (BFA) with an independent multi-variate Bernoulli $B(y|\nu)=\prod_j(\nu^{(j)})^{y^{(j)}}(1-\nu^{(j)})^{1-y^{(j)}}$.
- non-Gaussian factor analysis (NFA) with a non-Gaussian $q(y_t|\nu,\varphi_y)$.
- temporal factor analysis (TFA) that learns temporal relation $\overline{\nu}=b+\sum_{\tau=1}^{\kappa}B_\tau y_{t-\tau}$ by calling Box-⑥.

c) In Box-⑦, P_A is an operator that projects A onto P_AA with $(P_AA)^{\mathrm T}P_AA=I$, e.g., $P_A=I-AA^{\mathrm T}$.

Fig. 8 A unified Ying-Yang alternation procedure for typical learning algorithms on various factor analysis. (a) Main program; (b) sub-programs

but becomes a function $\Gamma^{y|x}=(A\Sigma^{-1}A^{\mathrm T}+\Lambda^{-1})^{-1}+\rho^2$, which results in the third term in $R(h,\theta)$. Maximizing $H(\theta)$ includes minimizing the term $\Sigma_t\mathrm{Tr}[(\rho^2+\varepsilon_t\varepsilon_t^{\mathrm T})\Pi^{y|x}]$ that drives not only $y_t-\mu(x,W)=\varepsilon_t\to0$ but also $\rho\to0$. Thus, $\Gamma_t\to0$ gradually reduces its regularization role in Eq. (16) such that some $\lambda_j\to0$ may gradually emerge for automatic model selection by Eq. (4). There is no such a scenario in the EM algorithm because $\Gamma_t=\Gamma(\theta^{\mathrm{old}})$ will not tend to zero and its role in Eq. (16) will not disappear, which shields the elements of Λ to be pushed towards zero.

We notice that $\Gamma_t=\rho^2+\varepsilon_t\varepsilon_t^{\mathrm T}\to0$ leads to $\Gamma^{y|x}=\Pi^{y|x-1}$ and $\mathrm{Tr}[(\Gamma^{y|x}+\varepsilon_t\varepsilon_t^{\mathrm T})\Pi^{y|x}]=m$. Thus, this term becomes no effect on getting $\nabla_\theta H_t(\theta)$ for renewing θ^{old} into θ^{new}. Alternatively, we may be also lead to this situation either by letting $\rho=0$ and $\mu(x_t,W)=y_t$ to directly get $\Gamma_t=0$ or maximizing $H(\theta)$ with respect to a structure free $p(y|x)$ that directly becomes $p(y|x)=\delta(y-y_t)$, i.e., $\varepsilon_t=0$, $\Gamma^{y|x}=0$, $\mathrm{Tr}[\Gamma^{y|x}\Pi^{y|x}]=0$. Conceptually, all the three cases should reach a same limit that learning finally converges towards. Also, all the three are featured with the automatic model selection nature by Eq. (5) that includes a part of maximizing

$$\int p(y)\ln q(y|\theta_y)\,\mathrm dy,$$

with $p(y)=\frac{1}{N}\sum_t p(y|x_t)$ for $p(y|x)=\delta(y-y_t)$ or otherwise $p(y)=\int p(y|x)p_h(x)\mathrm dx$, which tends to

$\int p(y) \ln p(y) \mathrm{d}y$, and its further maximization pushes the variance λ_j of $y^{(j)} \to 0$ if the dimension is redundant.

However, the case that $\Gamma_t = \rho^2 + \varepsilon_t \varepsilon_t^{\mathrm{T}} \to 0$ do have the following two nontrivial differences:

1) As $\varepsilon_t \to 0$ and $\rho \to 0$ gradually, $\Gamma_t \to 0$ leads to a better performance than directly setting $\Gamma_t = 0$ that makes learning to be trapped in a local optimum.

2) Even in the limit case $\Gamma_t = 0$, having $\mathrm{Tr}[\Gamma^{y|x}\Pi^{y|x}] = m$ is still better than $\mathrm{Tr}[\Gamma^{y|x}\Pi^{y|x}] = 0$, because it follows from $H(\theta)$ in Eq. (17) that we get a term $\mathrm{Tr}[\Gamma^{y|x}\Pi^{y|x}] = m$ to be included in the following criterion $J(m) = -H(\theta)$ as shown in Fig. 5(b) for selecting m in a two stage implementation.

$$\begin{aligned} J(m) = \frac{1}{2} & \left[\ln |\Sigma| + h^2 \mathrm{Tr}\left[\Sigma^{-1}\right] \right. \\ & \left. + m + n_f(\Theta) + \ln|\Lambda| + m \ln(2\pi \mathrm{e}) \right], \\ \text{with } n_f(\Theta) = & \, m + md - \frac{1}{2} m(m+1). \end{aligned} \tag{18}$$

As shown in Fig. 8, we summarize the EM algorithm and the BYY harmony learning, as well as extensions to non-Gaussian factors and temporal dependence, in a unified Ying-Yang alternation procedure with major parts in a same expression by Eq. (16) that is implemented by the Ying step in Fig. 8(a), while differences simply characterized by options in the Yang step, especially via the last line calling the Box-③ $\mathrm{Choice}_{\mathrm{FA}}$. Specifically, the first line calls the Box-① $\mathrm{Yang}_{\mathrm{for}\, y}$ for getting $y_t = \arg\max_y [q(x_t|y)q(y)]$ and the corresponding $\Pi^{y|x}$ from which we further get Γ_t via the 5th line that calls the Box-③. Moreover, $\varepsilon_t \to 0$ and $\rho \to 0$ are controlled in help of calling the Box-② $\mathrm{renew}_{W,\rho}$.

Considering the term $R(h, \theta_j)$ in Eq. (17), we are lead to data smoothing regularization. First, we let Δ_Σ in Eq. (16) replaced by $\Delta_\Sigma + h^2 I$, which leads to a smoothed EM algorithm for factor analysis. Second, we update h in the Yang step by the third line in help of calling the Box-④ in Fig. 7(c), which introduces data smoothing regularization into the BYY harmony learning.

FA has been further extended by taking temporal dependence in consideration, resulting in temporal factor analysis (TFA) [83,84]. A detailed study on modeling temporal dependence will be further addressed in Sect. 5.1. One useful simplification is letting the mean vector of ν given by a linear regression of the past values of y, as shown on the top of Fig. 8(a). In this case, the task of getting y_t by the Yang step closely relates to the classical state space and Kalman filtering [5,85]. Beyond the Kalman filtering, TFA also learns parameters of Ying machine by the Ying step in Fig. 8(a).

FA is also extended to non-Gaussian factor analysis (NFA) with a Gaussian $q(y_t|\theta_y) = G(y_t|\nu, \Lambda)$ replaced by a non-Gaussian $q(y_t|\nu, \varphi_y)$. One example is binary factor analysis (BFA) with a binary vector y in a multivariate Bernoulli distribution $B(y_t|\nu)$, also see Appendix C 1)g). Typical examples with a real vector y include $q(y_t|\nu, \varphi_y)$ that is either a Gaussian mixture or a product of each component distribution in a Gaussian mixture. Further details are referred to Refs. [17,49,83], as well as to Ref. [18] for a systematic review. For the Yang step, Box-① in Fig. 8(b) considers $q(y_t|\nu, \varphi_y)$ in general with y_t obtained from $\max_y \pi(y, \theta_y)$. For $B(y|\nu)$, $\max_y \pi(y, \theta_y)$ is a quadratic combinatorial optimization which can be effectively handled by the algorithms investigated in Ref. [86].

The last but not the least, all the above algorithms can be further extended to a mixture of multiple individual ones located at multi-locations $\{\mu_j\}$, with parameters in each local model learned to be responsible locally for a cloud of samples. One typical example is local factor analysis (LFA) (see a review in Ref. [18]), which may

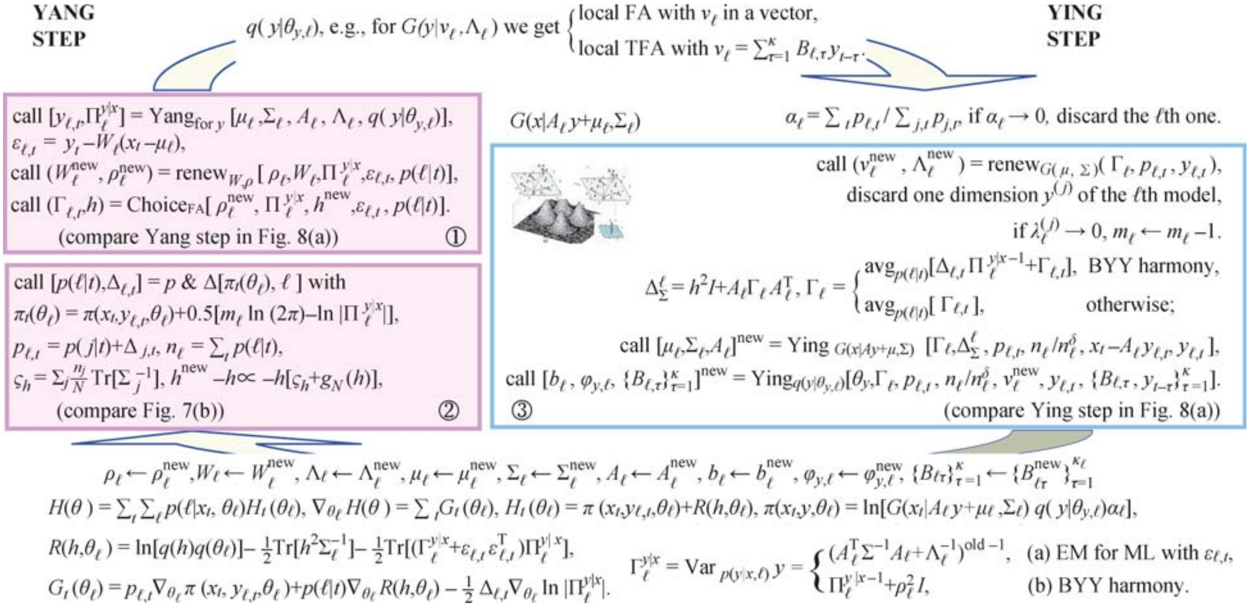

YANG STEP
$q(y|\theta_{y,\ell})$, e.g., for $G(y|\nu_\ell, \Lambda_\ell)$ we get $\begin{cases} \text{local FA with } \nu_\ell \text{ in a vector,} \\ \text{local TFA with } \nu_\ell = \sum_{\tau=1}^{\kappa} B_{\ell,\tau} y_{t-\tau}. \end{cases}$
YING STEP
call $[y_{\ell,t}, \Pi_\ell^{y|x}] = \mathrm{Yang}_{\mathrm{for}\, y}[\mu_\ell, \Sigma_\ell, A_\ell, \Lambda_\ell, q(y|\theta_{y,\ell})]$,
$\varepsilon_{\ell,t} = y_t - W_\ell(x_t - \mu_\ell)$,
call $(W_\ell^{\mathrm{new}}, \rho_\ell^{\mathrm{new}}) = \mathrm{renew}_{W,\rho}[\rho_\ell, W_\ell, \Pi_\ell^{y|x}, \varepsilon_{\ell,t}, p(\ell|t)]$,
call $(\Gamma_{\ell,t}, h) = \mathrm{Choice}_{\mathrm{FA}}[\rho_\ell^{\mathrm{new}}, \Pi_\ell^{y|x}, h^{\mathrm{new}}, \varepsilon_{\ell,t}, p(\ell|t)]$.
(compare Yang step in Fig. 8(a)) ①
call $[p(\ell|t), \Delta_{\ell,t}] = p \,\&\, \Delta[\pi_t(\theta_\ell), \ell]$ with
$\pi_t(\theta_\ell) = \pi(x_t, y_{\ell,t}, \theta_\ell) + 0.5[m_\ell \ln(2\pi) - \ln|\Pi_\ell^{y|x}|]$,
$p_{\ell,t} = p(j|t) + \Delta_{j,t}, n_\ell = \sum_t p(\ell|t)$,
$\varsigma_h = \sum_j \frac{n_j}{N} \mathrm{Tr}[\Sigma_j^{-1}], h^{\mathrm{new}} - h \propto -h[\varsigma_h + g_N(h)]$,
(compare Fig. 7(b)) ②
$G(x|A_\ell y + \mu_\ell, \Sigma_\ell)$
$\alpha_\ell = \sum_t p_{\ell,t} / \sum_{j,t} p_{j,t}$, if $\alpha_\ell \to 0$, discard the ℓth one.
call $(\nu_\ell^{\mathrm{new}}, \Lambda_\ell^{\mathrm{new}}) = \mathrm{renew}_{G(\mu,\Sigma)}(\Gamma_\ell, p_{\ell,t}, y_{\ell,t})$,
discard one dimension $y^{(j)}$ of the ℓth model,
if $\lambda_\ell^{(j)} \to 0$, $m_\ell \leftarrow m_\ell - 1$.
$\Delta_\Sigma^\ell = h^2 I + A_\ell \Gamma_\ell A_\ell^{\mathrm{T}}, \Gamma_\ell = \begin{cases} \mathrm{avg}_{p(\ell|t)}[\Delta_{\ell,t} \Pi_\ell^{y|x-1} + \Gamma_{\ell,t}], & \text{BYY harmony,} \\ \mathrm{avg}_{p(\ell|t)}[\Gamma_{\ell,t}], & \text{otherwise;} \end{cases}$
call $[\mu_\ell, \Sigma_\ell, A_\ell]^{\mathrm{new}} = \mathrm{Ying}_{G(x|Ay+\mu,\Sigma)}[\Gamma_\ell, \Delta_\Sigma^\ell, p_{\ell,t}, n_\ell/n_\ell^\delta, x_t - A_\ell y_{\ell,t}, y_{\ell,t}]$,
call $[b_\ell, \varphi_{y,\ell}, \{B_{\ell,\tau}\}_{\tau=1}^{\kappa}]^{\mathrm{new}} = \mathrm{Ying}_{q(y|\theta_{y,\ell})}[\theta_y, \Gamma_\ell, p_{\ell,t}, n_\ell/n_\ell^\delta, \nu_\ell^{\mathrm{new}}, y_{\ell,t}, \{B_{\ell,\tau}, y_{t-\tau}\}_{\tau=1}^{\kappa}]$.
③ (compare Ying step in Fig. 8(a))
$\rho_\ell \leftarrow \rho_\ell^{\mathrm{new}}, W_\ell \leftarrow W_\ell^{\mathrm{new}}, \Lambda_\ell \leftarrow \Lambda_\ell^{\mathrm{new}}, \mu_\ell \leftarrow \mu_\ell^{\mathrm{new}}, \Sigma_\ell \leftarrow \Sigma_\ell^{\mathrm{new}}, A_\ell \leftarrow A_\ell^{\mathrm{new}}, b_\ell \leftarrow b_\ell^{\mathrm{new}}, \varphi_{y,\ell} \leftarrow \varphi_{y,\ell}^{\mathrm{new}}, \{B_{\ell\tau}\}_{\tau=1}^{\kappa} \leftarrow \{B_{\ell\tau}^{\mathrm{new}}\}_{\tau=1}^{\kappa_\ell}$
$H(\theta) = \sum_t \sum_\ell p(\ell|x_t, \theta_\ell) H_t(\theta_\ell), \nabla_{\theta_\ell} H(\theta) = \sum_t G_t(\theta_\ell), H_t(\theta_\ell) = \pi(x_t, y_{\ell,t}, \theta_\ell) + R(h, \theta_\ell), \pi(x_t, y, \theta_\ell) = \ln[G(x_t|A_\ell y + \mu_\ell, \Sigma_\ell) q(y|\theta_{y,\ell}) \alpha_\ell]$,
$R(h, \theta_\ell) = \ln[q(h) q(\theta_\ell)] - \frac{1}{2}\mathrm{Tr}[h^2 \Sigma_\ell^{-1}] - \frac{1}{2}\mathrm{Tr}[(\Gamma_\ell^{y|x} + \varepsilon_{\ell,t}\varepsilon_{\ell,t}^{\mathrm{T}})\Pi_\ell^{y|x}]$,
$G_t(\theta_\ell) = p_{\ell,t} \nabla_{\theta_\ell} \pi(x_t, y_{\ell,t}, \theta_\ell) + p(\ell|t) \nabla_{\theta_\ell} R(h, \theta_\ell) - \frac{1}{2} \Delta_{\ell,t} \nabla_{\theta_\ell} \ln|\Pi_\ell^{y|x}|$.
$\Gamma_\ell^{y|x} = \mathrm{Var}_{p(y|x,\ell)}\, y = \begin{cases} (A_\ell^{\mathrm{T}} \Sigma^{-1} A_\ell + \Lambda_\ell^{-1})^{\mathrm{old}\,-1}, & \text{(a) EM for ML with } \varepsilon_{\ell,t}, \\ \Pi_\ell^{y|x-1} + \rho_\ell^2 I, & \text{(b) BYY harmony.} \end{cases}$

Fig. 9 A unified Ying-Yang alternation procedure for typical learning algorithms on multi-location extensions

also be obtained from a perspective of GM in Fig. 1(b) by getting each Gaussian in a constrained structure $G(x|\mu_j, A_j\Lambda_jA_j + \Sigma_j)$ via a factor analysis. Particularly, when A_j degenerates into a vector a_j, each FA becomes a hyperplane with its normal vector a_j, which leads to the task of line detection illustrated in Fig. 6(a). That is, this special case of LFA can be regarded as an improvement of RHT for line detection from noisy data.

In implementation, we can combine the learning algorithms in Fig. 7(b) and Fig. 8(a) into the unified Ying-Yang alternation procedure shown in Fig. 9. Specifically, the Box-① of Yang step and the Box-③ of Ying step updates each individual model in a way similar to those in Fig. 8(a), while the first line of Ying step takes the roles same as the first line of Ying step in Fig. 8(a).

Moreover, the Box-② takes the roles similar to the Box-① in Fig. 7(b), except that $\pi_t(\theta_l)$ in the Box-② becomes different from the one in Fig. 7(b) by an additional term $\frac{1}{2}[\cdots]$. Also, $G_t(\theta_l)$ gets an additional third term at the bottom of Fig. 9. The extra terms come from equivalently re-expressing $\ln\int G(x_t|A_ly + \mu_l, \Sigma_l)q(y|\theta_{y,l,})\mathrm{d}y$ in term of $\ln[G(x_t|A_ly+\mu_l, \Sigma_l)q(y|\theta_{y,l,})]$. Improvement can also be obtained by a counterpart of Eq. (13) for model selection via a two stage implementation:

$$J\left(k, \{m_j\}_{j=1}^k\right) = \frac{1}{2}\sum_{j=1}^k \alpha_j \left\{\ln|\Sigma_j| + h^2\mathrm{Tr}\left[\Sigma_j^{-1}\right] + m_j\right\} - \sum_{j=1}^k \alpha_j \ln\alpha_j + \frac{1}{2}n_f(\Theta) + J_y\left(k, \{m_j\}_{j=1}^k\right),$$
$$J_y\left(k, \{m_j\}_{j=1}^k\right) = \begin{cases} \dfrac{1}{2}\displaystyle\sum_{j=1}^k \alpha_j \{\ln|\Lambda_j| + m_j\ln(2\pi\mathrm{e})\}, & \text{(a) } q\left(y\left|\theta_{y,j}\right.\right) \text{ for a real vector,} \\ -\displaystyle\sum_{j=1}^k \alpha_j \sum_{\ell=1}^{m_j}\left[\nu_j^{(\ell)}\ln\nu_j^{(\ell)} + \left(1-\nu_j^{(\ell)}\right)\ln\left(1-\nu_j^{(\ell)}\right)\right], & \text{(b) } q\left(y\left|\theta_{y,j}\right.\right) = B\left(y\left|\nu_j\right.\right), \end{cases} \tag{19}$$

where $n_f(\Theta) = dk+k-1+0.5d(d+1)k+md+\sum_j[m_j + m_jd - 0.5m_j(m_j+1)]$.

The Box-③ also covers those previously discussed extensions of FA, namely, NFA, binary factor analysis (BFA), TFA.

3.3 Insights from the perspective of A5 circling

From the perspective of three nested levels of A5 circling we expect well balanced circling within each layer and across layers, in accordance to Proposition 4 in Appendix B.

For a circling balance within each layer, the circling flow should be neither too watery nor too full to be jammed at anyone of five actions, from which we further get some insights on whether computational burdens are appropriately allocated among the actions A-2, A-3 and A-4.

Taking the L$^{1\mathrm{st}}$-A5 circling in Fig. 1(b) and the algorithms in Fig. 7 as an example, the A-2 action by a WTA-BYY considers merely one candidate j^* by $p_{j,t} = \delta_{jj^*}$, which reduces the computing burden of A-3 and A-4 to a least. However, it also makes the L$^{1\mathrm{st}}$-A5 circling lost balance at A-3 and A-4 (too weak) and thus gets learning easy to be trapped in a local optimum. On the other hand, when k is too large, the computing burden of A-3 and A-4 become too loaded to be jammed. Therefore, the BYY harmony learning uses apex approximation to consider a subset $C_\kappa(x_t)$ that consists of κ climax neighbors (see Eq. (38)) for an appropriate balance.

For the FA problem in Fig. 1(c) and the algorithms in Fig. 8, similarly we observe that either of directly setting $p(y|x) = G(y|y_{t,}\Pi^{y|x\,-1})$ or $p(y|x) = \delta(y - y_t)$ leads straightly Γ_t=0, which makes the flow from the A-2 action to the A-4 action become too watery and thus vulnerable to be trapped at a point y_t that could be far from the global optimal one. In contrast, the BYY harmony learning tackles this problem by gradually driving $\Gamma_t = \rho^2I + \varepsilon_t\varepsilon_t^{\mathrm{T}} \to 0$.

Extending the FA problem beyond that a vector y comes from Gaussian, the flow from the A-2 action to the A-4 action generally becomes too loaded because both $p(y|x)$ and its apex become analytically intractable. One way to handle is seeking a $p(y|x)$ to analytically approximate a Gaussian as if it is in the case of Fig. 1(c). The other way is iteratively seeking the apex of $p(y|x)$ by an optimization algorithm and then estimating $p(y|x)$ on an apex-zone $C_t^\kappa(x)$, e.g., for a BFA with $B(y|\nu)$, we get y_t from $\max_y \pi(y, \theta_y)$ in Box-① of Fig. 8(b) and then considers maximizing $H(\theta)$ with

$$\begin{aligned} H_t(\theta) = &\sum_{y\in C_t^\kappa(x_t)} \ln\left[G\left(x_t\left|Ay+\mu, \Sigma\right.\right)B\left(y|\nu\right)\right] \\ &+ \ln\left[q(h)q(\theta)\right] - \frac{1}{2}\mathrm{Tr}\left[h^2\Sigma^{-1}\right], \\ C_t^\kappa(x_t) = &\left\{y : y \text{ differs from } y_t \text{ by one bit}\right\}. \end{aligned} \tag{20}$$

In addition to seeking a circling balance within each layer, we also seek a circling balance across layers. As discussed in Sect. 2.3, the bottleneck of $\max_\Theta F(\Theta)$

can be tackled by a series of smaller circling $\Theta^{\rm new} = \Theta^{\rm old} + \eta g_F(\Theta)$, that is, a jamming in an upper layer circling can be resolved by a series of smaller circles in its immediate lower layer. Still, it needs a step size η. If η is too small, there will be a huge number of small circling that takes an impractical long period to implement. If η is too large, the iterating process will become unstable and diverge. Interestingly, we observe that the problem of η has been avoided by the EM algorithm with $\Theta^{\rm old}$ updated into $\Theta^{\rm new}$ via a series of smaller circles in a L$^{\rm 1st}$-A5 circling for the first level inverse problem $x_t \to j_t$ per sample x_t, which provides the L$^{\rm 2nd}$-A5 circling more than just the value of $F(\Theta)$ and $g_F(\Theta)$, but also $p_{j,t}$ that makes a move from $\Theta^{\rm old}$ to $\Theta^{\rm new}$ without using η. Also as discussed in Sect. 2.3, the bottleneck of getting $\boldsymbol{k}^* = \arg\min_{\boldsymbol{k}} J(\boldsymbol{k})$ at the A-4 action of the L$^{\rm 3rd}$-A5 circling is tackled by a series of L$^{\rm 2nd}$-A5 circling for updating Θ, via either a two stage implementation or automatic model selection with one force of driving $\psi(\theta_{\boldsymbol{k}}) \to 0$ by Eq. (4). By the EM algorithm, the driving force $\psi(\theta_{\boldsymbol{k}}) \to 0$ comes from an external prior $q(\Theta|\Xi)$ within the same level L$^{\rm 2nd}$ during implementation of ML learning. In contrast, the BYY harmony learning gets this driving force $\psi(\theta_{\boldsymbol{k}}) \to 0$ not only from $q(\Theta|\Xi)$ but also from Eq. (5) intrinsically via a series of L$^{\rm 2nd}$-A5 circling in help of the L$^{\rm 1st}$-A5 circling.

The last but not the least, it is also insightful to observe different ways for a series of L$^{\rm 1st}$-A5 circling to be nested within an L$^{\rm 2nd}$-A5 circling. At one extreme case, an L$^{\rm 2nd}$-A5 circling is started after completing a series of L$^{\rm 1st}$-A5 circling that goes through every sample in X_N, which is usually called a batch learning or learning in a batch. At the other extreme case, an L$^{\rm 2nd}$-A5 circling is simultaneously made with a series of L$^{\rm 2nd}$-A5 circling, in a way that Θ is updated in a small step as an L$^{\rm 1st}$-A5 circling implemented per sample x_t, which is usually called adaptive learning. Moreover, there could be many variants between the two extreme cases.

4 Bayesian Ying-Yang learning

4.1 Ying-Yang best harmony principle

The ancient Yin-Yang or Ying-Yang harmony philosophy came from more than 3000 years ago [53,87–89]. As restated in Appendix B, it can be regarded as a Meta theory for a system that survives or interacts with its environment or world. This system consists of two different but complement parts, one is called Yang that samples and gathers from its external world or called Yang domain (e.g., we get X_N in Fig. 2(d)) and transforms them via a Yang pathway (e.g., those bottom-up arrows in Fig. 2) into an inner domain; while the other is Ying that consists of not only the inner domain (called Ying domain, e.g., $q(R)$ in Fig. 2(d)) that accumulates, integrates, digests, extracts, and condenses whatever came from Yang, but also a Ying pathway (e.g., those top-down arrows in Fig. 2) from which either the most appropriate one or ones chosen from the Ying domain will generate the corresponding best reconstructions back to the Yang domain. Three inverse problems in Sect. 1.2 are summarized by Eq. (1) from this Ying-Yang system perspective, as shown in Fig. 2(d).

Ying is primary and its structure is designed according to tasks of the system, while Yang is secondary and its structure is designed from the Ying structure according to a variety preservation principle for a Ying-Yang balance. With inputs from and outputs to its world, a Ying-Yang system behaves under a best harmony principle. On the system by Eq. (1), it means that Ying and Yang both adapt each other to reach a best agreement in a most tacit way (consuming a least amount of information communication) or that Ying and Yang become a best matching pair in a most compact form with a least complexity. This principle is mathematically implemented by maximizing a harmony functional by Eq. (2), which is one typical example of the following harmony functional in an even general form.

Considering two systems described respectively by P, Q that are both σ-finite measures on the same measure space (X, Σ), and let μ be another reference σ-finite measure that describes a volume or capacity (e.g., a Lebesgue measure) about this space (X, Σ) itself, we observe how the nature of each system changes on every differential piece of this space in help of the Radon-Nikodym derivative [90] $\mathrm{d}P/\mathrm{d}\mu, \mathrm{d}Q/\mathrm{d}\mu$, each of which represents the density of each measure on that piece. A local harmoniousness between two systems may be described by a product $f(\mathrm{d}Q/\mathrm{d}\mu)\mathrm{d}P/\mathrm{d}\mu$ with a scale adjustment by $f(r)$ that monotonically increases with r, e.g., $f(r) = \ln r$. As a whole, we have the following harmony functional for the activities of two systems:

$$H_\mu(P\|Q) = \int_X \frac{\mathrm{d}P}{\mathrm{d}\mu} f\left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)\mathrm{d}\mu = \int_X f\left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)\mathrm{d}P, \tag{21}$$

which becomes large as two measures on every differential piece are coherently large, for which both $\mathrm{d}P$ and $\mathrm{d}Q$ are large while the volume of differential piece is small, i.e., two measures are concentrated on small volume pieces. One intuitive example is considering to distribute two types of resources over the space (X, Σ) according to $\mathrm{d}P/\mathrm{d}\mu, \mathrm{d}Q/\mathrm{d}\mu$, subject to those existing constraints of each system itself. In order to get a joint task well done, we maximize $H_\mu(P||Q)$ to make two resources coherently distributed and concentrated on wherever really necessary for a best use of resources.

This triple-relation among $\mathrm{d}P$, $\mathrm{d}Q$, and $\mathrm{d}\mu$ includes two typical bi-relations as its degenerated cases. One is

$\mathrm{d}Q = \mathrm{d}P$ that

$$H_\mu\left(P\,\|P\right) = \int_X \frac{\mathrm{d}P}{\mathrm{d}\mu} f\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)\mathrm{d}\mu = \int_X f\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)\mathrm{d}P, \tag{22}$$

which becomes large when those large values of the measure P are concentrated on small volume pieces. That is, $H_\mu(P||P)$ describes a compactness of the measure density of P. The other bi-relation is obtained by letting $\mathrm{d}\mu = \mathrm{d}P$ that leads to

$$H_P\left(P\|Q\right) = \int_X f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P, \tag{23}$$

which describes a closeness between the two systems. For a concave function $f(r)$ with $f(1) = 0$, $-H_P(P||Q)$ becomes actually the f-divergence [91].

Moreover, when μ is a Lebesgue measure, and P, Q are probability measures on the probability space (X, Σ) (e.g., the resources we deal with are chances or probabilities), maximizing $H_\mu(P||Q)$ by Eq. (21) makes $\mathrm{d}Q/\mathrm{d}\mu$ match $\mathrm{d}P/\mathrm{d}\mu$, i.e., each differential volume (or an event to occur) gets a chance $\mathrm{d}Q$ from one system to be same as $\mathrm{d}P$ from the other system, which tends to $H_\mu(P||P)$ by Eq. (22). Further maximizing $H_\mu(P||P)$ makes two systems $\mathrm{d}P/\mathrm{d}\mu$, $\mathrm{d}Q/\mathrm{d}\mu$ located as compacted as possible subject to those existing constraints on P, Q. That is, each differential volume (or an event to occur) gets either no chance or a chance as higher as possible from both systems. Two systems best match in term of not only equal distributions but also as higher as possible chances concentrated wherever there is a match while as less as possible chances wasted in other places. In other words, two systems tend towards one agreement as deterministic or least-uncertain as possible. In a contrast, maximizing $H_P(P||Q)$ by Eq. (23) makes $\mathrm{d}P/\mathrm{d}\mu$, $\mathrm{d}Q/\mathrm{d}\mu$ match in distribution, but not be necessarily pushed to a compact form, since it has no consideration of each differential volume.

Furthermore, when $f(r) = \ln r$, $p = \mathrm{d}P/\mathrm{d}\mu$, $q = \mathrm{d}Q/\mathrm{d}\mu$, $H_\mu(P||Q)$ by Eq. (21), $H_\mu(P||P)$ by Eq. (22), and $H_P(P||Q)$ by Eq. (23) become $H(p||q)$, $H(p||p)$, and $\mathrm{KL}(p||q)$ respectively in a relationship as follows:

$$\begin{aligned} &H\left(p\,\|q\right) = \int p\left(X\right)\ln q\left(X\right)\mathrm{d}X = H\left(p\,\|p\right) - \mathrm{KL}\left(p\,\|q\right), \\ \text{or}\quad &\mathrm{KL}\left(p\,\|q\right) = H\left(p\,\|p\right) - H\left(p\,\|q\right), \end{aligned} \tag{24}$$

where $H(p||p)$ is a negative entropy, and $\mathrm{KL}(p||q)$ is the Kullback-Leibler divergence. Echoing the beginning of this subsection, Ying and Yang seeks a best agreement via minimizing $\mathrm{KL}(p||q)$ in a most tacit manner via minimizing the information $-H(p||p)$ that is transferred by Yang. Alternatively, this tacit agreement may also be observed directly from maximizing $H(p||q)$, which has a separable nature that $\max_q H(p||q)$ for a fixed p leads to $q = p$ for best matching and that $\max_p H(p||q)$ for a fixed q leads to $p(x) = \delta(x - c)$ for least complexity. This best matching nature has been widely used in those best data matching approaches via the special cases that p is fixed at given data, i.e., $p(X) = \delta(X - X_N)$. In these cases, we simply have $H(p||q) = \ln q(X_N)$, which leads to the marginal Bayes learning (e.g., BIC, MDL) for $q(X_N) = q(X_N|\boldsymbol{k})$ and the ML learning, as discussed in Fig. 4(d).

Also, one must not confuse $H(p||q)$ by Eq. (24) with an information theoretic term called cross entropy though it has apparently a same expression. Actually, the nature that $\max_p H(p||q)$ for a fixed q leads to $p(x) = \delta(x - c)$ has not been involved in the cross entropy related studies because it was regarded a useless degenerated case. The name of cross entropy measure was originated from the Kullback and Leibler [92] by $\mathrm{KL}(p||q)$ for the "discrepancy" between p and q. In the literature, it has been used in two ways. One is widely studied in the literature of signal processing and information theory under the name of minimum cross entropy (MCE) [93] for estimating p subject to a set of constraints with a given reference distribution q, which includes the maximum entropy approach as a special case [94,95]. This way is different from $H(p||q)$ we studied here. The other way is studied in the literature of statistical learning for estimating q by minimizing $\mathrm{KL}(p||q) = H(p||p) - H(p||q)$ given a reference distribution p which is equivalent to maximizing $H(p||q)$ to getting q (since p is fixed), i.e., a special case of the best data matching studies.

Promisingly, the above apparent useless singular nature becomes useful and important when p, q are given by a Ying-Yang pair by Eq. (1), which leads to the BYY harmony learning by Eq. (2). Because $p = p(R|X)p(X)$ includes $p(X) = p(X|X_N, h)$ given by Eq. (6), $\max_p H(p||q)$ for a fixed q can not push $p(R|X)p(X)$ simply to one extreme δ format, but instead to push $p(R|X)$ into a most compact form under the constraint of $p(X) = p(X|X_N, h)$ and also certain structure of $p(R|X)$ (if any). On the other hand, $\max_q H(p||q)$ for a fixed p forces the Ying machine $q(X|R)q(R)$ to best match $p(R|X)p(X)$ and thus become more compact too. Due to a finite size N and other existing constraints (if any), the limit $q(X|R)q(R) = p(R|X)p(X)$ may not be really reached. Still we consider a trend towards this equality by which $H(p||q)$ in Eq. (2) becomes the negative entropy, and its further maximization will minimize the system complexity, which makes the Ying-Yang pair in a least complexity.

Shown in Fig. 10 is a simple geometrical illustration of the BYY harmony learning for an intuitive insight. The Ying machine is a parametric manifold designed according to the nature of learning tasks subject to a least redundancy principle (see the next subsection),while the manifold of Yang machine consists of two submanifolds, one is almost specified by X_N and may vary subject to merely one free parameter $h > 0$, and the other for $p(R|X, \Theta_q, \Theta_p)$ comes from the Ying manifold subject to a variety preservation principle (see the next subsection), as shown in Fig. 10(a). Instead of that $p(R|X, \Theta_q, \Theta_p)$ is obtained from a fixed instance of $q(X|R)q(R)$, the structure of $p(R|X, \Theta_q, \Theta_p)$ is a functional of $q(X|R)q(R)$, not only possessing its own variables Θ_p but also sharing the common variables Θ_q of $q(X|R)q(R)$.

The learning process consists of two steps in alternation. As shown in Fig. 10(b), Yang step focuses on a peak zone of $p(R|X, \Theta_q, \Theta_p)$ centered around newly estimated Y^*, L^*, Θ_p^* as functions of Θ_q, while the Ying step adjusts Θ_q^* to project the Ying manifold towards the peak zone of the Yang manifold, and accordingly the peak zone also changes as a function of Θ_q too, during which both manifolds shrink due to a least complexity nature. This learning procedure has a quite unique feature. As illustrated in Fig. 5(b), the manifold shrinking will cause $H(p||q)$ tends infinity, i.e., the process becomes diverging, which was regarded as a bad thing in a conventional sense. Here, it however acts as signals that the related dimensions should be discarded as the corresponding extra scale representative (SR) parameters approach zeros.

The nature of Ying-Yang best harmony by Eq. (2) is also observed from the relation of its gradient flow $\nabla_\varphi H(p||q)$ to the differential flow $\int p(R|X_N) \cdot \nabla_\varphi \ln[q(X_N|R)q(R)]\mathrm{d}R$, while the latter actually indicates the updating flow of the M-step in the EM algorithm for the maximum likelihood learning and Bayesian learning, where and hereafter in this paper, we use $\nabla_\varphi f$ to denote the gradient of f with respect to φ, a general notation that could be flexibly Θ or one of its subsets, and we simply have $\nabla_\varphi f{=}0$ when f is irrelevant to φ. Further discussions on this general formulation are referred to those made after Eq. (29) in Ref. [2] and after Eq. (15) in Ref. [50]. For simplicity, we consider the cases that there is no prior $q(\Theta|\Xi)$ and that $p(\Theta|X)$ is free of any structure and thus determined by maximizing $H(p||q)$ in Eq. (2). In these cases, parameter learning is made by $\max_{\Theta,h} H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$ as shown in Fig. 5(b).

Following the variety preservation principle to be introduced in Sect. 4.2 and considering the Yang path by Eq. (27) at $D_p(X) = D_q(X)$, i.e., the following Bayesian structure $p(Y|X, \theta_{y|x}) = q(Y|X, \theta)$, we have

$$\begin{aligned}\nabla_\varphi H\left(p\|q, \Theta, \boldsymbol{k}, \Xi\right) &= \int p\left(Y|X, \theta_{y|x}\right)\left[1 + \Delta\pi\left(X, Y\right)\right]\nabla_\varphi \pi(X, Y) p\left(X\left|X_N, h\right.\right)\mathrm{d}X\mathrm{d}Y,\\ p\left(Y|X, \theta_{y|x}\right) &= q(Y|X, \theta), \quad q(Y|X, \theta) = q\left(X|Y, \theta_{x|y}\right) q\left(Y\left|\theta_y\right.\right)/q\left(X|\Theta\right),\\ \Delta\pi(X, Y) &= \pi(X, Y) - \int p\left(Y|X, \theta_{y|x}\right)\pi\left(X, Y\right)\mathrm{d}Y, \quad \pi\left(X, Y\right) = \ln\left[q\left(X|Y, \theta_{x|y}\right) q\left(Y\left|\theta_y\right.\right)\right],\end{aligned} \tag{25}$$

where $\pi(X, Y)$ describes the fitness of an inner representation Y on the observation X, and $\Delta\pi(X, Y)$ indicates whether this Y fits X better than the average of all the possible choices of Y. Without losing generality, we simply let $p(X|X_N, h)$ given by Eq. (6) with $h = 0$ to further examine different scenarios of $\Delta\pi(X, Y)$.

Let $\Delta\pi(X, Y) = 0, \nabla_\varphi H(p||q, \Theta, \boldsymbol{k}, \Xi)$ actually becomes the updating flow of the M-step in the EM algorithm for the ML learning [20]. Usually $\Delta\pi(X, Y) \neq 0$, i.e., the gradient flow $\nabla_\varphi \pi(X, Y)$ under all possible choices of Y is integrated via a weighting not just by $p(Y|X, \theta_{y|x})$ but also by a modification of a relative fitness $1 + \Delta\pi(X, Y)$. If $\Delta\pi(X, Y) > 0$, updating goes along the same direction of the EM learning with an increased strength. If $0 > \Delta\pi(X, Y) > -1$, i.e., the fitness is worse than the average and the current Y is doubtful, updating still goes along the same direction of the EM learning but with a reduced strength. When

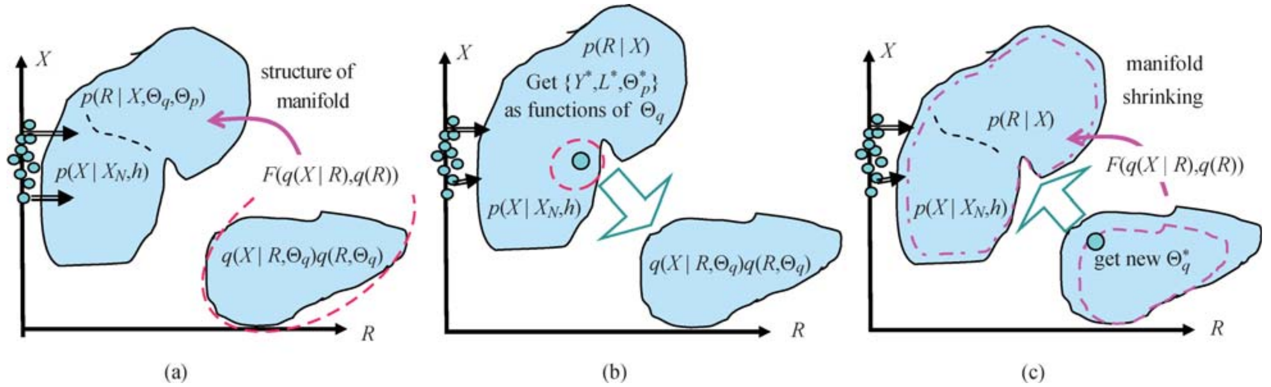


Fig. 10 A geometrical illustration of Ying-Yang harmony. (a) Manifold of Yang machine consists of two submanifolds, one specified by X_N and the other being a functional of the Ying manifold subject to a variety preservation; (b) Yang step focuses on a peak zone centered around newly estimated Y^*, L^*, Θ_p^* as functions of Θ_q; (c) Ying step updates Θ_q^* to be projected towards the peak zone of Yang manifold, and both Yang manifold and Ying manifold shrink as a function of Θ_q^*

$-1 > \Delta\pi(X,Y)$, updating reverses the direction of the EM learning and actually becomes de-learning. In other words, the BYY harmony learning shares a mechanism similar to RPCL learning [64–66].

Also, we may observe the model selection nature from the roles of $q(\Theta|\Xi)$ and $q(Y|\Theta)$, as previously discussed in Sects. 2.1 and 2.2. First, we consider $R = \{\Theta\}$ and $p(X) = \delta(X - X_N)$, by which Eq. (2) becomes $\int p(\Theta|X_N)\ln[q(X_N|\Theta)q(\Theta)]\mathrm{d}\Theta$. Its maximization with respect to a structural free Yang $p(\Theta|X_N)$ becomes $p(\Theta|X_N) = \delta(\Theta - \Theta^*)$ and $\Theta^* = \arg\max_\Theta \ln[q(X_N|\Theta)q(\Theta)]$, i.e., we are lead to those studies of Bayes learning, with its model selection nature discussed in Sect. 2.1 and especially around Fig. 4(c). To be further discussed after Eq. (33) in Sect. 4.2, $\ln[q(X_N|\Theta^*)q(\Theta^*)]$ becomes coincided with the normalized maximum likelihood (NML) used in the MDL encoding if we let $p(\Theta|X_N) = \delta(\Theta - \Theta^*)$ with $\Theta^* = \arg\max_\Theta \ln q(X_N|\Theta)$ and $q(\Theta)$ given by a so-called IBC priori. Moreover, considering $p(\Theta|X_N)$ by a Gaussian with its mean Θ^* and its covariance matrix given by the inverted Fisher information matrix of $\ln[q(X_N|\Theta)q(\Theta)]$, we are also lead to $H(p||q,\boldsymbol{k},\Xi)$ in Fig. 5(b) with

$$\begin{aligned}H(p\|q,\Theta^*,h,\boldsymbol{k},\Xi) = {} & \ln q(X_N|\Theta^*) - \frac{1}{2}d_k(\Xi|\Theta^*)\\ & + \ln[q(h)q(\Theta^*)]\\ & - \frac{1}{2}h^2\mathrm{Tr}[\Sigma_L(X_N)],\end{aligned} \tag{26}$$

where the details are delayed to Eq. (36) and $\Sigma_L(X_N)$ to Eq. (37) (see Sect. 4.3), which extends Bayes learning. Second, we consider $R = \{\Theta, Y\}$ and $p(X) = \delta(X - X_N)$. As already introduced in Sect. 2.2 and Fig. 5, the mechanism of pushing the Ying-Yang pair into a most compact form can be also observed from $q(R) = q(Y|\Theta)q(\Theta|\Xi)$ that puts $q(Y|\Theta)$ and $q(\Theta|\Xi)$ in the positions of equal importance to be considered for model selection, i.e., the BYY best harmony includes the maximization by Eq. (5) as a part, which leads to those three favorable features discussed at the end of Sect. 2.2.

Though both $H(p||q)$ and $\mathrm{KL}(p||q)$ can be regarded as special cases of the general form of harmony functional by Eq. (21), it follows from Eq. (24) that $\mathrm{KL}(p||q)$ is not a special case of $H(p||q)$. Instead, $H(p||q)$ and $\mathrm{KL}(p||q)$ share some common special cases and also differ in that minimizing $\mathrm{KL}(p||q)$ only covers a partial purpose of maximizing $H(p||q)$. With p, q given by Eq. (1), it follows from Eq. (24) that BYY best matching by $\mathrm{KL}(p||q)$ leads to the early studies in 1995 made under the name of Bayesian Kullback Ying-Yang (BKYY) learning in Ref. [47], which provides a framework that unifies a number of typical learning methods. Also, Eq. (24) of Ref. [47] actually initialized an effort on the BYY harmony learning via getting a criterion for selecting the cluster number. During 1996–1999, this criterion evolved into a model selection criterion $J_2(k)$ for Gaussian mixture, factor analysis, and other learning models, and finally reached the generic formulation $H(p||q)$ for Ying-Yang best harmony. Details are referred to Ref. [96], especially its Sect. II(B) and the footnote on its first page.

In a summary, the novelty and favorable natures of BYY best harmony, as well as its relation to BYY best matching, can be observed from the following aspects:

- As stated at the beginning of Sect. 4.1, Ying-Yang best harmony means that both adapt each other to reach a best match via a least amount of information communication, that is, maximizing $H(p||q)$ minimizes $\mathrm{KL}(p||q)$ in a most tacit manner via minimizing the information $-H(p||p)$ that is transferred by Yang.
- As discussed after Eq. (24), Ying-Yang best harmony means that Ying-Yang becomes a best matching pair in a most compact form with a least complexity, observable from the nature that $\max_q H(p||q)$ for a fixed p leads to $q = p$ for best matching and $\max_p H(p||q)$ for a fixed q leads to $p(x) = \delta(x - c)$ for least complexity. This least complexity is regarded as useless singular case in the studies of cross entropy but becomes a favorable nature in a BYY system.
- From the perspective of Radon-Nikodym derivative, the harmony functional measures a triple-relation among $\mathrm{d}P$, $\mathrm{d}Q$, and $\mathrm{d}\mu$, while the KL-divergence is its degenerated case at $\mathrm{d}\mu = \mathrm{d}P$ for measuring a bi-relation, see Eqs. (21)–(23).
- From Eq. (25), we observe how the updating flow of the M-step in the EM algorithm for the maximum likelihood learning and Bayesian learning are modified into the gradient flow $\nabla_\varphi H(p||q)$ with a mechanism similar to RPCL learning.
- As illustrated in Fig. 10, the manifold of Yang machine is a functional of the Ying manifold subject to a variety preservation principle, and the manifolds shrink during BYY harmony learning, resulting in a least complexity nature.
- As stated in Sect. 2.2 and Fig. 5(b), the BYY harmony learning puts $q(Y|\Theta)$ and $q(\Theta|\Xi)$ in the positions of equal importance to be considered for model selection, including the maximization by Eq. (5) as a part, which leads to those three favorable features discussed at the end of Sect. 2.2.
- Referred to Fig. A2 and Appendix A, the concepts of maximizing harmony, correlation, and similarity all join together to form one evolution stream originated from the concept of seeking common points or mutual agreement; while the concepts of minimizing divergence, fitting error, least distance form the other evolution stream originated from the concept of least difference. Two steams are equivalent in certain cases but also different in other cases.

- Another information-theoretic perspective of BYY harmony learning can be found in Sect. II(C), Sect. II(E), and Fig. 3 of Ref. [49], which provides a three level encoding scheme for optimal communication, being different from both the conventional MDL and the bit back MDL [71].
- As to be introduced in Sect. 5.2, BYY best matching has a bottom-up hierarchical decoupling nature that makes the tasks of learning hierarchical layers decoupled to be made sequentially bottom-up or the tasks of handling latent variables, parameter learning, and model selection decoupled to be made sequentially step by step. Though this nature makes implementation easy, learning within one layer become insensitive to the components of the lower layers. So, it is poor on determining a hierarchical configuration. Also, as discussed in Fig. 5, it makes the role of $q\left(Y, L|\Theta, \boldsymbol{k}\right)$ not considered in the task of model selection. Without such a nature, the BYY best harmony learning makes automatic model selection possible on each layer and each step.

In addition, readers are referred to the roadmap shown in Fig. A2. BYY best matching provides a unified framework for typical existing learning methods, while BYY best harmony provides a framework of new approaches with a favorable new mechanism for model selection.

4.2 Bayesian Ying-Yang system, variety preservation principle, and induced bias cancellation

To set up a BYY system as shown in Fig. 2(d), we encounter tasks similar to those discussed in Sect. 1.1 and Fig. 2. Here, a family of infinite many structures $\{S_{\boldsymbol{k}}\}$ means a family of BYY systems with each $S_{\boldsymbol{k}}$ sharing a same Ying-Yang system architecture but in a different scale. This system architecture is set up via designing the structures of the Ying and Yang machines, which actually consists of designing each of four components in Eq. (2). Usually, $p(X) = p(X|X_N, h)$ is given by Eq. (6), with a unknown smoothing parameter h. What need to be designed actually consists of the structures of $q(X|R), q(R)$ and $p(R|X)$.

The Ying domain $q(R)$ is considered first. The key point is the representation form of R, which is task-dependent, e.g., $R = \{j, \theta\}$ for the GM in Fig. 1(b) and $R = \{y, \theta\}$ for the FA in Fig. 1(c). As shown in Fig. 6(b), we generally have three levels of inner representations. The first level is a STM domain that accommodates either or both of $L = \{j\}$ of labels and $Y = \{y\}$ of vectors as inner representations of samples, the second level is $\Theta = \{\theta\}$ as a collective inner representation of the entire set X_N, and an even higher level consists of either or both of hyperparameters Ξ of $q(\Theta|\Xi)$ and scales k of $S_{\boldsymbol{k}}$. In a compliment of ancient Ying-Yang philosophy (see Propositions 1 and 2 in Appendix B), Ying is primary and be a capacity of accommodating, accumulating, integrating and digesting whatever came from Yang, featured by simplicity (see Ref. [S2] in Fig. B1(b) of Appendix B). Therefore, we prefer that the number of variables and parameters of R should be as less as possible. Already, this is a model selection task that can only be partly considered in the following design guides.

Least redundancy principle Ying machine accommodates inner representations and generates reconstructions to fit observed data via structures with least redundancy. As discussed in Sect. 2.2, a major part of $\boldsymbol{k}$ is the scale set $\boldsymbol{k}_Y$ of the STM domain, which provides a lower bound for the number n_{YL} of parameters in $q(Y, L|\Theta)$. Thus, it is preferred that this number n_{YL} should be as close as possible to $\boldsymbol{k}_Y$. Then, considering both $q(Y, L|\Theta)$ and the structure underlying X_N, we further design the structure of $q(X|Y, L, \Theta)$ that again consists of a set of individual simple structures in a simple combination.

E.g., for the GM in Fig. 1(b) we have the parameters $q(l) = \alpha_l, l = 1, 2, \ldots, k$, with $\boldsymbol{k}_Y = k - 1$ due to $\sum_l \alpha_l = 1$, and for the FA in Fig. 1(c) we have parameters $\lambda_l, l = 1, 2, \ldots, m$, with $\boldsymbol{k}_Y = m$. Generally, when y is a multiple dimensional vector, we may consider a $q(y)$ with mutual independent dimensions. Also, we design $q(X|Y, L, \Theta)$ via Gaussian based linear regressions of X conditional on Y. Readers are referred to Refs. [2,51] for a number of detailed structures. In some situation, it is unnecessary to design $q(Y, L|\Theta)$ and $q(X|Y, L, \Theta)$ separately. Instead, we may design $q(X, R)$, especially $q(X, Y, L|\Theta)$ via an integrated parametric model as a whole but still attempting to follow the above principle, e.g., we may consider $q(X, Y, L|\Theta)$ as a whole to be a σ-finite measure.

Next, the structure of $p(R|X)$ is designed from $q(X|Y, L, \Theta)q(Y, L|\Theta)$. It follows from Propositions 1 and 2 in Appendix B that the circling from Yang to Ying is a converging or digesting process, with enough but not excess inputs from Yang to what needed by Ying, which motivates the following design principle.

Variety preservation (VP) principle (or called uncertainty preservation) Yang machine preserves the dynamism or variety of Ying machine for the inner representation of X. That is, Yang should provide with an enough but not excess variety on those candidate assumptions upon X, for a subsequent processing or decision by Ying machine.

Let D_q to be the domain of R by Ying machine, we consider

$$p\left(R|X\right) \leqslant q\left(R|X\right), \text{ for each } R \in D_\rho^*\left(X\right),$$
$$q\left(R|X\right) = \frac{q\left(X|R\right) q(R)}{\displaystyle\int_{R \in D_\rho^*(X)} q\left(X|R\right) q(R) \mathrm{d}R},$$

$$D^*_\rho(X) = \{R \in D_q : q(R|X) + \rho \geqslant q(R^*|X)\}, \\ \text{for } \rho \geqslant 0, \text{ and } R^* = \arg\max_R [q(X|R)q(R)]. \tag{27}$$

The realm outside of $D^*_\rho(X)$ is relaxed to be free of the constraint by Eq. (27). This $D^*_\rho(X)$ consists of the apex point R^* and its certain neighborhood. Thus, we call $D^*_\rho(X)$ apex zone or climax neighborhood, which forms a dynamic focus for us to avoid an excess variety. Moreover, $D^*_\rho(X)$ is controlled by a scalar ρ to form a spectrum ranging between two extremes:

1) When $\rho = 0$, $D^*_{\rho=0}(X)$ consists of only the apex point R^*. It follows from Eq. (27) that $p(R|X) \leqslant q(R|X) = \delta(R - R^*)$, and maximizing $H(p||q, \boldsymbol{k}, \Xi)$ with respect to $p(R|X)$ further leads to $p(R|X) = \delta(R - R^*)$, i.e., the Yang is given by the maximum a posteriori (MAP) estimate of the Ying machine. This is equivalent to the cases that Yang is free of structure or the corresponding BYY system is said to have a backward architecture [47–49,83].

2) When $\rho > 0$ becomes large enough such that $D^*_\rho(X) = D_q$, we have the strongest preservation $p(R|X) = q(R|X)$ for every R in D_q, i.e., the Yang is given by the Bayesian inverse of the Ying machine, which is a typical example of the cases that the corresponding BYY system is said to have a bidirectional architecture [2,3,47–50,83]. Actually, Eq. (27) let us focus on a particular subset of a bidirectional BYY architecture that compliments to Proposition 2 in Appendix B.

In implementation, we decompose $p(R|X)$ into components as shown in Fig. 6(b), i.e.,

$$p(R|X) = q(\Theta|\Xi)p(Y|X, L)p(L|X).$$

We consider each component separately. First, Eq. (27) directly applies to $L = \{j\}$ of labels, i.e., we have

$$p(L|X) \leqslant q(L|X) \text{ for each } L \in D^*_\delta(X). \tag{28}$$

Second, for $Y = \{y\}$ of vectors and Θ of real parameters, it becomes more convenient to measure the variety or uncertainty by the information in term of the second order statistics

$$\mathrm{Var}_{p(R|X)}[\mathrm{vec}(Y)] \geqslant \mathrm{Var}_{q(R|X)}[\mathrm{vec}(Y)], \\ \mathrm{Var}_{p(R|X)}[\mathrm{vec}(\Theta)] \geqslant \mathrm{Var}_{q(R|X)}[\mathrm{vec}(\Theta)], \tag{29}$$

where $\mathrm{Var}_{p(u)}[u]$ denotes the covariance matrix of a vector u. This can be regarded as an extension of the celebrated Cramér-Rao inequality to the Ying-Yang system [49(p889)]. For two positive definite matrices A and B, $A \geqslant B$ means $u^{\mathrm{T}}Au \geqslant u^{\mathrm{T}}Bu$ for any u. A simple example is given as follows:

$$A = B + \rho^2 \text{ for a diagonal matrix } \rho > 0, \\ \text{because } u^{\mathrm{T}}Au = u^{\mathrm{T}}Bu + u^{\mathrm{T}}\rho^2 u > u^{\mathrm{T}}Bu. \tag{30}$$

We may rather conveniently obtain

$$\mathrm{Var}_{q(R|X)}[\mathrm{vec}(Y)] = \Pi^{Y|X\,-1}, \\ \mathrm{Var}_{p(R|X)}[\mathrm{vec}(\Theta)] = \Pi^{\Theta\,-1}, \\ \Pi^{Y|X} = -\frac{\partial^2 \ln[q(X|R)q(R)]}{\partial\mathrm{vec}[Y]\partial\mathrm{vec}[Y]^{\mathrm{T}}}, \\ \Pi^{\Theta} = -\frac{\partial^2 \ln[q(X|R)q(R)]}{\partial\mathrm{vec}[\Theta]\partial\mathrm{vec}[\Theta]^{\mathrm{T}}}. \tag{31}$$

Finally, the last component to be designed is $q(\Theta|\Xi)$. We partition Θ into two subset Θ^a and Θ^b. The subset Θ^b is supported by a prior distribution $q(\Theta^b|\Xi)$ with unknown hyperparameters Ξ, and also associated with a posteriori conjugate distribution $p(\Theta^b|X, \Xi)$ such that $\int p\left(\Theta^b\middle| X, \Xi\right) \ln q\left(\Theta^b\middle| \Xi\right) \mathrm{d}\Theta^b$ is solved analytically:

Those priors in the literature of Bayesian approaches may be adopted accordingly, e.g., a Laplace prior for a regression or interpolation task [60–62] and a Dirichlet prior for the GM problem shown in Fig. 1(b) [67,68]. If it is difficult to handle such distributions, we may ignore this type of priors in Θ, that is, let Θ^b to be empty.

The subset Θ^a is simply denoted as Θ hereafter whenever there is no confusion. It may have either no priori or an improper priori without hyperparameters Ξ. The case without a priori is regarded as an extreme case of an improper priori with $\ln q(\Theta) = 0$. One widely used improper priori is Jeffrey prior [97]. Moreover, a data sensitive improper priori $q(\Theta)$ was proposed for regularizing the irregularity of a finite size of samples, under the name of data smoothing [73,74,80], especially Eq. (18) in Ref. [73] and Eq. (7) in Ref. [74], and of normalization [80,98–100]. Two key points are given as follows:

1) It has been shown empirically that a good choice of $q(h)$ is simply

$$q(h) \propto \frac{1}{\sum_{t=1}^{N} p_h(u_t)}, \; p_h(u) = \frac{1}{N}\sum_{t=1}^{N} G\left(u|u_t, h^2 I\right). \tag{32}$$

2) The above is just a special case of the following parametric model $q(u|\Theta)$ induced priori:

$$q(\Theta) \propto 1/\sum_{t=1}^{N} q(u_t|\Theta), \tag{33}$$

which came from replacing the integral in $H(p||q)$ by Eq. (24) with a summation over a set of samples via turning $q(u_t|\Theta)$ on these samples into a discrete distribution $q(u_t|\Theta)/Z(\Theta)$ with $Z(\Theta) = \sum_t q(u_t|\Theta)$, e.g., see Eqs. (21) and (22) in Ref. [99]. Alternatively, it has also be re-explained that a finite size of samples makes $Z(\Theta) \neq 1$ that imposes an implicit prior with some bias, and Eq. (33) aims at canceling this induced bias, called induced bias cancelation (IBC). Readers are further referred to Sect. 3.4.3 in Ref. [77] for a recent overview and Sect.

23.7.4 in Ref. [100] for historical remarks. Even interestingly, we consider $H(p||q)$ by Eq. (2) at a degenerated case that $R = \{\Theta\}$, $p(X) = \delta(X - X_N)$ and $p(\Theta|X_N) = \delta(\Theta - \Theta^*)$ with $\Theta^* = \arg\max_\Theta \ln q(X_N|\Theta)$, we have that $H(p||q) = \ln[q(X_N|\Theta^*)q(\Theta^*)] = \ln[q(X_N|\Theta^*)/Z(\Theta^*)]$ becomes coincided with the normalized maximum likelihood (NML) used in the MDL encoding [101].

The above understanding about IBC motivates a generic consideration for getting a priori that consists of two typical roles. One is adding an informative priori $q(\Theta^b|\boldsymbol{k}, \Xi)$, usually with a hyperparameter set Ξ, associated with a posteriori conjugate distribution. The other is canceling out certain bias introduced implicitly by using a parametric model on a small size of samples, e.g., the above $q(h)$ and $q(\Theta|\boldsymbol{k})$ in Eq. (32) and Eq. (33).

4.3 BYY implementation: Apex approximation, primal gradient flow, and alternative maximization

Still, we use the notation $S_{\boldsymbol{k}}(\Theta)$ to refer a BYY system, with its configuration $S_{\boldsymbol{k}}$ specified by designing and with Θ consisting of unknown parameters in both Ying machine and Yang machine. Moreover, the scale $\boldsymbol{k}$ features the complexity of R, including the scale $\boldsymbol{k}_Y$ as a primary part. Both parameter learning for determining Θ and model selection for selecting an appropriate scale $\boldsymbol{k}$ are accomplished via maximizing the harmony functional $H(p||q)$ by Eq. (2). Partitioning Θ into two subsets, with one denoted by Θ^b and the other still denoted by Θ, we consider a prior $q(\Theta)q(\Theta^b|\Xi)$ and correspondingly the posteriori $p\left(\Theta|X, \Xi\right) p\left(\Theta^b|X, \Xi\right)$. The smoothing parameter h may be considered either in the subset Θ^b or the subset Θ. With $p(X|X_N, h)$ given by Eq. (6) and $q(\Theta) = \prod_L q(\Theta_L)$, it follows from Eq. (2) that we rewrite $H(p||q) = H(p||q, \boldsymbol{k}, \Xi)$ as follows:

$$
\begin{aligned}
& H\left(p\|q, \boldsymbol{k}, \Xi\right) \\
& \quad = H_b\left(\Xi, \boldsymbol{k}\right) \\
& \qquad + \int p\left(\Theta|X, \Xi\right) p\left(X|X_N, h\right) H\left(p\|q, \Theta, \boldsymbol{k}, \Xi\right) \mathrm{d}\Theta \mathrm{d}X, \\
& H_b\left(\Xi, \boldsymbol{k}\right) = \int p\left(\Theta^b|X, \Xi\right) \ln q\left(\Theta^b|\Xi\right) \mathrm{d}\Theta^b, \\
& H(p\|q, \Theta, \boldsymbol{k}, \Xi) \\
& \quad = \sum_L \int p(Y|X, L) p(L|X) p(X|X_N, h) \\
& \qquad \times \ln[q(X|Y, L, \Theta_{X|YL}) q(Y, L|\Theta_{YL}) q(\Theta_L)] \mathrm{d}X \mathrm{d}Y.
\end{aligned}
\tag{34}
$$

With h included in Θ, $H(p||q, \Theta, h, \boldsymbol{k}, \Xi)$ given in Fig. 5(b) is a special case of the above $H(p||q, \Theta, \boldsymbol{k}, \Xi)$ after dropping L. Also, an improper priori $q(\Theta)$ can be either no priori (i.e., $\ln q(\Theta) = 0$) or a Jeffreys prior [97]. Also, it can be the IBC priori by Eq. (33), which leads to $Z(\Theta) = -\ln q(\Theta)$ that was studied under the term of normalization regularization [48,49,80,83,96,98–100].

Computing difficulties are encountered for the integral over Θ and the integral over Y for $H(p||q, \Theta, \boldsymbol{k}, \Xi)$. To get rid of it, we consider a Ying-Yang alteration procedure, featured with apex approximation and primal gradient based search.

Apex approximation is made via the following Taylor expansion around u^* up to the second order:

$$
\begin{aligned}
& \int p\left(u\right) Q\left(u\right) \mathrm{d}u \approx Q\left(u^*\right) - \frac{1}{2} \mathrm{Tr}\left[\left(\Gamma^u + \varepsilon_u \varepsilon_u^{\mathrm{T}}\right) \Omega\left(u^*\right)\right], \\
& \qquad u^* = \arg\max_u Q\left(u\right), \quad \varepsilon_u = u^\mu - u^*,
\end{aligned}
\tag{35}
$$

where u^μ, Γ^u are the mean and the covariance of $p(u)$, and $\Omega(u) = -\partial^2 Q(u)/\partial u \partial u^{\mathrm{T}}$. When $Q(u)$ is a quadratic function of u, not only this $\approx$ becomes $=$, but also Eq. (35) applies to the cases that u takes discrete values, with $\Omega(u)$ obtained by regarding that the domain of u is expanded to a real domain.

Using it on the integral over Θ in Eq. (34), we get $H(p||q, \boldsymbol{k}, \Xi)$ as shown in Fig. 5(b) with

$$
\begin{aligned}
& d_k\left(\Xi\left|\Theta^*\right.\right) = \mathrm{Tr}\left[\left\{\Gamma(\Xi) + \varepsilon(\Xi)\varepsilon(\Xi)^{\mathrm{T}}\right\} \Omega\left(\Theta^*\right)\right], \\
& \varepsilon(\Xi) = \mathrm{vec}\left[\Theta^\mu(\Xi) - \Theta^*\right], \ \Theta^\mu(\Xi) = E_{p(\Theta|X,\Xi)}\Theta, \\
& \Gamma(\Xi) = \mathrm{Var}_{p(\Theta|X,\Xi)}\left[\mathrm{vec}\left[\Theta\right]\right], \\
& \Omega(\Theta^*) = -\frac{\partial^2 H\left(p\|q, \Theta, \boldsymbol{k}\right)}{\partial \mathrm{vec}[\Theta] \partial \mathrm{vec}[\Theta]^{\mathrm{T}}}.
\end{aligned}
\tag{36}
$$

That is, we approximate the integral over Θ by its apex zone around Θ^*, which is referred as apex approximation on the support of Θ. Particularly, we have $H_b(\Xi, \boldsymbol{k}) = 0$ if there is no hyper-parameter Ξ and $d_{\boldsymbol{k}}(\Xi|\Theta) = n_f(\Theta)$. It follows from Eq. (31) that $\Gamma(\Xi) = \Omega^{-1}(\Theta^*)$ and $\varepsilon(\Xi) = 0$. Also, it follows from Eq. (29) and Eq. (30) that we may consider $\Gamma(\Xi) = \Omega^{-1}(\Theta^*) + \delta^2$ for a diagonal matrix $\delta > 0$, with or without considering a priori $q(\delta)$ in $q(\Xi)$, and $d_{\boldsymbol{k}}(\Xi|\Theta) = n_f(\Theta) + \mathrm{Tr}[\delta^2 \Omega(\Theta)]$, where δ may also be learned during maximizing $H(p||q, \Theta, \boldsymbol{k}, \Xi)$.

$H(p||q, \Theta, \boldsymbol{k}, \Xi)$ is maximized by a gradient ascending via $\nabla_\Theta H(p||q, \Theta, \boldsymbol{k}, \Xi)$. Two typical roads are featured by choosing the order of handling ∇_Θ and $\sum_L \int$. One is making ∇_Θ first and then approximating the integral over Y and the summation $\sum_L$. The other is removing all the integrals first and then computing $\nabla_\Theta H(p||q, \Theta, \boldsymbol{k}, \Xi)$. In sequel, we focus on the latter.

Again, we remove the integral over Y by apex approximation. Similar to Eq. (35) we also consider a Taylor expansion around u^μ up to the second order and get $\int p(u)Q(u)\mathrm{d}u = Q(u^\mu) - \frac{1}{2}\mathrm{Tr}[\Gamma H_Q(u^\mu)]$, from which we remove the integral over X in Eq. (34). Taking h out of Θ and putting it explicitly in parallel to Θ_L, i.e., $q(\Theta) = q\left(h\left|X_N\right.\right) \prod_L q(\Theta_L)$, some derivation further turns $H(p||q, \Theta, \boldsymbol{k}, \Xi)$ by Eq. (34) into

$$\begin{aligned}
&H\left(p||q,\Theta,\boldsymbol{k},\Xi\right)=\sum_L p\left(L\left|X_N\right.\right)H_L\left(\Theta\right),\\
&H_L\left(\Theta\right)=\pi_L\left(X_N,Y_L^*,\Theta\right)+R_L\left(X_N,Y_L^*,\Theta\right),\\
&\pi_L\left(X,Y,\Theta\right)=\ln\left[q\left(X\left|Y,L,\Theta_{X|YL}\right.\right)q\left(Y,L\left|\Theta_{YL}\right.\right)\right],\\
&R_L\left(X_N,Y_L^*,\Theta\right)\\
&\qquad=\ln\left[q\left(h\left|X_N\right.\right)q\left(\Theta_L\right)\right]-\frac{1}{2}h^2\mathrm{Tr}\left[\Sigma_L\left(X_N\right)\right]\\
&\qquad\quad-\frac{1}{2}\mathrm{Tr}\left[\left\{\Gamma_L^{Y|X}+\varepsilon_L\left(X_N\right)\varepsilon_L^{\mathrm{T}}\left(X_N\right)\right\}\Pi_L^{Y|X}\right],\\
&\varepsilon_L(X)=\mathrm{vec}\left[\mu_L(X)-Y_L^*\right],\ \mu_L(X)=E_{p(Y|X,L)}Y,\\
&Y_L^*=\arg\max_Y\pi_L\left[X_N,Y,\Theta\right],\\
&\Gamma_L^{Y|X}=\mathrm{Var}_{p(Y|X,L)}\left[\mathrm{vec}(Y)\right],\\
&\Sigma_L(X)=-\frac{\partial^2\pi_L\left(X,Y,\Theta\right)}{\partial\mathrm{vec}[X]\partial\mathrm{vec}[X]^{\mathrm{T}}},\\
&\Pi_L^{Y|X}=-\frac{\partial^2\pi_L\left(X,Y,\Theta\right)}{\partial\mathrm{vec}[Y]\partial\mathrm{vec}[Y]^{\mathrm{T}}}.
\end{aligned}\tag{37}$$

Further insights are obtained by taking the cases of Figs. 7–9 as examples. Considering that $X_N=\{x_t\}$ consists of i.i.d. samples and noticing $\ln\Pi_t p_t=\Sigma_t\ln p_t$, we observe that the above $H(p||q,\Theta,\boldsymbol{k},\Xi)$ becomes $H(\theta)$ by Eq. (10) after simply discarding terms related to Y. Considering the structure of $p(L|X)$ by Eq. (28) according to the variety preservation principle, maximizing $H(p||q,\Theta,\boldsymbol{k},\Xi)$ with respect to $p(L|X)$ leads us to

$$\begin{aligned}
&p\left(L\left|X\right.\right)=\chi_\kappa\left(L\right)q\left(L\left|X\right.\right),\\
&q\left(L\left|X\right.\right)=\frac{q\left(X\left|L,\Theta_L\right.\right)q\left(L\right)}{\sum\limits_{L\in C_\kappa(X_N)}q\left(X\left|L,\Theta_L\right.\right)q\left(L\right)},\\
&\chi_\kappa\left(L\right)=\begin{cases}1, & \text{for } L\in C_\kappa\left(X_N\right),\\ 0, & \text{for } L\notin C_\kappa\left(X_N\right).\end{cases}
\end{aligned}\tag{38}$$

That is, the variety preservation is considered with $p(L|X)=q(L|X)$ within $C_\kappa(X_N)$ in help of

$$\begin{aligned}
q\left(X\left|L,\Theta_L\right.\right)&=\int q\left(X\left|Y,L,\Theta_{X|YL}\right.\right)q\left(Y,L\left|\Theta_{YL}\right.\right)\mathrm{d}Y\\
&=\mathrm{e}^{\pi_L(X,Y_L^*,\Theta)}(2\pi)^{d_Y/2}\left|\Pi_L^{Y|X}\right|^{-1/2},\\
C_\kappa\left(X_N\right)&=\left\{L:\ \text{for the first }\kappa\text{ largests of }H_L(\Theta)\right\}.
\end{aligned}$$

From $\mathrm{d}H(p||q,\Theta,\boldsymbol{k},\Xi)=\sum_L[p(L|X_N)\mathrm{d}H_L(\Theta)+H_L(\Theta)\mathrm{d}p(L|X_N)]$, we get

$$\begin{aligned}
&\nabla_{\Theta_L}H\left(p||q,\Theta,\boldsymbol{k}\right)\\
&\quad=p^\delta\left(L\left|X_N\right.\right)\nabla_{\Theta_L}\pi_L\left(X_N,Y_L^*,\Theta\right)\\
&\qquad+p\left(L\left|X_N\right.\right)\nabla_{\Theta_L}R_L\left(X_N,Y_L^*,\Theta\right)\\
&\qquad-\frac{1}{2}\Delta\pi_L\left(X_N,Y_L^*\right)\nabla_{\Theta_L}\ln\left|\Pi_L^{Y|X}\right|,\\
&p^\delta\left(L\left|X_N\right.\right)=p\left(L\left|X_N\right.\right)+\Delta\pi_L\left(X_N,Y_L^*\right),\\
&\Delta\pi_L\left(X,Y\right)=p\left(L\left|X_N\right.\right)\Delta H_L(\Theta),\\
&\Delta H_L(\Theta)=H_L(\Theta)-\sum_L p\left(L\left|X_N\right.\right)H_L(\Theta).
\end{aligned}\tag{39}$$

For the case that $X_N=\{x_t\}$ consists of i.i.d. samples, Eq. (38) leads to the Box-① in Fig. 7. Moreover, it follows from $\mathrm{d}R_L(X_N,Y_L^*,\Theta)/\mathrm{d}h$ (actually its first two terms) that we lead to the Box-④ in Fig. 7.

Alternatively, $H(p||q,\Theta,\boldsymbol{k},\Xi)$ becomes $H(\theta)$ given by Eq. (17) after reducing Σ_L into a sum merely over a single term (i.e., $p(L|X)=1$ and $\Delta\pi_L(X,Y)=0$). For the structure of $p(Y|X,L)$, we consider $\mu_L(X)$ in a parametric form, e.g., $W(x-\mu)$ in Fig. 1(c) for FA, and also Eqs. (29)–(31). That is, we let

$$\begin{aligned}
&\mu_L(X)=\mu_L\left(X,W_L\right)\ \text{ and }\ \Gamma_L^{Y|X}=\Pi_L^{Y|X-1}+\rho^2\\
&\text{with }\mathrm{Tr}\left[\Gamma_L^{Y|X}\Pi_L^{Y|X}\right]=d_Y+\mathrm{Tr}\left[\rho^2\Pi_L^{Y|X}\right],
\end{aligned}\tag{40}$$

where ρ is a diagonal matrix and d_Y is the dimension of Y, from which the last term of $R_L(X_N,Y_L^*,\Theta)$ leads to the last term of $R(h,\theta)$ in Eq. (17). Also, we may consider a priori $q(\rho)$ in $q(\Theta)$.

Furthermore, $H(p||q,\Theta,\boldsymbol{k},\Xi)$ by Eq. (37) also leads to the one at the bottom of Fig. 9 as a combined case of Figs. 7 and 8. One additional issue is that the third term of the gradient of $H(p||q,\Theta,\boldsymbol{k},\Xi)$ by Eq. (39) takes in effect because we no longer have $\Delta\pi_L(X,Y)=0$, which contributes a correcting term to $\pi_t(\theta_l)$ within the Box-② and to Γ_l within the Box-③ in Fig. 9.

With the gradient $\nabla_\Theta H(p||q,\Theta,\boldsymbol{k},\Xi)$, we make a gradient based ascending for $\max_\Theta H(p||q,\Theta,\boldsymbol{k},\Xi)$. Instead of directly using $\nabla_\Theta H(p||q,\Theta,\boldsymbol{k},\Xi)$, we propose to use a technique called primal gradient flow. For a decomposition $\nabla_\Theta f(\Theta)=GE(\Theta)F$ with positive definite matrices G,F, we have $\mathrm{Tr}[\nabla_\Theta f(\Theta)E^{\mathrm{T}}(\Theta)]>0$ by noticing $\mathrm{Tr}[HH^{\mathrm{T}}]>0$ with $H=G^{\mathrm{T}/2}E(\Theta)F^{1/2}$. Thus, we use one of the following two updating formulae for increasing or maximizing $f(\Theta)$:

$$\begin{aligned}
&\text{a) }\Theta^{\text{new}}-\Theta^{\text{old}}\propto E(\Theta)\text{ in general, which is}\\
&\quad\ \text{called primal gradient flow,}\\
&\text{b) }\Theta^{\text{new}}=A^{-1}BC^{-1},\ \text{for }E(\Theta)=A\Theta C-B,\\
&\quad\ \text{if }A,C\text{ are both positive definite.}
\end{aligned}\tag{41}$$

It follows from Eq. (5.2) in Ref. [8] that the case b) with $\mathrm{cond}[H]=\mathrm{cond}[I]=1$ converges faster than the gradient updating $\Theta^{\text{new}}-\Theta^{\text{new}}\propto\nabla_\Theta f(\Theta)$ with $\mathrm{cond}[H]\gg1$, where $\mathrm{cond}[A]$ denotes the condition number of the matrix A. In help of Eq. (40), we obtain those updating equations in Figs. 7–9, particularly the Box-⑦ in Fig. 8 from the above case b).

Considering a learning system in a Ying-Yang pair, we are naturally motivated to make $\max_\Theta H(p||q,\Theta,\boldsymbol{k},\Xi)$ by the following alternative iteration:

Yang step: fixing all the unknowns in the Ying machine, we get Y_L^* by Eq. (37) and $p(L|X)$ by Eq. (38), as well as update W_L and ρ in Eq. (40).

Ying step: fixing the above just updated unknowns, we update all the unknowns in the Ying machine.

It provides a general procedure for developing EM-like algorithms for maximizing the general form $H_\mu(P||Q)$ by Eq. (21), which includes the well known EM algorithm for $H_P(P||Q)$ by Eq. (23) at the special setting $\mu = P$. Again, those algorithms in Figs. 7–9 are examples of this Ying-Yang alternation.

From the perspective of the A5 circling in Fig. 6(b), after getting sampling at A-1, a bottleneck is encountered in implementing A-2, A-3, A-4, i.e., the integral over Θ in Eq. (34) and the integral over Y in Fig. 5(b) for $H(p||q, \Theta, \boldsymbol{k}, \Xi)$, which is tackled with apex approximation for A-2 and primal gradient based ascending $H(p||q, \Theta, \boldsymbol{k}, \Xi)$ for A-3. The role of A-4 consists of detecting Eq. (4) for automatic model selection and of identifying convergence. Finally, the job of A-5 is making a validation, which is omitted here.

4.4 Unsupervised learning, semi-supervised learning, and supervised learning

Conventionally, a learning that only bases on input sample x_t is called unsupervised, for which a BYY system directly applies. Instead, a learning that bases on each input-output sample pair is called supervised, which is also covered by the BYY system. Specifically, there are two types of BYY system that implement supervised learning.

The first type is directly using the BYY system by its Yang machine when either or both of l_t and y_t are available per sample x_t, or when either or both of the pair $\{L_N, X_N\}$ and the pair $\{Y_N, X_N\}$ are available as a whole. What need to do is simply setting the Yang machine with either or both of

$$p(L|X) = \delta_{L,L_N} \text{ and } p(Y|X, L) = \delta(Y - Y_N).$$

Particularly, if both of them are used, the BYY best harmony and BYY best matching learning both equivalently degenerate to making the maximum likelihood learning on $\ln[q(X_N|Y_N, L_N, \Theta_{X|YL})q(Y_N, L_N|\Theta_{YL})]$ of the Ying machine.

To use the information provided by the input-output sample pairs and also use the estimation of Yang machine by unsupervised learning, a better consideration is given as follows:

$$\begin{aligned} &p(L|X) = (1-\varrho)\chi_\kappa(L)\, q(L\,|X) + \varrho\delta_{L,L_N}, \\ &p(Y|X, L) = (1-\varrho)G(Y|\mu_L(X), \Gamma_L^{Y|X}) \\ &\qquad + \varrho G(Y|Y_N, h_y^2 I), \end{aligned} \tag{42}$$

where $\mu_L(X), \Gamma_L^{Y|X}$ are same as in Eq. (37), and $\chi_\kappa(L), q(L\,|X)$ are same as in Eq. (38). Also, $h_y > 0$ is a data smoothing parameter that can be handled in a way similar to h, in help of one $q(h_y)$ together with $q(h)$ to be included in $q(\Theta)$.

The combining coefficient $0 \leqslant \varrho \leqslant 1$ may be either pre-specified or learned together with Θ in help of one priori $q(\varrho)$ included in $q(\Theta)$, e.g., $q(\varrho)$ is a uniform distribution or a beta distribution. It becomes either supervised when $\varrho = 1$ or unsupervised when $\varrho = 0$. Generally, we combine the two. This is a special task of combining classifiers and learning mixture-of-experts [102].

The above discussed directly applies to an even general case that we have two parts of samples, one is the above X_N with teaching pairing for which we use Eq. (42) while the other is X'_N without teaching pairing for which we can still use Eq. (42) by simply letting $\varrho = 0$.

A detailed insight can be obtained by considering the problem of Gaussian mixture introduced in Sect. 3.1. It follows from Eq. (42) that for a given pair (x_t, j_t) we simply have

$$\begin{aligned} &p(j\,|x_t, \theta) = (1-\varrho)q(j\,|x_t, \theta)\,\chi_{\kappa,t}(j) + \varrho\delta_{j,j_t}, \\ &p(j|t) = p\left(j\,\middle|x_t, \theta^{\text{old}}\right), \\ &p_{j,t} = p(j|t) + (1-\varrho)\Delta_{j,t}. \end{aligned} \tag{43}$$

We put the above $p(j\,|x_t, \theta)$ into Eq. (10) per sample pair (x_t, j_t). Correspondingly in the algorithm given in Fig. 7, we use the above $p_{j,t}$ to replace $p_{j,t} = p(j|t) + \Delta_{j,t}$ of the Yang step and use the above $p(j|t)$ in the position of $p(j|t)$ in the Ying step.

What discussed above belongs to a recent topic that becomes quite popular in the machine learning literature under the name of semi-supervised learning. One early exploration on Gaussian mixture was made in 1997 under the name of *semi-unsupervised learning*, see Eq. (7.14) in Ref. [73]. Interestingly, the BYY system provides an easy way to conduct such tasks. Also, we may consider other combining rules reviewed in Ref. [102], in addition to the linear combination in Eq. (42) and Eq. (43).

The second type of BYY supervised learning considers learning an input-output mapping $\xi_t \to \zeta_t$ via a set of pairs $\{\xi_t, \zeta_t\}$, which is implemented by the BYY system with each pair $\{\xi_t, \zeta_t\}$ jointly taking the position of an input sample $x_t = \{\xi_t, \zeta_t\}$ that is mapped into its corresponding inner representation $\{l_t, y_t\}$. Then, the mapping $\xi_t \to \zeta_t$ is obtained via a cascaded mapping $\xi_t \to \{l_t, y_t\} \to \zeta_t$.

Considering $X = \{\boldsymbol{\xi}, \boldsymbol{\zeta}\}$ with $X = \{x_t\}$, $\boldsymbol{\xi} = \{\xi_t\}$, and $\boldsymbol{\zeta} = \{\zeta_t\}$, we start at the special case Y = empty and consider the following decomposition

$$q(X|L, \Theta_{X|L}) = q(\boldsymbol{\zeta}|\boldsymbol{\xi}, L, \Theta_L)q(\boldsymbol{\xi}|L, \Theta_L). \tag{44}$$

We further let

$$p(L|\boldsymbol{\xi}) = p(L|\boldsymbol{\xi},\boldsymbol{\zeta})|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_L(\boldsymbol{\xi})},$$
$$\boldsymbol{\zeta}_L(\boldsymbol{\xi}) = \int \boldsymbol{\zeta}\ q(\boldsymbol{\zeta}|\boldsymbol{\xi},L,\Theta_L)\mathrm{d}\boldsymbol{\zeta}. \tag{45}$$

Putting them into Eq. (34) that is simplified into

$$H(p\|q,\Theta,\boldsymbol{k},\Xi) = \sum_L \int p(L|\boldsymbol{\xi})p(X|X_N,h) \times \ln[q(\boldsymbol{\zeta}|\boldsymbol{\xi},L,\Theta_L)q(\boldsymbol{\xi}|L,\Theta_L)q(L|\Theta_L)q(\Theta_L)]\mathrm{d}X,$$

which is maximized by the way as implemented in the previous subsections. Then, we get $\boldsymbol{\xi}\to\boldsymbol{\zeta}$ by

$$q(\boldsymbol{\zeta}|\boldsymbol{\xi}) = \sum_L p(L|\boldsymbol{\xi})q(\boldsymbol{\zeta}|\boldsymbol{\xi},L,\Theta_L). \tag{46}$$

For the i.i.d. samples, with $p_h(x_t)$ given by Eq. (6) we are lead to the following special case:

$$H(p\|q,\Theta,\boldsymbol{k},\Xi) = \sum_t \sum_{\ell_t} \int p(\ell_t|\xi_t)p_h(x_t) \times \ln[q(\zeta_t|\xi_t,\theta_{\ell_t})G(\xi_t|\mu_{\ell_t},\Sigma_{\ell_t})q(\ell_t)q(\theta_{\ell_t})]\mathrm{d}x_t,$$
$$p(\ell|\xi) = p(\ell|\xi,\zeta)|_{\zeta=\zeta_\ell(\xi)},\quad \zeta_\ell(\xi) = \int \zeta\ q(\zeta|\xi,\theta_\ell)\mathrm{d}\zeta,$$
$$q(\zeta|\xi) = \sum_\ell p(\ell|\xi)q(\zeta|\xi,\theta_\ell), \tag{47}$$

from which we are lead to RBF networks and alternative mixture of experts [50,80], as illustrated by the Box-⑨ in Fig. 11.

Learning algorithms can be obtained from the algorithm in Fig. 7 for Gaussian mixture with some modifications. We can directly use this algorithm for updating the part $G(\xi|\mu_\ell,\Sigma_\ell)\alpha_\ell$, with the first line of Yang step modified by $\pi_t(\theta_\ell)\ = \ln[q(x_t|\theta_\ell)\alpha_\ell]_{x_t=[\xi_t,\zeta_\ell(\xi_t)]}$. Also, the Ying step is added with a new part for updating $q(\zeta|\xi,\theta_\ell)$, which can be implemented either in a batch way that gets $\theta_\ell^{\mathrm{new}}$ by solving $\nabla_{\theta_\ell}\sum_t p_{\ell,t}q(\zeta_t|\xi_t,\theta_\ell)=0$ or adaptively (e.g., see Figs. 3(C), 3(D), and 3(E) in Ref. [50] there with $p_{j,t}$ in the position of $\eta_{j,t}$).

Second, we consider the special case L = empty and the following decomposition:

$$q(X|Y,\Theta_{X|Y}) = q(\boldsymbol{\zeta}|Y,\Theta_{X|Y})q(\boldsymbol{\xi}|Y,\Theta_{X|Y}),$$
$$p(Y|\boldsymbol{\xi}) = p(Y|\boldsymbol{\xi},\boldsymbol{\zeta})|_{\boldsymbol{\zeta}=\mathbf{0}}. \tag{48}$$

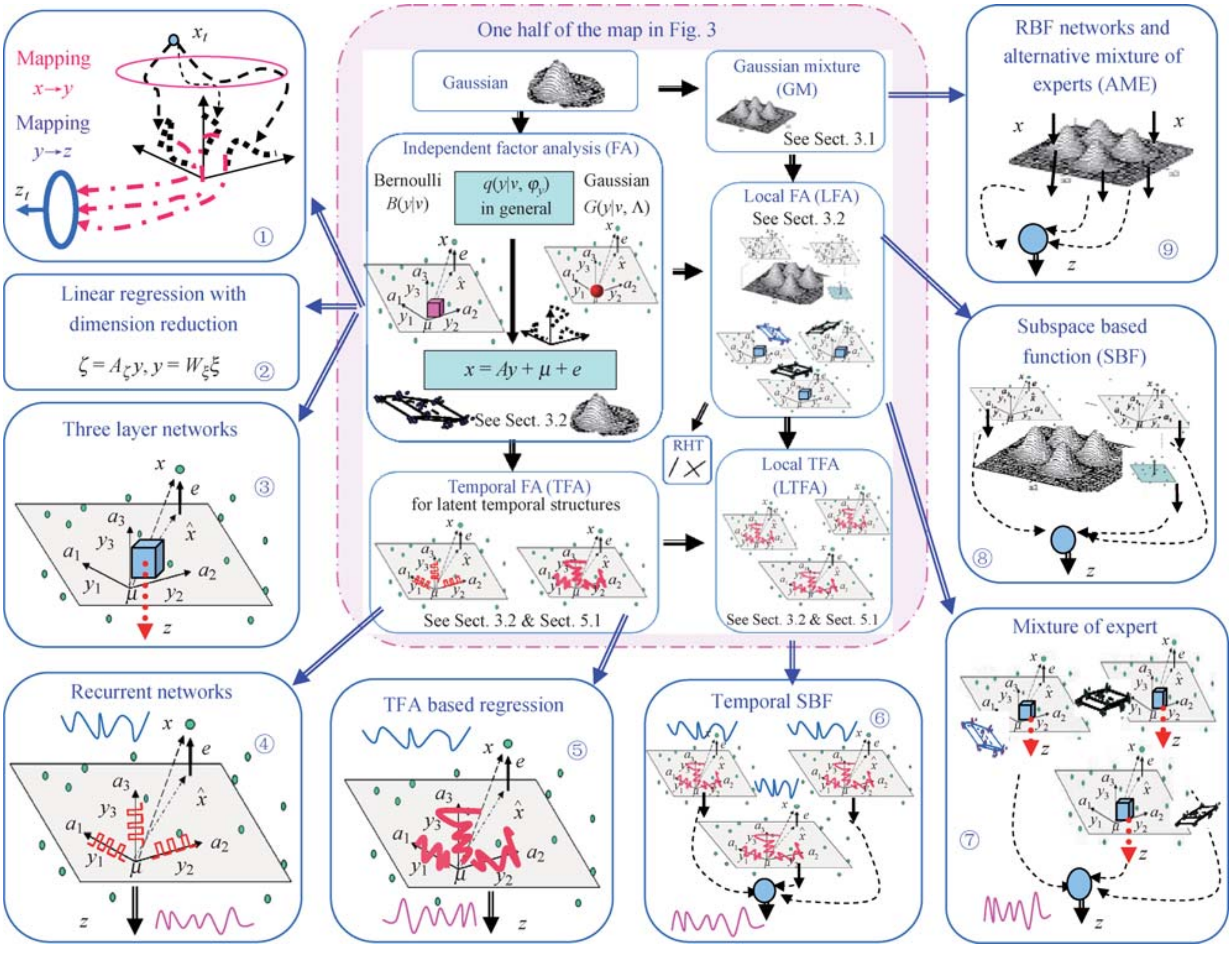


Fig. 11 A roadmap of typical supervised learning tasks

Putting them into Eq. (34) that becomes

$$\begin{aligned}&H(p\|q,\Theta,\boldsymbol{k},\Xi)\\&\quad=\int p(Y|\boldsymbol{\xi})p(X|X_N,h)\\&\qquad\times\ln[q(\boldsymbol{\zeta}|Y,\Theta_{X|Y})q(\boldsymbol{\xi}|Y,\Theta_{X|Y})\\&\qquad\times q(Y|\Theta_Y)q(\Theta)]\mathrm{d}X\mathrm{d}Y,\end{aligned}\tag{49}$$

which is also maximized by the way as implemented in the previous subsections. Then, we get $\boldsymbol{\xi}\to Y\to\boldsymbol{\zeta}$ by

$$q(\boldsymbol{\zeta}|\boldsymbol{\xi})=\int p(Y|\boldsymbol{\xi})q(\boldsymbol{\zeta}|Y,\Theta_{X|Y})\mathrm{d}Y.\tag{50}$$

For the i.i.d. samples, we consider the special case below:

$$\begin{aligned}&H(p\|q,\Theta,\boldsymbol{k},\Xi)\\&\quad=\sum_t\int p(y_t|\xi_t)p_h(x_t)\\&\qquad\times\ln[q(\zeta_t|y_t,\theta_{x|y})q(\xi_t|y_t,\theta_{x|y})\\&\qquad\times q(y_t|\theta_y)q(\theta)]\mathrm{d}y_t\mathrm{d}x_t,\\&\text{(a) }q(x|y,\theta_{x|y})=q(\zeta|y,\theta_{x|y})q(\xi|y,\theta_{x|y}),\\&\qquad p(y|\xi)=p(y|\xi,\zeta)_{\zeta=0},\\&\qquad\text{where}\\&\qquad q(x|y,\theta_{x|y})=G(x|Ay+\mu,\Sigma),\\&\qquad q(\xi|y,\theta_{x|y})=G(\xi|A_\xi y+\mu_\xi,\Sigma_\xi),\\&\qquad q(\zeta|y,\theta_{x|y})=G(\zeta|A_\zeta y+\mu_\zeta,\Sigma_\zeta),\\&\qquad x=[\xi,\zeta]^{\mathrm{T}},\ A^{\mathrm{T}}=[A_\xi,A_\zeta],\\&\qquad \mu=[\mu_\xi,\mu_\zeta]^{\mathrm{T}},\ \Sigma=\mathrm{diag}[\Sigma_\xi,\Sigma_\zeta],\\&\text{(b) }q(\zeta|\xi)=\int p(y|\xi)G(\zeta|A_\zeta y+\mu_\zeta,\Sigma_\zeta)\mathrm{d}y,\\&\qquad\zeta(\xi)=\int g(y)p(y|\xi)\mathrm{d}y=A_\zeta\mu(\xi,W)+\mu_\zeta,\\&\qquad\mu(\xi,W)=\int yp(y|\xi)\mathrm{d}y,\\&\qquad g(y)=\int\zeta\ G(\zeta|A_\zeta y+\mu_\zeta,\Sigma_\zeta)\mathrm{d}\zeta=A_\zeta y+\mu_\zeta,\end{aligned}\tag{51}$$

from which we are lead to three layer network $\xi\to y\to\zeta$ with hidden units being linear, sigmoid, and other nonlinear types [50,80,103], as illustrated by the Box-①, Box-②, and Box-③ in Fig. 11.

Learning algorithms can be obtained from the algorithm in Fig. 8 with a slight modification. What needs to do is letting $\mu(\xi,W)$ in the Yang step replaced by $\mu(\xi,W)_{\zeta=0}$ and letting $\Sigma\leftarrow\Sigma^{\mathrm{new}}$ at the bottom replaced by $\Sigma\leftarrow\mathrm{diag}[\Sigma^{\mathrm{new}}_\xi,\Sigma^{\mathrm{new}}_\zeta]$, i.e., only considering the block diagonal part of Σ^{new}. The mapping $\xi\to\zeta$ is implemented by $q(\zeta|\xi)$ or simply $\zeta(\xi)=A_\zeta\mu(\xi,W)+\mu_\zeta$.

Third, we extend the above case with L taken in consideration. That is, we consider

$$\begin{aligned}&q(X|Y,L,\Theta_{X|YL})=q(\boldsymbol{\zeta}|Y,L,\Theta_{X|YL})q(\boldsymbol{\xi}|Y,L,\Theta_{X|YL}),\\&p(Y|\boldsymbol{\xi},L)=p(Y|\boldsymbol{\xi},\boldsymbol{\zeta},L)|_{\boldsymbol{\zeta}=\mathbf{0}}.\end{aligned}\tag{52}$$

For each specific L, we get a mapping $\boldsymbol{\xi}\to Y\to\boldsymbol{\zeta}$ in the same way as given in Eq. (50). For the i.i.d. samples, we get its counterpart by modifying Eq. (51) via adding the subscript ℓ to all the components, e.g.,

$$\begin{aligned}&H(p\|q,\Theta,\boldsymbol{k},\Xi)\\&\quad=\sum_t\sum_{\ell_t}\int p(\ell_t|\xi_t)p(y_t|\xi_t)p_h(x_t)\\&\qquad\times\ln[G(\xi_t|A_{\xi,\ell}y_t+\mu_{\xi,\ell},\Sigma_{\xi,\ell})\\&\qquad\times G(\zeta_t|A_{\zeta,\ell}y_t+\mu_{\zeta,\ell},\Sigma_{\zeta,\ell})\\&\qquad\times q(y_t|\theta_{y,\ell_t})q(\ell_t)q(\theta_{\ell_t})]\mathrm{d}y_t\mathrm{d}x_t,\\&\text{(a) }q(x|y,\theta_{x|y,\ell})\\&\qquad=G(x|A_\ell y+\mu_\ell,\Sigma_\ell)\\&\qquad=G(\xi|A_{\xi,\ell}y+\mu_{\xi,\ell},\Sigma_{\xi,\ell})G(\zeta|A_{\zeta,\ell}y+\mu_{\zeta,\ell},\Sigma_{\zeta,\ell}),\\&\quad p(y|\xi,\ell)=p(y|\xi,\zeta,\ell)_{\zeta=0},\\&\text{(b) }q(\zeta|\xi,\theta_\ell)=\int p(y|\xi,\ell)G(\zeta|A_{\zeta,\ell}y+\mu_{\zeta,\ell},\Sigma_{\zeta,\ell})\mathrm{d}y,\\&\quad\zeta_\ell(\xi)=A_{\zeta,\ell}\mu_\ell(\xi,W_\ell)+\mu_{\zeta,\ell}.\end{aligned}\tag{53}$$

On the other hand, similar to Eq. (47) we have

$$\begin{aligned}&\text{(c) }q(x|\theta_\ell)=\frac{(2\pi)^{0.5m_\ell}}{|\Pi^{y|x}_\ell|^{0.5}}G(x|A_\ell y^*_\ell+\mu_\ell,\Sigma_\ell)q(y^*_\ell|\theta_{y,\ell}),\\&\qquad y^*_\ell=\arg\max_y[G(x|A_\ell y+\mu_\ell,\Sigma_\ell)q(y|\theta_{y,\ell})],\\&\text{(d) }p(\ell|x)=\frac{q(x|\theta_\ell)\alpha_\ell}{\sum_\ell q(x|\theta_j)\alpha_j},\\&\qquad p(\ell|\xi)=p(\ell|\xi,\zeta)|_{\zeta=\zeta_\ell(y^*_\ell)},\\&\text{(e) }q(\zeta|\xi)=\sum_\ell p(\ell|\xi)q(\zeta|\xi,\theta_\ell),\\&\qquad\zeta(\xi)=\sum_\ell p(\ell|\xi)[A_{\zeta,\ell}\mu_\ell(\xi,W_\ell)+\mu_{\zeta,\ell}],\end{aligned}\tag{54}$$

which combines each local three layer network $\xi\to y\to\zeta$ via $q(\zeta|\xi,\theta_\ell)$ given by Eq. (53).

As illustrated by the Box-⑧ in Fig. 11, when $q(y|\theta_{y,\ell})$ is a Gaussian $G(y|v_\ell,\Lambda_\ell)$, each $q(x|\theta_\ell)$ is a Gaussian featured by a subspace spanned by A_ℓ and located at μ_ℓ, which supports a cascaded linear regression from $\xi\to y$ by $W_\ell(x_t-\mu_\ell)_{\zeta=0}$ and then $y\to\zeta$ by $\zeta_\ell(y)=A_{\zeta,\ell}y+\mu_{\zeta,\ell}$. These subspace based local regression function are combined by $p(\ell|\xi)$, called subspace based functions. The details are referred to Ref. [50] (therein Figs. 5–7 and the subsection after Eq. (11)). When $q(y|\theta_{y,\ell})$ is a Bernoulli $B(y|v_\ell)$, each $A_{\zeta,\ell}\ \mu_\ell(\xi,W_\ell)+\mu_{\zeta,\ell}$ implements a three layer networks and the above $q(\zeta|\xi)$ or $\zeta(\xi)$ actually implements a mixture of experts (ME), as illustrated by the Box-⑦ in Fig. 11. Further details are referred to Refs. [2,17,50,80]. Moreover, when $q(y|\theta_{y,\ell})$ has a temporal dependence (e.g., one given on the top of Fig. 8), we are further lead to those extensions illustrated by the Box-④, Box-⑤, and Box-⑥ in Fig. 11.

Similarly, learning algorithms for typical types of $q(y|\theta_{y,\ell})$ can be obtained from the algorithm in Fig. 9. Taking the SBF functions as an example, we let $W_\ell(x_t-\mu_\ell)$ in the Yang step to be replaced by $W_\ell(x_t-\mu_\ell)_{\zeta=0}$ and $\pi(x_t,y_{\ell,t},\theta_\ell)$ to be replaced by $\pi(x_t,y_{\ell,t},\theta_\ell)_{\zeta=\zeta_\ell(y_{\ell,t})}$, and also let $\Sigma_\ell \leftarrow \Sigma_\ell^{\rm new}$ at the bottom replaced by $\Sigma_\ell \leftarrow {\rm diag}[\Sigma_{\ell,\xi}^{\rm new}, \Sigma_{\ell,\zeta}^{\rm new}]$.

Inherited from the BYY harmony learning nature of automatic model selection, the number of basis functions, experts, hidden units, and subspace dimensions can be determined during implementing these supervised learning algorithms. This favorable nature is shared generally by efforts along a road of decomposing $q(X|Y,L,\Theta_{X|L})$ via the partition $X=\{\boldsymbol{\xi},\boldsymbol{\zeta}\}$ for $\boldsymbol{\xi}\to\boldsymbol{\zeta}$ in help of maximizing $H(p||q,\Theta,\boldsymbol{k})$ by Eq. (34).

5 Hierarchical, temporal BYY harmony learning and HMM examples

5.1 Hierarchical and temporal BYY harmony learning

The BYY harmony learning can also be extended to learning hierarchical models. Here, an introduction is made on extending a Gaussian mixture into a mixture of hierarchical components as shown at the center of in Fig. 12(b), i.e., each component of a finite mixture of $q(x|\theta)$ is itself a mixture of finite components, and each component in a higher layer is a mixture of a number of components in its lower layer. The samples are input at the bottom layer as a part of Yang machine.

We have $Y=\text{empty}, L=\{i,\ell,j\}$ in a tree configuration with $q(\Theta)\propto 1$ (i.e., ignored). Considering a set $\{x_t\}$ of i.i.d. samples that inputs to the tree of three layers shown at the center of in Fig. 12(b), $H(p||q,\Theta,\boldsymbol{k})$ by Eq. (34) is simplified into $H(\theta)$ on the top shown of Fig. 12(a), given by the flow from the bottom up to the top. At the bottom, it consists of the likelihood $\pi_t(\theta_{j|\ell\, i})$ of a component on a sample and the corresponding regularization information. After a sum by the weight $p(j|l,i,x_t)$, the flow from each component is integrated into one upper layer component. Finally, on the top layer, the flow $H_t(\theta)$ for each sample is summed up over all the samples. Specifically, the detailed form of $H_t(\theta)$ is given as follows:

$$H_t(\theta)=\sum_{i,\ell,j}\int p\left(j|\ell,i,x_t\right)p\left(\ell|i,x_t\right)p\left(i|x_t\right)$$
$$\times p_h\left(x_t\right)\ln\left[q\left(x_t|\theta_{j|\ell\, i}\right)\alpha_{j|\ell\, i}\alpha_{\ell|i}\alpha_i q(h|X_N)\right]{\rm d}x_t.$$

The Ying-Yang alternating algorithm given in Fig. 12(b) is developed from $\nabla_\Theta H(p||q,\Theta,\boldsymbol{k})$. In help of the chain rule for derivatives, we can obtained the gradients hierarchically as follows:

$$\begin{aligned}
\nabla_{\phi_i}H_t(\theta)&=p\left(i|x_t\right)\nabla_{\phi_i}H_t\left(\theta_i\right)+\Delta_{i,t}\nabla_{\phi_i}\pi_t\left(\theta_i\right),\\
\nabla_{\phi_i}\pi_t\left(\theta_i\right)&=\nabla_{\phi_i}\ln\alpha_i+\nabla_{\phi_i}\ln q\left(x|\theta_i\right),\\
\nabla_{\phi_i}\ln q\left(x|\theta_i\right)&=\sum_\ell p\left(\ell|i,x_t\right)\nabla_{\phi_i}\pi_t\left(\theta_{\ell|i}\right),\\
\nabla_{\phi_i}\pi_t\left(\theta_{\ell|i}\right)&=\nabla_{\phi_i}\ln\alpha_{\ell|i}+\nabla_{\phi_i}\ln q\left(x|\theta_{\ell|i}\right),\\
\nabla_{\phi_i}\ln q\left(x|\theta_{\ell|i}\right)&=\sum_j p\left(j|\ell,i,x_t\right)\nabla_{\phi_i}\pi_t\left(\theta_{j|\ell i}\right),\\
\nabla_{\phi_i}\pi_t\left(\theta_{j|\ell i}\right)&=\nabla_{\phi_i}\ln\alpha_{j|\ell i}+\nabla_{\phi_i}\ln q\left(x|\theta_{j|\ell i}\right).
\end{aligned}\tag{55}$$

Accordingly, we can implement $\max_\Theta H(p||q,\Theta,\boldsymbol{k})$ hierarchically as shown in Fig. 12(b). The Yang step updates the Bayesian posteriors $p(j|l,i,x_t)$, $p(l|i,x_t)$, $p(i|x_t)$ and the corresponding $\Delta_{j|li,t}$, $\Delta_{i|l,t}$, $\Delta_{i,t}$ from the bottom up to the top, which are transferred horizontally to the Ying step on each layer. Then, from the top down to the bottom, the Ying step updates α_i, $\alpha_{i|l}$, $\alpha_{j|li}$, and finally updates the components at the bottom, e.g., $\mu_{j|li}$, $\Sigma_{j|li}$ for each Gaussian.

Setting $\Delta_{j|li,t}=0$, $\Delta_{i|l,t}=0$, $\Delta_{i,t}=0$, the Ying-Yang iteration by Fig. 12(b) actually degenerates back to the EM algorithm for a hierarchical mixture. With the correcting terms $\Delta_{i,t}\neq 0$, $\Delta_{i|l,t}\neq 0$, $\Delta_{j|li,t}\neq 0$, the Ying-Yang iteration implements the BYY harmony learning, during which the number of components in each layer is determined by automatic model selection. That is, those extra components are discarded if the corresponding $\alpha_i\to 0$, $\alpha_{i|l}\to 0$, and $\alpha_{j|li}\to 0$.

A typical example of using hierarchical mixture of Gaussians shown in Fig. 12 is tree-based clustering in context dependent phone modeling in HMM based acoustic models [104]. This is commonly made by decision trees in help of a greedy iterative node splitting algorithm. However, the depth of a tree or the number of Gaussian components is controlled heuristically. Recently in Ref. [105], BYY harmony learning has been applied for a part of this purpose. It has been shown that the number of Gaussian components has been reduced with considerably improvements on recognition word error rate (WER). The Ying-Yang alternating given in Fig. 12 provides a further tool for this application.

In addition to a hierarchical relation bottom up from observed samples, temporal dependence among samples can also be introduced into a BYY system by modeling relation of hidden representations across times. Two types of temporal BYY system are considered by Sect. II(C) in Ref. [84]. One considers $H\left(p||q,\Theta,\boldsymbol{k}\right)$ in Eq. (34) by the following sum with each $H(x_t,y_t)$ given by an instantaneous Ying-Yang pair:

$$H\left(p||q,\Theta,\boldsymbol{k}\right)=\sum_t H(x_t,y_t)+\ln\left[q(h|X_N)q(\Theta)\right],$$

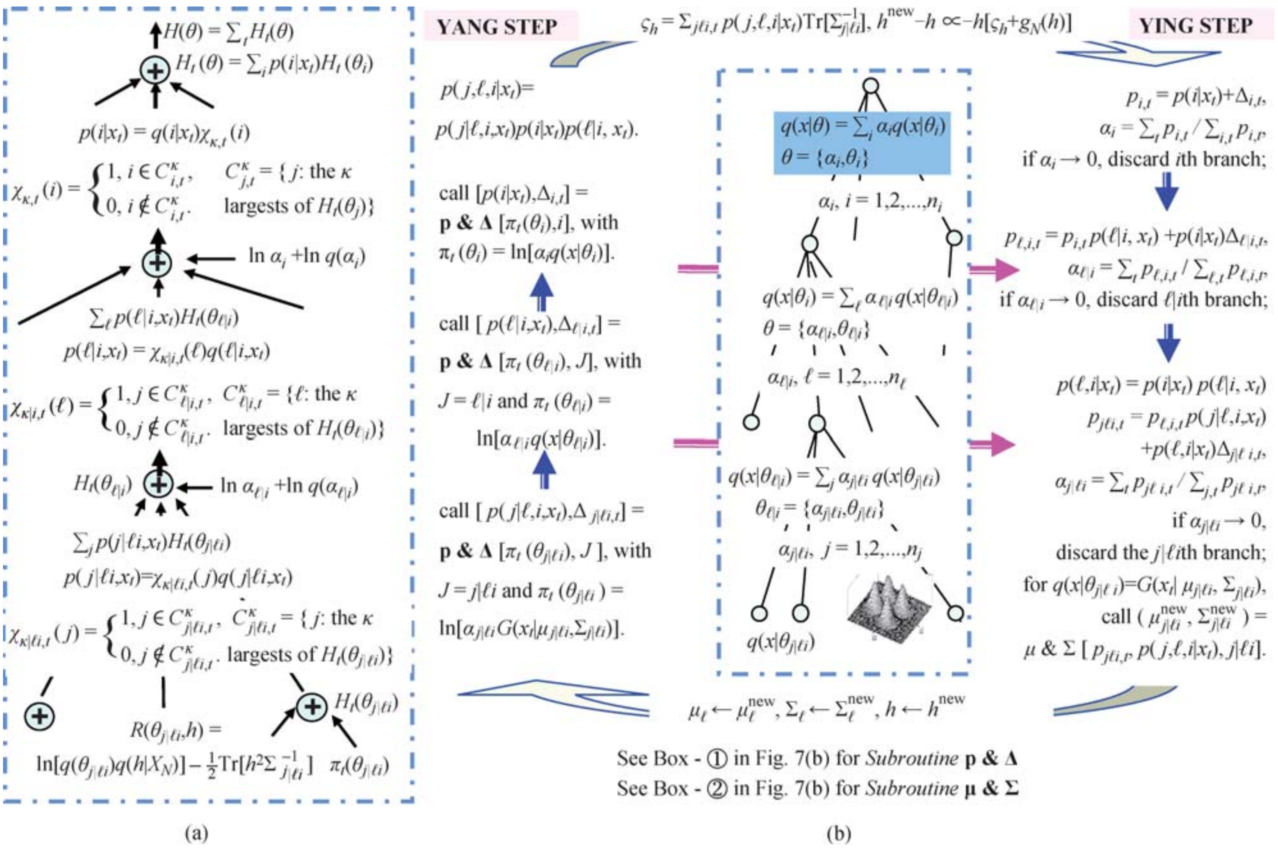

$H(\theta)=\sum_t H_t(\theta)$
$H_t(\theta)=\sum_i p(i|x_t)H_t(\theta_i)$
$p(i|x_t)=q(i|x_t)\chi_{\kappa,t}(i)$
$\chi_{\kappa,t}(i)=\begin{cases}1, i\in C^{\kappa}_{i,t}, & C^{\kappa}_{j,t}=\{j\text{: the } \kappa\\ 0, i\notin C^{\kappa}_{i,t}. & \text{largests of } H_t(\theta_j)\}\end{cases}$
$\ln\alpha_i+\ln q(\alpha_i)$
$\sum_\ell p(\ell|i,x_t)H_t(\theta_{\ell|i})$
$p(\ell|i,x_t)=\chi_{\kappa|i,t}(\ell)q(\ell|i,x_t)$
$\chi_{\kappa|i,t}(\ell)=\begin{cases}1, j\in C^{\kappa}_{\ell|i,t}, & C^{\kappa}_{\ell|i,t}=\{\ell\text{: the } \kappa\\ 0, j\notin C^{\kappa}_{\ell|i,t}. & \text{largests of } H_t(\theta_{\ell|i})\}\end{cases}$
$H_t(\theta_{\ell|i})$
$\ln\alpha_{\ell|i}+\ln q(\alpha_{\ell|i})$
$\sum_j p(j|\ell i,x_t)H_t(\theta_{j|\ell i})$
$p(j|\ell i,x_t)=\chi_{\kappa|\ell i,t}(j)q(j|\ell i,x_t)$
$\chi_{\kappa|\ell i,t}(j)=\begin{cases}1, j\in C^{\kappa}_{j|\ell i,t}, & C^{\kappa}_{j|\ell i,t}=\{j\text{: the } \kappa\\ 0, j\notin C^{\kappa}_{j|\ell i,t}. & \text{largests of } H_t(\theta_{j|\ell i})\}\end{cases}$
$H_t(\theta_{j|\ell i})$
$R(\theta_{j|\ell i},h)=$
$\ln[q(\theta_{j|\ell i})q(h|X_N)]-\frac{1}{2}\mathrm{Tr}[h^2\Sigma^{-1}_{j|\ell i}]$
$\pi_t(\theta_{j|\ell i})$
(a)
YANG STEP
$\varsigma_h=\sum_{j\ell i,t}p(j,\ell,i|x_t)\mathrm{Tr}[\Sigma^{-1}_{j|\ell i}]$, $h^{\text{new}}-h\propto -h[\varsigma_h+g_N(h)]$
YING STEP
$p(j,\ell,i|x_t)=$
$p(j|\ell,i,x_t)p(i|x_t)p(\ell|i,x_t)$.
call $[p(i|x_t),\Delta_{i,t}]=$
p & Δ $[\pi_t(\theta_i),i]$, with
$\pi_t(\theta_i)=\ln[\alpha_i q(x|\theta_i)]$.
call $[p(\ell|i,x_t),\Delta_{\ell|i,t}]=$
p & Δ $[\pi_t(\theta_{\ell|i}),J]$, with
$J=\ell|i$ and $\pi_t(\theta_{\ell|i})=$
$\ln[\alpha_{\ell|i}q(x|\theta_{\ell|i})]$.
call $[p(j|\ell,i,x_t),\Delta_{j|\ell i,t}]=$
p & Δ $[\pi_t(\theta_{j|\ell i}),J]$, with
$J=j|\ell i$ and $\pi_t(\theta_{j|\ell i})=$
$\ln[\alpha_{j|\ell i}G(x_t|\mu_{j|\ell i},\Sigma_{j|\ell i})]$.
$q(x|\theta)=\sum_i\alpha_i q(x|\theta_i)$
$\theta=\{\alpha_i,\theta_i\}$
$\alpha_i,\ i=1,2,\ldots,n_i$
$q(x|\theta_i)=\sum_\ell\alpha_{\ell|i}q(x|\theta_{\ell|i})$
$\theta=\{\alpha_{\ell|i},\theta_{\ell|i}\}$
$\alpha_{\ell|i},\ \ell=1,2,\ldots,n_\ell$
$q(x|\theta_{\ell|i})=\sum_j\alpha_{j|\ell i}q(x|\theta_{j|\ell i})$
$\theta_{\ell|i}=\{\alpha_{j|\ell i},\theta_{j|\ell i}\}$
$\alpha_{j|\ell i},\ j=1,2,\ldots,n_j$
$q(x|\theta_{j|\ell i})$
$p_{i,t}=p(i|x_t)+\Delta_{i,t}$,
$\alpha_i=\sum_t p_{i,t}/\sum_{i,t}p_{i,t}$,
if $\alpha_i\rightarrow 0$, discard ith branch;
$p_{\ell,i,t}=p_{i,t}p(\ell|i,x_t)+p(i|x_t)\Delta_{\ell|i,t}$,
$\alpha_{\ell|i}=\sum_t p_{\ell,i,t}/\sum_{\ell,t}p_{\ell,i,t}$,
if $\alpha_{\ell|i}\rightarrow 0$, discard $\ell|i$th branch;
$p(\ell,i|x_t)=p(i|x_t)\,p(\ell|i,x_t)$
$p_{j\ell i,t}=p_{\ell,i,t}p(j|\ell,i,x_t)$
$+p(\ell,i|x_t)\Delta_{j|\ell\, i,t}$,
$\alpha_{j|\ell i}=\sum_t p_{j\ell\, i,t}/\sum_{j,t}p_{j\ell\, i,t}$,
if $\alpha_{j|\ell i}\rightarrow 0$,
discard the $j|\ell i$th branch;
for $q(x|\theta_{j|\ell\, i})=G(x_t|\mu_{j|\ell i},\Sigma_{j|\ell i})$,
call $(\mu^{\text{new}}_{j|\ell i},\Sigma^{\text{new}}_{j|\ell i})=$
μ & Σ $[p_{j\ell i,t},p(j,\ell,i|x_t),j|\ell i]$.
$\mu_\ell\leftarrow\mu^{\text{new}}_\ell$, $\Sigma_\ell\leftarrow\Sigma^{\text{new}}_\ell$, $h\leftarrow h^{\text{new}}$
See Box - ① in Fig. 7(b) for *Subroutine* **p & Δ**
See Box - ② in Fig. 7(b) for *Subroutine* **μ & Σ**
(b)

Fig. 12 Hierarchical BYY harmony learning. (a) Hierarchical harmony flows; (b) main program

$$H(x,y)=\sum_\ell\int p(\ell,y|x)G(x|Ex,h^2I)\pi_\ell(x,y,\theta)\mathrm{d}x\mathrm{d}y,$$
$$\pi_\ell(x,y,\theta)=\ln[q(x|y,\ell,\theta_{x|y,\ell})q(y,\ell)], \tag{56}$$

where temporal relation is encoded by $q(y_t,\ell_t|\omega_{t-1},\theta_{y,\ell})$ via $q(y_t,\ell_t)=\int q(y_t,\ell_t|\omega_{t-1},\theta_{y,\ell})q(\omega_{t-1})\mathrm{d}\omega_{t-1}$ with $\omega_{t-1}=\{y_{t-\tau},\ell_{t-\tau}\}^{\kappa}_{\tau=1},\kappa\geqslant 1$. One example is the temporal extension of factor analysis (FA) on the top of Fig. 8 where we drop ℓ and have $q(y|\theta_y)=G(y|\nu,\Lambda)$.

The other type describes temporal dependence among $Y=\{y_t\},L=\{\ell_t\}$ by a finite order Markovian, especially the first order. One typical example is

$$q\left(\boldsymbol{Y},\boldsymbol{L}\,|\theta_y\right)=q\left(y_0,\ell_0\right)\prod_{t\geqslant 1}q\left(y_t,\ell_t\,|y_{t-1},\ell_{t-1},\theta_y\right),$$
$$q\left(X_N|\,\boldsymbol{Y},\boldsymbol{L},\theta_{x|y}\right)=\prod_{t\geqslant 1}q(x_t|y_t,\ell_t,\theta_{x|y}), \tag{57}$$

which are put into Eq. (34) for implementing the BYY harmony learning. There are two ways to remove the integral over X. One starts from Eq. (37) via putting Eq. (57) into $\pi_L\left(X,Y,\Theta\right)$. The other is simplifying Eq. (34) in help of a temporal decoupling nature of $\ln[q(X|Y,L,\Theta_{X|YL})q(Y,L|\Theta_{YL})]$, resulting in

$$H(p||q,\theta,\boldsymbol{k})=\sum_{t=1}^{N}[H_t(\Theta,\boldsymbol{k})+\ln\left[q(h|X_N)q(\Theta)\right]],$$
$$H_t(\Theta,\boldsymbol{k})=\sum_{\ell_t,\ell_{t-1}}\int p(\ell_t,y_t;\ell_{t-1},y_{t-1}|x_t)\times G(x_t|Ex_t,h^2I)\pi_{\ell_t}(x_t,y_t,y_{t-1},\theta)\mathrm{d}x_t\mathrm{d}y_t\mathrm{d}y_{t-1},$$
$$\pi_{\ell_t}(x_t,y_t,y_{t-1},\theta)=\ln[q(x_t|y_t,\ell_t,\theta_{x|y,\ell_t})q(y_t|y_{t-1},\ell_t,\theta_{y,\ell_t})\times q(\ell_t|\ell_{t-1},Q)], \tag{58}$$

with $q(x_t|y_t,\ell_t,\theta_{x|y,\ell_t})$ and $q(y_t|y_{t-1},\ell_t,\theta_{y,\ell_t})$ in different specific structures, we are lead to different models of the first order Markovian based temporal BYY harmony learning. One typical example is hidden Markov models (HMMs). The detailed discussions are delayed to the next subsection. Here, we introduce another typical example by dropping $L=\{\ell_t\}$ in Eq. (58).

Using the same technique in Eq. (37), we remove the integral over x_t and make Eq. (58) simplified into

$$H\left(p||q,\theta,\boldsymbol{k}\right)=\sum_{t=1}^{N}\left[H_t(\Theta,\boldsymbol{k})+R(h,\Theta)\right],$$
$$R(h,\Theta)=\ln\left[q(h|X_N)q(\Theta)\right]-\frac{1}{2}\mathrm{Tr}\left[h^2\Sigma^{-1}\right],$$
$$\Sigma^{-1}=-\nabla^2_{xx^{\mathrm{T}}}\pi\left(x,y_t\,|y_{t-1},\theta\right),$$
$$H_t(\Theta,\boldsymbol{k})=\int p\left(y_t,y_{t-1}\,|x_t\right)\pi\left(x_t,y_t|y_{t-1},\theta\right)\mathrm{d}y_t\mathrm{d}y_{t-1},$$

$$\pi\left(x_t, y_t | y_{t-1}, \theta\right) = \ln\left[q\left(x_t | y_t, \theta_{x|y}\right) q\left(y_t\,| y_{t-1}, \theta_y\right)\right]. \tag{59}$$

The maximization of the above $H\left(p||q, \theta, \boldsymbol{k}\right)$ is implemented by the schematic algorithm as shown in Fig. 13, in help of getting $\nabla_\Theta H(p||q, \Theta, \boldsymbol{k})$. Specifically, the integral over y_t, y_{t-1} can be handled in one of two ways. One is first making the gradient operation and then attempts to compute the integral over y_t, y_{t-1}. The other way is first removing the integral over y_t, y_{t-1} in help of apex approximation by Eq. (35) around $Y^*_{t,t-1}$ given by the Ying step in Fig. 13, in a way similar to Eq. (37), from which we have $H_t(\Theta, k) = \pi(x_t, y^*_t, y^*_{t-1}, \theta) - \frac{1}{2}\mathrm{Tr}[(\Gamma^{y|x}_{t,t-1} + \varepsilon_{t,t-1}\varepsilon^{\mathrm{T}}_{t,t-1})\Pi^{y|x}_{t,t-1}]$ and $\varepsilon_{t,t-1} = Y^*_{t,t-1} - \bar{Y}_{t,t-1}$, with $\Gamma^{y|x}_{t,t-1}, \bar{Y}_{t,t-1}$ obtained in two choices given by the Yang step in Fig. 13. The choice (a) is same as the one used in Eq. (37), with $p(y_t, y_{t-1}|x_t) = q(y_t, y_{t-1}|x_t)$ designed from the following Bayesian inverse:

$$q(y_t, y_{t-1}|x_t) = \frac{q(y_t, y_{t-1}, x_t)}{\displaystyle\int q(y_t, y_{t-1}, x_t)\mathrm{d}y_t\mathrm{d}y_{t-1}}. \tag{60}$$

The choice (b) is equivalent to the choice (a) when the above $q(y_t, y_{t-1}|x_t)$ is Gaussian, In general, the choice (b) provides an approximation to the choice (a).

Particularly, when $q(y_t|\theta_y) = G(y_t|By_{t-1}, \Lambda)$ and $q(x_t|y_t, \theta_{x|y}) = G(x_t|Ay_t, \Sigma)$, we are lead to temporal factor analysis (TFA) [17,73(Sect. 5),83,84] that is an extension of factor analysis (FA) in Fig. 8 with $q(y_t|\theta_y) = G(y_t|\nu, \Lambda)$. Alternatively, it may also be regarded as a state space model widely studied in the literature of control theory and signal processing [5].

5.2 Bottom-up decoupling versus temporal decoupling

A further insight can be obtained via a discussion on a bottom-up hierarchical decoupling nature of BYY best matching and a temporal decoupling nature of BYY best harmony. We start from the following example of BYY best matching

$$\begin{aligned}\mathrm{KL}\left(p||q\right) &= \sum_t \sum_{i,\ell,j} \int p\left(j|\ell, i, x_t\right) p\left(\ell|i, x_t\right) p\left(i|x_t\right) \\ &\quad \times p\left(x_t\right) \ln \frac{p\left(j|\ell, i, x_t\right) p\left(\ell|i, x_t\right) p\left(i|x_t\right) p\left(x_t\right)}{q\left(x_t|\theta_{j|\ell\ i}\right) \alpha_{j|\ell\ i}\alpha_{\ell|i}\alpha_i}\mathrm{d}x_t,\end{aligned}$$

with respect to $p(j|l, i, x_t)p(l|i, x_t)p(i|x_t)$, first we minimize $\mathrm{KL}(p||q)$ respect to $p(j|l, i, x_t)$, resulting in

$$\begin{aligned}\mathrm{KL}\left(p||q\right) = \sum_t \sum_{i,\ell} \int p\left(\ell|i, x_t\right) p\left(i|x_t\right) p\left(x_t\right) \\ \times \ln \frac{p\left(\ell|i, x_t\right) p\left(i|x_t\right) p\left(x_t\right)}{q\left(x_t|\theta_{\ell|i}\right) \alpha_{\ell|i}\alpha_i}\mathrm{d}x_t,\end{aligned}$$

which is further minimized with respect to $p(l|i, x_t)$, resulting in

$$\mathrm{KL}\left(p||q\right) = \sum_t \sum_i \int p\left(i|x_t\right) p\left(x_t\right) \ln \frac{p\left(i|x_t\right) p\left(x_t\right)}{q\left(x_t|\theta_i\right) \alpha_i}\mathrm{d}x_t.$$

Its minimization with respect to $p(i|x_t)$ further leads to $\mathrm{KL}(p||q) = \sum_t \int p(x_t) \ln[p(x_t)/q(x_t|\theta)]\mathrm{d}x_t$.

In other words, the BYY best matching of a three layer will reduce the problem to the BYY best matching of a two layer problem as if we are dealing with each component $q(x|\theta_{l|i})$ without involving the third layer. Also, we get a similar scenario from the second layer to the first layer, and then to the top layer. That is, we get a bottom-up decoupling nature for $\mathrm{KL}(p||q)$ based BYY best matching. This nature makes $\mathrm{KL}(p||q)$ within one layer become insensitive to the number of components from the lower layers. So, it is poor on determining the number of branches of each node, and thus a tree configuration.

We encounter a similar bottom-up hierarchical decoupling nature when we consider the minimization of

$$\mathrm{KL}\left(p\|\,q\right) = \int p\left(R|\,X\right) p\left(X\right) \ln \frac{p\left(R|\,X\right) p\left(X\right)}{q\left(\,X|\,R\right) q\left(R\right)}\mathrm{d}R\mathrm{d}X,$$

featured by the hierarchy of inner representation $R = \{\{(Y, L), \Theta\}, \boldsymbol{k}\}$, i.e., we consider the minimization of

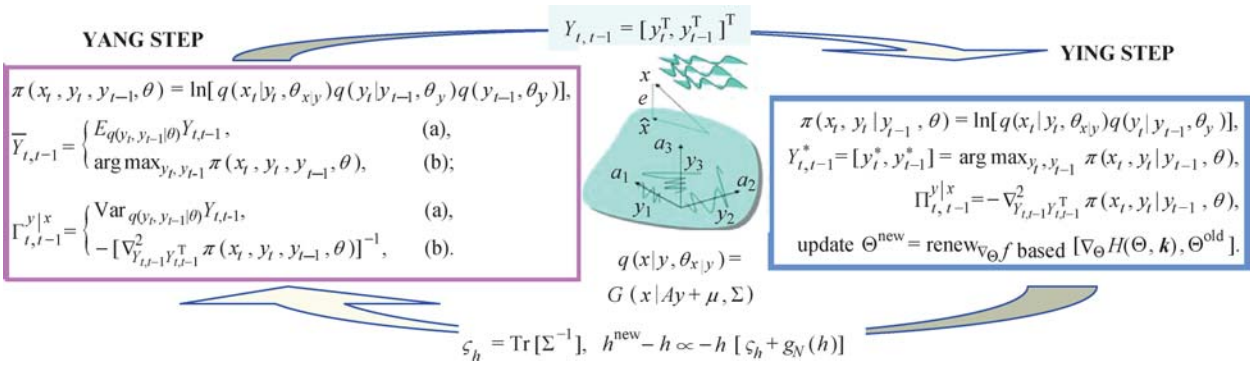

YANG STEP
$Y_{t,t-1} = [y^{\mathrm{T}}_t, y^{\mathrm{T}}_{t-1}]^{\mathrm{T}}$
YING STEP
$\pi(x_t, y_t, y_{t-1}, \theta) = \ln[q(x_t|y_t, \theta_{x|y})q(y_t|y_{t-1}, \theta_y)q(y_{t-1}, \theta_y)]$,
$\overline{Y}_{t,t-1} = \begin{cases} E_{q(y_t, y_{t-1}|\theta)}Y_{t,t-1}, & \text{(a)}, \\ \arg\max_{y_t, y_{t-1}} \pi(x_t, y_t, y_{t-1}, \theta), & \text{(b)}; \end{cases}$
$\Gamma^{y|x}_{t,t-1} = \begin{cases} \mathrm{Var}_{q(y_t, y_{t-1}|\theta)}Y_{t,t-1}, & \text{(a)}, \\ -[\nabla^2_{Y_{t,t-1}Y^{\mathrm{T}}_{t,t-1}} \pi(x_t, y_t, y_{t-1}, \theta)]^{-1}, & \text{(b)}. \end{cases}$
x
e
$\hat{x}$
a_3
y_3
a_1
a_2
y_1
y_2
$q(x|y, \theta_{x|y}) = G(x|Ay + \mu, \Sigma)$
$\pi(x_t, y_t|y_{t-1}, \theta) = \ln[q(x_t|y_t, \theta_{x|y})q(y_t|y_{t-1}, \theta_y)]$,
$Y^*_{t,t-1} = [y^*_t, y^*_{t-1}] = \arg\max_{y_t, y_{t-1}} \pi(x_t, y_t|y_{t-1}, \theta)$,
$\Pi^{y|x}_{t,t-1} = -\nabla^2_{Y_{t,t-1}Y^{\mathrm{T}}_{t,t-1}} \pi(x_t, y_t|y_{t-1}, \theta)$,
update $\Theta^{\mathrm{new}} = \mathrm{renew}_{\nabla_\Theta f}$ based $[\nabla_\Theta H(\Theta, \boldsymbol{k}), \Theta^{\mathrm{old}}]$.
$\varsigma_h = \mathrm{Tr}[\Sigma^{-1}]$, $h^{\mathrm{new}} - h \propto -h\,[\varsigma_h + g_N(h)]$

Fig. 13 Temporal factor analysis by temporal BYY harmony learning

$$
\begin{aligned}
&\mathrm{KL}(p||q)\\
&\quad=\int p(\boldsymbol{k}|X)p(\Theta|X,\boldsymbol{k})p(Y,L|\Theta,\boldsymbol{k},X)p(X)\\
&\qquad\times\ln\frac{p(\boldsymbol{k}|X)p(\Theta|X,\boldsymbol{k})p(Y,L|\Theta,\boldsymbol{k},X)p(X)}{q(X|R)q(Y,L|\Theta,\boldsymbol{k})q(\Theta,\boldsymbol{k})q(\boldsymbol{k})}\mathrm{d}R\mathrm{d}X,
\end{aligned}
$$

with respect to $p(\boldsymbol{k}|X)p\,(\Theta|X,\boldsymbol{k})\,p\,(Y,L|\Theta,\boldsymbol{k},X)$. We first minimize $\mathrm{KL}(p||q)$ respect to $p(Y,L|\Theta,\boldsymbol{k},X)$, which leads to

$$
\begin{aligned}
\mathrm{KL}\,(p||q)&=\sum_{\boldsymbol{k}}\int p\,(\boldsymbol{k}|X)\,p\,(\Theta|X,\boldsymbol{k})\,p\,(X)\\
&\qquad\times\ln\frac{p\,(\boldsymbol{k}|X)\,p\,(\Theta|X,\boldsymbol{k})\,p\,(X)}{q\,(X|\Theta,\boldsymbol{k})\,q\,(\Theta|\boldsymbol{k})\,q\,(\boldsymbol{k})}\mathrm{d}\Theta\mathrm{d}X,\\
p(Y,L|\Theta,\boldsymbol{k},X)&=\frac{q\,(X|R)\,q\,(Y,L|\Theta,\boldsymbol{k})}{q\,(X|\Theta,\boldsymbol{k})},\\
q\,(X|\Theta,\boldsymbol{k})&=\sum_{L}\int q\,(X|R)\,q\,(Y,L|\Theta,\boldsymbol{k})\,\mathrm{d}Y,
\end{aligned}
\tag{61}
$$

which is further minimized with respect to $p\,(\Theta|X,\boldsymbol{k})$, resulting in

$$
\begin{aligned}
\mathrm{KL}\,(p||q)&=\sum_{\boldsymbol{k}}\int p\,(\boldsymbol{k}|X)\,p\,(X)\ln\frac{p\,(\boldsymbol{k}|X)\,p\,(X)}{q\,(X|\boldsymbol{k})\,q\,(\boldsymbol{k})}\mathrm{d}X,\\
p\,(\Theta|X,\boldsymbol{k})&=\frac{q\,(X|\Theta,\boldsymbol{k})\,q\,(\Theta|\boldsymbol{k})}{q\,(X|\boldsymbol{k})},\\
q\,(X|\boldsymbol{k})&=\int q\,(X|\Theta,\boldsymbol{k})\,q\,(\Theta|\boldsymbol{k})\,\mathrm{d}\Theta.
\end{aligned}
\tag{62}
$$

Its minimization with respect to $p\,(\boldsymbol{k}|X)$ further leads to $\mathrm{KL}(p||q)=\int p\,(X)\ln[p\,(X)\,/q\,(X)]\mathrm{d}X$ with $q\,(X)=\sum_{\boldsymbol{k}}q\,(X|\boldsymbol{k})\,q\,(\boldsymbol{k})$.

In other words, the BYY best matching can solve three levels of inverse problems one by one from the bottom upwards, such that the second inverse problem is decoupled from the first one that is summarized into $q\,(X|\Theta,\boldsymbol{k})$ by Eq. (61), a typical example is given in Fig. 5(a), and then the third inverse problem is decoupled from the second one that is summarized into $q\,(X|\boldsymbol{k})$ by in Eq. (62), a typical example is given in Fig. 4(d). This bottom-up decoupling nature makes the tasks of handling latent variables, parameter learning, and model selection decoupled to be conducted sequentially step by step. Though this nature makes implementation easy and thus regarded as being favorable traditionally, it makes the role of $q\,(Y,L|\Theta,\boldsymbol{k})$ not considered in the task of model selection while model selection is made only via appropriate priori $q\,(\Theta|\boldsymbol{k})\,q\,(\boldsymbol{k})$, as previously discussed about Fig. 5.

We have a very different scenario for the BYY best harmony. There is no such a bottom-up decoupling nature for $H(p||q)$. From the relation $H(p||q)=H(p||p)-\mathrm{KL}(p||q)$ by Eq. (24), it follows that

$$
\begin{aligned}
H(p||q)&=\sum_{t}\sum_{i,\ell,j}\int p\,(j|\ell,i,x_t)\,p\,(\ell|i,x_t)\,p\,(i|x_t)\,p\,(x_t)\ln\left[q\left(x_t\left|\theta_{j|\ell\ i}\right.\right)\alpha_{j|\ell\ i}\alpha_{\ell|i}\alpha_i\right]\mathrm{d}x_t\\
&=\sum_{t}\sum_{i,\ell,j}\int p\,(j|\ell,i,x_t)\,p\,(\ell|i,x_t)\,p\,(i|x_t)\,p\,(x_t)\ln p\,(j|\ell,i,x_t)\mathrm{d}x_t\\
&\quad+\sum_{t}\sum_{i,\ell}\int p\,(\ell|i,x_t)\,p\,(i|x_t)\,p\,(x_t)\ln\left[q\left(x\left|\theta_{\ell|i}\right.\right)\alpha_{\ell|i}\alpha_i\right]\mathrm{d}x_t;\\
H(p||q)&=\int p\,(\boldsymbol{k}|\,X)\,p\,(\Theta|\,X,\boldsymbol{k})\,p\,(Y,L|\,\Theta,\boldsymbol{k},X)\,p\,(X)\ln\,p\,(Y,L|\,\Theta,\boldsymbol{k},X)\,\mathrm{d}R\mathrm{d}X\\
&\quad+\sum_{\boldsymbol{k}}\int p\,(\boldsymbol{k}|\,X)\,p\,(\Theta|\,X,\boldsymbol{k})\,p\,(X)\ln[q\,(X|\Theta,\boldsymbol{k})\,q\,(\Theta|\boldsymbol{k})\,q\,(\boldsymbol{k})]\mathrm{d}\Theta\mathrm{d}X.
\end{aligned}
$$

We observe that a best Ying-Yang harmony of a three layer system consists of not only a best Ying-Yang harmony of a two layer system, but also the above first term that relates to minimizing the complexity of the third layer. We get a similar scenario from the second layer to the first layer, and then to the top layer. Hence, the BYY best harmony learning makes an automatic model selection on determining the number of components of each layer, and finally a tree configuration. Similarly, we observe that a best Ying-Yang harmony of the entire BYY system is not just making a best Ying-Yang harmony of the two upper levels via $q\,(X|\Theta,\boldsymbol{k})$, but also including a term that relates to minimizing the complexity of the bottom level via $p\,(Y,L|\,\Theta,\boldsymbol{k},X)$.

Interestingly, the BYY best harmony also has a temporal decoupling nature that a BYY best harmony implementation of Eq. (37) with a first order Markovian temporal dependence can be decoupled into a summation of terms, with each term only involving a temporal dependence from $t-1$ to t. This nature comes from the temporal decoupling nature of $\ln[q(X|Y,L,\Theta_{X|YL})q(Y,L|\Theta_{YL})]$, which is not difficult to be extended into a finite order Markovian dependence among $\{y_t,\ell_t\}$. However, this nature is not applicable to $\mathrm{KL}(p\|q,\Theta)$, that is, the BYY best matching counterpart of $H(p||q,\Theta,\boldsymbol{k})$ in Eq. (34), because $\ln[p(Y|X,L,\Theta_L)p(L|X,\Theta_L)]$ does not have the same temporal decoupling nature.

5.3 Learning hidden Markov models (HMMs) and discriminative learning for HMM mixture

As illustrated at the center of Fig. 14(b), an HMM model considers a hierarchy of three levels, i.e., removing the top level in Fig. 12(b). The temporal dependence is considered on the third level with a sequence $L = \{l_t\}$ described by a joint distribution $q(L) = q(\ell_0)\prod_{t\geqslant 1} q(\ell_t|\ell_{t-1},Q)$ with a first order Markov nature, featured by a transfer probability matrix $Q=[q_{i|j}]$ with $q_{i|j} = q(\ell_t = i|\ell_{t-1}=j,Q)$. At each time t, a discrete random number l_t takes a fixed number of values, called states, while the observed sample x_t depends on the state that l_t takes. Given that $l_t = l$ takes a particular state, x_t becomes irrelevant to l_{t-1}. Specifically, it takes an emission probability $\alpha_{j|l}$ that the state l emits an output package $q(x_t|\theta_{j|l})$. Jointly, an observed sample x_t is emitted from the state l in a finite mixture as follows:

$$q(x_t|\theta_\ell) = \sum_j q(x_t\,|\theta_{j|\ell})\,\alpha_{j|\ell},$$

and usually we have $\theta_{j|l} = \theta_j$ without depending on the state l. The simplest case is $q(x_t|\theta_{j|l}) = \delta(x_t - a_j)$ with a_j denoting a label or color. In this case, the corresponding model is a classic HMM model. Widely applied to acoustic models in speech processing [105], one typical extension is Gaussian mixture based HMM by which $q(x_t|\theta_{j|l})$ is a Gaussian and thus each state l is associated with a Gaussian mixture [104]. Alternatively, this case can be regarded as an extension of adding a level $L=\{l_t\}$ on the top of the tree in Fig. 7(a). Recently, factor analysis have also been made on each local Gaussian with a structured covariance matrix [106]. This case can be regarded as an extension of adding another level $L=\{l_t\}$ on the top of local FA in Fig. 9(b).

Precisely, a first order HMM model is featured by a Ying machine that describes the distribution $q(X_N|\theta) = \sum_{L,J} q(X_N,J|L)\,q(L)$ and $q(X_N,J|L) = \prod_{t\geqslant 1} q(x_t\,|\theta_{j_t|\ell_t})\,\alpha_{j_t|\ell_t}$. Putting it into Eq. (37) and noticing that y is simply j and correspondingly that $q(y_t|y_{t-1},\ell_t,\theta_{y,\ell_t}) = q(j_t|\ell_t) = \alpha_{j_t|l_t}$ and $q(x_t|y_t,\ell_t,\theta_{x|y,\ell_t}) = q(x_t|\theta_{j_t|l_t})$, we get a special case of Eq. (37) as follows:

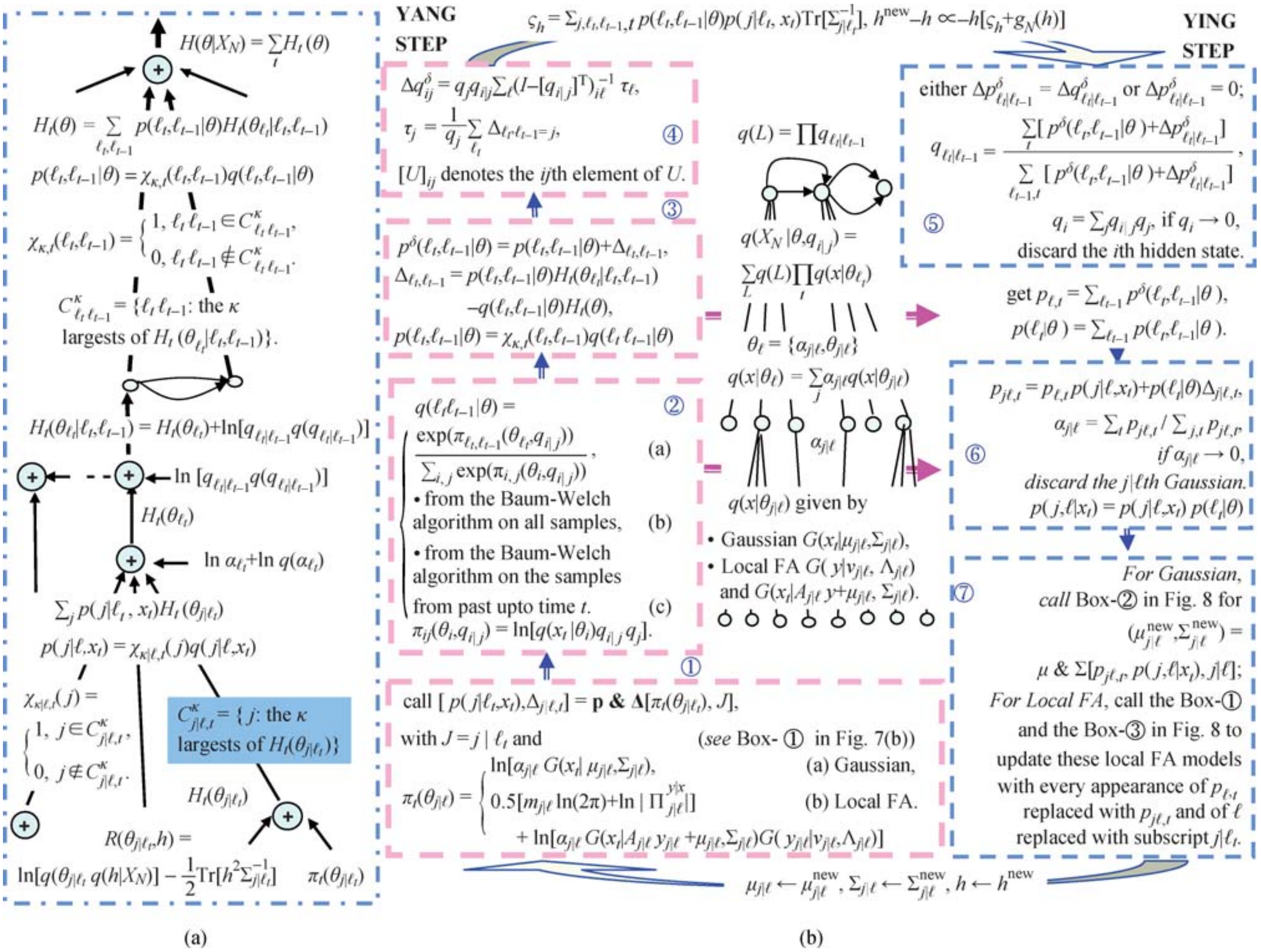
$$\begin{aligned}
&H(p||q,\theta,\boldsymbol{k}) = H(\theta\,|X_N) = \sum_t H_t(\theta),\\
&H_t(\theta) = \sum_{\ell_t,\ell_{t-1}} p(\ell_t,\ell_{t-1}|\theta)\,H_t(\theta_{\ell_t}\,|\ell_t,\ell_{t-1}),\\
&H_t(\theta_{\ell_t}\,|\ell_t,\ell_{t-1}) = H_t(\theta_{\ell_t}) + \ln\left[q_{\ell_t|\ell_{t-1}}q\left(q_{\ell_t|\ell_{t-1}}\right)\right],\\
&H_t(\theta_{\ell_t}) = \sum_j p(j\,|\ell_t,x_t)\,H_t\left(\theta_{j|\ell_t}\right) + \ln\alpha_{\ell_t} + \ln q(\alpha_{\ell_t}).
\end{aligned}\tag{63}$$

Considering the variety preservation principle and apex approximation by Eq. (38), the Yang machine is designed via designing the following two components:

1) $p(j_t|l_t,x_t) = \chi_{\kappa|l,t}(j_t)q(j_t|l_t,x_t)$ is designed from the Bayesian inverse of $q(j_t|l_t,x_t)$, as given in the bottom layer in Fig. 14(a). We have $\chi_{\kappa|l,t}(j_t)=1$ for ones in the apex zone, otherwise $\chi_{\kappa|l,t}(j_{t,})=0$. In other word, $p(j_t|l_t,x_t)$ is same as $q(j_t|l_t,x_t)$ in the apex zone.

2) $p(\ell_t,\ell_{t-1}|\theta) = p(l_t,l_{t-1}|X_N)$ is designed from the Bayesian inverse of $q(l_t,l_{t-1}|X_N)$. At time t, it means that the probability of occurring l_t is considered also in help of future sample x_t from t to N. This is possible when the entire set is available and considered, i.e., learning in batch. For the case that has only a set X_{0t} of samples up to present, we may consider $p(l_t,l_{t-1}|\theta)$ from the Bayesian inverse of $q(l_t,l_{t-1}|X_{0t})$. Both $q(l_t,l_{t-1}|X_N)$ and $q(l_t,l_{t-1}|X_{0t})$ are computationally quite involved and usually handled by the well known Baum-Welch algorithm [104]. If we further believe that the information of x_t from 0 to $t-1$ is already contained in l_{t-1}, one simplified design considers $p(l_t,l_{t-1}|\theta)$ from the Bayesian inverse of $q(l_t,l_{t-1}|x_{t,})$. In a summary, the Yang machine is designed from $q(l_t,l_{t-1}|\theta)$ of the Ying machine that has the following three typical choices:

$$q(\ell_t,\ell_{t-1}|\,\theta) = \begin{cases}
q(\ell_t,\ell_{t-1}|\,X_N) = \dfrac{q(\ell_t,\ell_{t-1},X_N)}{\sum_{\ell_t,\ell_{t-1}} q(\ell_t,\ell_{t-1},X_N)}, & \text{(a) entire set,}\\[2ex]
q(\ell_t,\ell_{t-1}|\,X_{0t}) = \dfrac{q(\ell_t,\ell_{t-1},X_{0t})}{\sum_{\ell_t,\ell_{t-1}} q(\ell_t,\ell_{t-1},X_{0t})}, & \text{(b) up to now,}\\[2ex]
q(\ell_t,\ell_{t-1}|\,x_t) = \dfrac{q(x_t\,|\theta_{\ell_t})\,q(\ell_t\,|\ell_{t-1})\,q(\ell_{t-1})}{\sum_{\ell_t,\ell_{t-1}} q(x_t\,|\theta_{\ell_t})\,q(\ell_t\,|\ell_{t-1})\,q(\ell_{t-1})}, & \text{(c) on sample.}
\end{cases}\tag{64}$$



$H(\theta|X_N)=\sum_t H_t(\theta)$
$H_t(\theta)=\sum_{\ell_t,\ell_{t-1}} p(\ell_t,\ell_{t-1}|\theta)H_t(\theta_{\ell_t}|\ell_t,\ell_{t-1})$
$p(\ell_t,\ell_{t-1}|\theta)=\chi_{\kappa,t}(\ell_t,\ell_{t-1})q(\ell_t,\ell_{t-1}|\theta)$
$\chi_{\kappa,t}(\ell_t,\ell_{t-1})=\begin{cases}1, & \ell_t\ell_{t-1}\in C^{\kappa}_{\ell_t\ell_{t-1}},\\ 0, & \ell_t\ell_{t-1}\notin C^{\kappa}_{\ell_t\ell_{t-1}}.\end{cases}$
$C^{\kappa}_{\ell_t\ell_{t-1}}=\{\ell_t\ell_{t-1}$: the κ largests of $H_t(\theta_{\ell_t}|\ell_t,\ell_{t-1})\}$.
$H_t(\theta_{\ell_t}|\ell_t,\ell_{t-1})=H_t(\theta_{\ell_t})+\ln[q_{\ell_t|\ell_{t-1}}q(q_{\ell_t|\ell_{t-1}})]$
$\ln[q_{\ell_t|\ell_{t-1}}q(q_{\ell_t|\ell_{t-1}})]$
$H_t(\theta_{\ell_t})$
$\ln\alpha_{\ell_t}+\ln q(\alpha_{\ell_t})$
$\sum_j p(j|\ell_t,x_t)H_t(\theta_{j|\ell_t})$
$p(j|\ell,x_t)=\chi_{\kappa|\ell,t}(j)q(j|\ell,x_t)$
$\chi_{\kappa|\ell,t}(j)=\begin{cases}1, & j\in C^{\kappa}_{j|\ell,t},\\ 0, & j\notin C^{\kappa}_{j|\ell,t}.\end{cases}$
$C^{\kappa}_{j|\ell,t}=\{j$: the κ largests of $H_t(\theta_{j|\ell_t})\}$
$H_t(\theta_{j|\ell_t})$
$R(\theta_{j|\ell_t},h)=$
$\ln[q(\theta_{j|\ell_t}q(h|X_N)]-\frac{1}{2}\mathrm{Tr}[h^2\Sigma^{-1}_{j|\ell_t}]$
$\pi_t(\theta_{j|\ell_t})$
(a)
YANG STEP
$\varsigma_h=\sum_{j,\ell_t,\ell_{t-1},t}p(\ell_t,\ell_{t-1}|\theta)p(j|\ell_t,x_t)\mathrm{Tr}[\Sigma^{-1}_{j|\ell_t}]$, $h^{\mathrm{new}}-h\propto -h[\varsigma_h+g_N(h)]$
YING STEP
$\Delta q^{\delta}_{ij}=q_j q_{i|j}\Sigma_\ell(I-[q_{i|j}]^{\mathrm{T}})^{-1}_{i\ell}\,\tau_\ell$,
$\tau_j=\frac{1}{q_j}\sum_{\ell_t}\Delta_{\ell_t,\ell_{t-1}=j}$,
④
$[U]_{ij}$ denotes the ijth element of U.
③
$p^{\delta}(\ell_t,\ell_{t-1}|\theta)=p(\ell_t,\ell_{t-1}|\theta)+\Delta_{\ell_t,\ell_{t-1}}$,
$\Delta_{\ell_t,\ell_{t-1}}=p(\ell_t,\ell_{t-1}|\theta)H_t(\theta_{\ell_t}|\ell_t,\ell_{t-1})$
$-q(\ell_t,\ell_{t-1}|\theta)H_t(\theta)$,
$p(\ell_t,\ell_{t-1}|\theta)=\chi_{\kappa,t}(\ell_t,\ell_{t-1})q(\ell_t\ell_{t-1}|\theta)$
②
$q(\ell_t\ell_{t-1}|\theta)=$
$\frac{\exp(\pi_{\ell_t,\ell_{t-1}}(\theta_{\ell_t},q_{i|j}))}{\sum_{i,j}\exp(\pi_{i,j}(\theta_i,q_{i|j}))}$, (a)
• from the Baum-Welch algorithm on all samples, (b)
• from the Baum-Welch algorithm on the samples from past upto time t. (c)
$\pi_{ij}(\theta_i,q_{i|j})=\ln[q(x_t|\theta_i)q_{i|j}q_j]$.
①
call $[p(j|\ell_t,x_t),\Delta_{j|\ell,t}]=\mathbf{p}\ \&\ \mathbf{\Delta}[\pi_t(\theta_{j|\ell_t}),J]$,
with $J=j\,|\,\ell_t$ and (see Box-① in Fig. 7(b))
$\pi_t(\theta_{j|\ell})=\begin{cases}\ln[\alpha_{j|\ell}G(x_t|\mu_{j|\ell},\Sigma_{j|\ell}), & \text{(a) Gaussian,}\\ 0.5[m_{j|\ell}\ln(2\pi)+\ln|\Pi^{y|x}_{j|\ell}|] & \text{(b) Local FA.}\\ \quad+\ln[\alpha_{j|\ell}G(x_t|A_{j|\ell}y_{j|\ell}+\mu_{j|\ell},\Sigma_{j|\ell})G(y_{j|\ell}|v_{j|\ell},\Lambda_{j|\ell})]\end{cases}$
$q(L)=\prod q_{\ell_t|\ell_{t-1}}$
$q(X_N|\theta,q_{i|j})=\sum_L q(L)\prod_t q(x|\theta_{\ell_t})$
$\theta_\ell=\{\alpha_{j|\ell},\theta_{j|\ell}\}$
$q(x|\theta_\ell)=\sum_j\alpha_{j|\ell}q(x|\theta_{j|\ell})$
$\alpha_{j|\ell}$
$q(x|\theta_{j|\ell})$ given by
• Gaussian $G(x_t|\mu_{j|\ell},\Sigma_{j|\ell})$,
• Local FA $G(y|v_{j|\ell},\Lambda_{j|\ell})$ and $G(x_t|A_{j|\ell}y+\mu_{j|\ell},\Sigma_{j|\ell})$.
either $\Delta p^{\delta}_{\ell_t|\ell_{t-1}}=\Delta q^{\delta}_{\ell_t|\ell_{t-1}}$ or $\Delta p^{\delta}_{\ell_t|\ell_{t-1}}=0$;
$q_{\ell_t|\ell_{t-1}}=\frac{\sum_t[p^{\delta}(\ell_t,\ell_{t-1}|\theta)+\Delta p^{\delta}_{\ell_t|\ell_{t-1}}]}{\sum_{\ell_{t-1},t}[p^{\delta}(\ell_t,\ell_{t-1}|\theta)+\Delta p^{\delta}_{\ell_t|\ell_{t-1}}]}$,
⑤
$q_i=\sum_j q_{i|j}q_j$, if $q_i\to 0$, discard the ith hidden state.
get $p_{\ell,t}=\sum_{\ell_{t-1}}p^{\delta}(\ell_t,\ell_{t-1}|\theta)$,
$p(\ell_t|\theta)=\sum_{\ell_{t-1}}p(\ell_t,\ell_{t-1}|\theta)$.
⑥
$p_{j\ell,t}=p_{\ell,t}p(j|\ell,x_t)+p(\ell_t|\theta)\Delta_{j|\ell,t}$,
$\alpha_{j|\ell}=\sum_t p_{j\ell,t}/\sum_{j,t}p_{j\ell,t}$,
if $\alpha_{j|\ell}\to 0$,
discard the $j|\ell$th Gaussian.
$p(j,\ell|x_t)=p(j|\ell,x_t)\,p(\ell_t|\theta)$
⑦
For Gaussian, call Box-② in Fig. 8 for $(\mu^{\mathrm{new}}_{j|\ell},\Sigma^{\mathrm{new}}_{j|\ell})=\mu\ \&\ \Sigma[p_{j\ell,t},p(j,\ell|x_t),j|\ell]$;
For Local FA, call the Box-① and the Box-③ in Fig. 8 to update these local FA models with every appearance of $p_{\ell,t}$ replaced with $p_{j\ell,t}$ and of ℓ replaced with subscript $j|\ell_t$.
$\mu_{j|\ell}\leftarrow\mu^{\mathrm{new}}_{j|\ell}$, $\Sigma_{j|\ell}\leftarrow\Sigma^{\mathrm{new}}_{j|\ell}$, $h\leftarrow h^{\mathrm{new}}$
(b)

Fig. 14 A unified Ying-Yang alternation procedure for three typical types of HMM models. (a) Hierarchical harmony flows; (b) main program

The choice (c) may be alternatively obtained from Eq. (58) by noticing that y is simply j and correspondingly that $q(y_t|y_{t-1},\ell_t,\theta_{y,\ell_t}) = q\,(j_t|\ell_t) = \alpha_{j_t|l_t}$ and $q(x_t|y_t,\ell_t,\theta_{x|y,\ell_t}) = q(x_t|\theta_{j_t|l_t})$. We are again lead to Eq. (64) but only with $p(l_t,l_{t-1}|\theta)$ in the form of $p(l_t,l_{t-1}|x_t,)$. Here, we have two different perspectives of interpretation. The one bases on Eq. (37) that comes from Eq. (34) with the integral over X approximately removed first, while the second bases on Eq. (58) that comes from Eq. (34) with the Markovian temporal dependence considered first.

Further considering apex approximation by Eq. (38), on the top layer in Fig. 14(a) we get $p(l_t,l_{t-1}|\theta)$ from $q(l_t,l_{t-1}|\theta)$ in one of three choices in Eq. (64), with the apex zone consisting of κ best candidates for transferring $l_{t-1}\to l_t$. The maximization of $H(p||q,\Theta,k)$ is implemented by the algorithm given in Fig. 14(b), which comes from getting

$$
\begin{aligned}
&\nabla_\varphi H_t(\theta)\\
&\quad=\sum_{\ell_t,\ell_{t-1}}\big[p\left(\ell_t,\ell_{t-1}|\theta\right)\nabla_\phi H_t\left(\theta_{\ell_t}|\,\ell_t,\ell_{t-1}\right)\\
&\qquad+H_t\left(\theta_{\ell_t}|\,\ell_t,\ell_{t-1}\right)\chi_{\kappa,t}\left(\ell_t,\ell_{t-1}\right)\nabla_\phi q\left(\ell_t,\ell_{t-1}|\,\theta\right)\big].
\end{aligned}
$$

When $q(l_t,l_{t-1}|\theta)$ in the first two choices of Eq. (64), it is quite involved and costly to compute its gradient. For simplicity, we get $\nabla_\varphi q(l_t,l_{t-1}|\theta)$ via $q(l_t,l_{t-1}|\theta)$ in choice (c), while $p(l_t,l_{t-1}|\theta)$ can be any one of the three cases. For a better understanding on Box-② and Box-③ in Fig. 14(b), one may observe that the counter part of Eq. (39) becomes

$$
\begin{aligned}
&\sum_{\ell_t,\ell_{t-1}} H_t\left(\theta_{\ell_t}|\,\ell_t,\ell_{t-1}\right)\chi_{\kappa,t}\left(\ell_t,\ell_{t-1}\right)\nabla_\varphi q\left(\ell_t,\ell_{t-1}|\,\theta\right)\\
&\quad=\sum_{\ell_t,\ell_{t-1}}\Delta_{\ell_t,\ell_{t-1}}\nabla_\varphi\pi_{\ell_t,\ell_{t-1}}\left(\theta_{\ell_t},q_{\ell_t|\ell_{t-1}}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
&\nabla_\varphi\pi_{\ell_t,\ell_{t-1}}\left(\theta_{\ell_t},q_{\ell_t|\ell_{t-1}}\right)\\
&\quad=\nabla_\varphi\ln q\left(x\,|\theta_\ell\right)+\nabla_\varphi\ln q_{\ell_t|\ell_{t-1}}+\nabla_\varphi\ln q_{\ell_{t-1}},\\
&\nabla_\varphi\ln q\left(x\,|\theta_\ell\right)=\sum_j p\left(j\,|\ell,x_t\right)\nabla_\varphi\pi_t\left(\theta_{j|\ell}\right),\\
&\nabla_\varphi\pi_t\left(\theta_{j|\ell}\right)=\nabla_\varphi\ln\alpha_{j|\ell}+\nabla_\varphi\ln q\left(x\,|\theta_{j|\ell}\right).
\end{aligned}
$$

Similar to the ones in Eq. (55) for the bottom layer in Fig. 12, the equation of $\nabla_\varphi\pi_{\ell_t,\ell_{t-1}}$ corresponds to the bottom layer in Fig. 14(b). Specifically, the Box-① p & Δ gets the Bayesian posterior $p(j|l,x_t)$ and $\Delta_{j|l,t}$, which are combined with $p_{l,t}$ and $p(l_t|\theta)$ from the upper layer to call the Box-⑥ for updating $\alpha_{j|l}$ and the Box-⑦ for updating $q(x_t|\theta_{j|l})$.

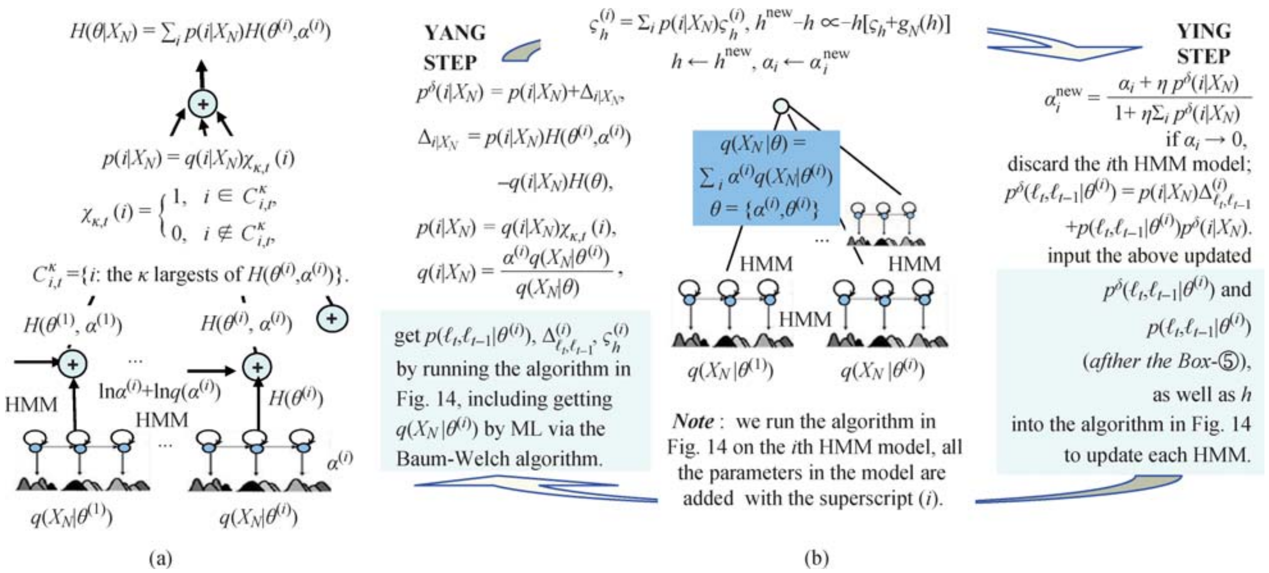
Being different from Fig. 12, the first line in Eq. (65) includes temporal dependence between l_t, l_{t-1}, for which each term in the summation for $H_t(\theta)$ needs considering l_t, l_{t-1} jointly, as given on the top of Fig. 14(a). Specifically, the Box-② implements Eq. (64), and the Box-③ gets $p^{\delta}(l_t, l_{t-1}|\theta)$. Moreover, the Box-④ computes a correcting term Δq^{δ}_{ij} used in the Box-⑤ for updating Q, which comes from approximately considering $\nabla_{\varphi} \ln q_{\ell_{t-1}}$ by $\nabla_{\varphi} \ln q_j$, see Appendix C 4)e). For simplicity, we may also choose to ignore it simply by letting $\Delta q^{\delta}_{ij} = 0$. Following the Box-⑤, we get $p_{l,t}$ and $p(l_t|\theta)$ and send them downward to the Box-⑥ for updating $\alpha_{j|l}$ and the Box-⑦ for updating $q(x_t|\theta_{j|l})$.

Again, setting $\Delta_{jli,t}$=0 and $\Delta_{l_t l_{t-1}}$=0, the Ying-Yang iteration by Fig. 14(b) actually becomes equivalent to perform the Baum-Welch algorithm and the EM algorithm for the maximum likelihood learning on the HMM model. With the correcting terms $\Delta_{jli,t} \neq 0$ and $\Delta_{l_t l_{t-1}} \neq 0$, the Ying-Yang iteration implements the BYY harmony learning, during which the number of states and the number of components in the bottom layer is determined by automatic model selection. That is, those extra ones are discarded if the corresponding $q_i \to 0$ and $\alpha_{i|l} \to 0$, respectively.

In the degenerated case that l takes one value only, i.e., there is only one state, the Box-②, the Box-③, the Box-④ and the Box-⑤ are all degenerated to be no effect with $p_{l,t} = 1$ and $p(l_t|\theta) = 1$, while the Box-①, the Box-⑥, and the Box-⑦ jointly become equivalent to the algorithms in Fig. 7(b) and Fig. 9(b). Moreover, the Box-⑦ can be extended to cover those extensions of FA discussed in Sect. 3.2.

Learning HMM model in Fig. 14 can be further extended to a mixture of multiple HMM models as shown in Fig. 15(a). This is equivalent to adding one top node on the Fig. 14(a), which consists of a number of descendants. Each descendant is one HMM model $q(X_N|\theta^{(i)})$ as shown in the center of Fig. 14(b) for describing X_N, with the superscript (i) added to its parameter set θ for a notation purpose, i.e., we get $\theta^{(i)}$. Correspondingly, we are lead to a three layer hierarchical tree that can be regarded as an extension of the hierarchical tree shown in the center of Fig. 12(b), with the ith node of the middle layer extended to covering temporal dependence $q(l_t|l_{t-1})$. Similar to the top layer in Fig. 12(a), the harmony flow $H(\theta|X_N)$ about the added top layer is given on the top of Fig. 15(a). The counterpart of updating equations for the top layer in Fig. 12(b) is given in the upper layer of Fig. 15(b). One difference is that adaptive updating equation is provided for updating α_i, since the batch way needs to be made over a number of different sample sets instead of just one set X_N. It is very time consuming and also practically not available. Adaptive updating avoids this problem, it modifies α_i each time as $p^{\delta}(i|X_N)$ is updated by a small step size $\gamma > 0$.

Setting $\Delta_{i|X} = 0$, $\Delta_{jli,t} = 0$ and $\Delta_{l_t l_{t-1}} = 0$, the Ying-Yang iteration by Fig. 15(b) becomes equivalent to perform the Baum-Welch algorithm for the maximum likelihood learning. With $\Delta_{i|X} \neq 0$, $\Delta_{jli,t} \neq 0$ and $\Delta_{l_t l_{t-1}} \neq 0$, the Ying-Yang iteration implements the BYY harmony learning, during which not only those extra states and extra components in the bottom layer are discarded as the corresponding $q_i \to 0$ and $\alpha_{i|l} \to 0$, respectively, but also an extra HMM model is discarded if the corresponding $\alpha_i \to 0$. Moreover, it follows Eq. (A.2) (see Appendix A) that maximizing $H(\theta|X_N)$ is equivalent to maximizing



$H(\theta|X_N) = \sum_i p(i|X_N)H(\theta^{(i)},\alpha^{(i)})$
$p(i|X_N) = q(i|X_N)\chi_{\kappa,t}(i)$
$\chi_{\kappa,t}(i) = \begin{cases} 1, & i \in C^{\kappa}_{i,t}, \\ 0, & i \notin C^{\kappa}_{i,t}, \end{cases}$
$C^{\kappa}_{i,t}$={i: the κ largests of $H(\theta^{(i)},\alpha^{(i)})$}.
$H(\theta^{(1)}, \alpha^{(1)})$
$H(\theta^{(i)}, \alpha^{(i)})$
HMM
$\ln\alpha^{(i)}+\ln q(\alpha^{(i)})$
HMM
$H(\theta^{(i)})$
$\alpha^{(i)}$
$q(X_N|\theta^{(1)})$
$q(X_N|\theta^{(i)})$
(a)
YANG STEP
$\varsigma^{(i)}_h = \Sigma_i p(i|X_N)\varsigma^{(i)}_h$, $h^{\text{new}} - h \propto -h[\varsigma_h + g_N(h)]$
$h \leftarrow h^{\text{new}}$, $\alpha_i \leftarrow \alpha_i^{\text{new}}$
YING STEP
$p^{\delta}(i|X_N) = p(i|X_N)+\Delta_{i|X_N}$,
$\Delta_{i|X_N} = p(i|X_N)H(\theta^{(i)},\alpha^{(i)}) - q(i|X_N)H(\theta)$,
$p(i|X_N) = q(i|X_N)\chi_{\kappa,t}(i)$,
$q(i|X_N) = \dfrac{\alpha^{(i)}q(X_N|\theta^{(i)})}{q(X_N|\theta)}$,
get $p(\ell_t,\ell_{t-1}|\theta^{(i)})$, $\Delta^{(i)}_{\ell_t,\ell_{t-1}}$, $\varsigma^{(i)}_h$ by running the algorithm in Fig. 14, including getting $q(X_N|\theta^{(i)})$ by ML via the Baum-Welch algorithm.
$q(X_N|\theta) = \Sigma_i \alpha^{(i)}q(X_N|\theta^{(i)})$
$\theta = \{\alpha^{(i)},\theta^{(i)}\}$
HMM
HMM
HMM
$q(X_N|\theta^{(1)})$
$q(X_N|\theta^{(i)})$
Note : we run the algorithm in Fig. 14 on the ith HMM model, all the parameters in the model are added with the superscript (i).
$\alpha_i^{\text{new}} = \dfrac{\alpha_i + \eta\, p^{\delta}(i|X_N)}{1+\eta\Sigma_i p^{\delta}(i|X_N)}$
if $\alpha_i \to 0$, discard the ith HMM model;
$p^{\delta}(\ell_t,\ell_{t-1}|\theta^{(i)}) = p(i|X_N)\Delta^{(i)}_{\ell_t,\ell_{t-1}} + p(\ell_t,\ell_{t-1}|\theta^{(i)})p^{\delta}(i|X_N)$.
input the above updated $p^{\delta}(\ell_t,\ell_{t-1}|\theta^{(i)})$ and $p(\ell_t,\ell_{t-1}|\theta^{(i)})$ (*afther the Box-⑤*), as well as h into the algorithm in Fig. 14 to update each HMM.
(b)

Fig. 15 Discriminative learning of multiple HMM models

$$\sum_{i} p\left(i|X_N\right) \ln p\left(i|X_N\right) + \ln q\left(X_N|\theta_N\right)$$

that consists of the second term for maximizing likelihood on this mixture of multiple HMM models and the first term for making these HMM models become more discriminative. In other words, the Ying-Yang iteration by Fig. 15(b) actually provides a discriminative learning for multiple HMM models, which is an alternative solution for discriminative learning on HMM based acoustic models in speech processing [107,108].

6 Summarizing remarks and further topics

An intelligent system is featured with two types of intelligent abilities, namely, Type I from inside to outside or top down for its knowledge about the world and Type II from outside to inside or bottom up for solving problem and adapting the system itself. To be specific, Type II deals with three levels of nested inverse problems, consisting of $X_N \to Y, L$ for perception, encoding, and solving problem, $X_N \to \Theta$ for parameter learning, and $X_N \to \boldsymbol{k}$ for model selection, which are summarized as a mapping $X \to R$ with $R = \{Y, L, \Theta, \boldsymbol{k}\}$; while Type I deals with $R \to X$. In a probability theory framework, two mappings are described in a joint distribution of X, R in two types of Bayesian decompositions $q(X|R)q(R)$ and $p(R|X)p(X)$, which can be understood from a perspective of ancient Chinese Ying-Yang philosophy. We have $p(X, R) = p(R|X)p(X)$ as Yang machine, and $q(X, R) = q(X|R)q(R)$ as Ying machine, and call this pair the BYY system.

To build up a BYY system, we first design its architecture under a guideline that Ying machine accommodates inner representations and generates reconstructions to fit observed data via structures with least redundancy, while the Yang is designed according to the principle of variety preservation for a Ying-Yang balance, in a compliment to Proposition 2 in Appendix B. First, $q(R)$ is considered according to the representation form of $R = \{Y, L, \Theta, \boldsymbol{k}\}$ in three layers as shown in Fig. 6(b). Specifically, Y, L depend on the natures of tasks (e.g., clustering, encoding, feature extraction, etc.). Also, a temporal and hierarchical dependence among observations is accommodated via a corresponding structure in Y, L, e.g., as shown in Figs. 12–14. Both Θ and $\boldsymbol{k}$ depend on the parametric structures used in the system, based on pre-knowledge about data clouds X_N. Next, $q(X|R)$ is designed for an appropriate mapping from R to fit X_N, in a structure that consists of a set of individual simple structures in a simple combination. Finally, the Yang machine is designed with input data X_N as $p(X)$, and $p(R|X)$ in a structure that preserves the dynamism or variety of R by the Ying machine to maintain a balanced information flow from Yang to Ying.

A BYY system operates in term of a circling flow $X_N \to R$ with $(Y, L \to \Theta \to \boldsymbol{k}) \to X_N$ featured by three nested levels of five action circling, namely, acquisition → assumption → accumulation and amalgamation → apex-seeking → affirmation, which was previously proposed under the name of A5 paradigm [77]. Interestingly, it coincides with the famous ancient Chinese WuXing theory [53]. Also, improving HT by RHT and making $\max_\Theta F(\Theta)$ by gradient based line search are even interesting examples of changing A-2 to improve a bottleneck of A-4, which well coincides with the Ke-Cheng-Hui law.

Generally, the WuXing theory provides a general guide to keep the A5 circling within an intelligent system well balanced, in a sense that each action should be neither too weak to sustain the system nor too loaded to jam the circling. Specifically, we consider to pick samples adaptively versus in a batch by A-1, to make apex approximation by A-2, to conduct an effective searching or evidence integrating (e.g., primal gradient flow) by A-3, to well detect a peak or a convergence by A-4, and to get a reliable verification by A-5.

All together, learning all the unknowns in a BYY system and implementing all the levels of A5 circling are governed under the principle of Ying-Yang best harmony. This principle is mathematically implemented by maximizing a harmony functional $H(p||q)$ that tends to $p = q$ with q in a most compact form, which is different from the minimization of the well known KL divergence $\mathrm{KL}(p||q)$ that only ensures the tendency towards $p = q$. This BYY harmony learning provides a new road that leads to improved model selection criteria, Ying-Yang alternating algorithms with automatic model selection, and a coordinated implementation of Ying based model selection and Yang based data smoothing regularization.

Maximizing $H(p||q)$ versus minimizing $\mathrm{KL}(p||q)$ are different but closely related, which can be understood from an number of perspectives. Both maximizing $H(p||q)$ and minimizing $\mathrm{KL}(p||q)$ can be represented as special cases of Radon-Nikodym theorem based harmony functional by Eq. (21). Moreover, as shown in Fig. A2 and further discussed in Appendix B, a BYY system with a Ying-Yang best matching by minimizing $\mathrm{KL}(p||q)$ provides a unified framework for typical existing learning methods, including not only the best-inner-encoding intended stream of minimum mutual information or maximum information transfer studies, but also the best-data-matching intended stream of maximum likelihood and marginal likelihood based studies (e.g., ML, AIC, BIC, MDL, etc.), as well as their further progresses in help of variational approximation methods. Also, an approximate implementation of the second stream leads to the maximum a posteriori (MAP) based Bayes learning and MML. In addition, all these studies can be extended in help with a data smoothing regularization. On the other hand, a BYY system with a Ying-Yang best harmony by maximizing $H(p||q)$ provides a framework of new approaches with a favorable new mechanism for model selection. Also, it shares some common special cases with the ones of Ying-Yang best matching, e.g., maximum marginal likelihood based studies (e.g., AIC, BIC, MDL, etc.), the MAP based Bayes learning and MML, etc.

Learning for the unknowns in the Yang machine and Ying machine can be decoupled by alternatively updating one with the other fixed, which provides a Ying-Yang alternation procedure that summarizes not only algorithms for implementing BYY harmony learning but also the EM algorithm for the maximum likelihood learning, as well as RPCL learning and MAP based competitive learning in a unified procedure with most parts in a same expression while options elected in a few setting choices.

On one hand, we believe that the Chinese ancient Yin-Yang and WuXing Meta theories provide guidelines for modern information theoretic studies. On the other hand, we regard that Bayesian Ying-Yang learning proposed in Sect. 4, especially the Radon-Nikodym theorem based harmony functional by Eq. (21) as a mathematical formulation of the ancient Yin-Yang philosophy from an information theoretic perspective. Moreover, it follows from $H_\mu(P||P)$ by Eq. (22) that information may also have a more general interpretation. Given a σ-finite measure P that describes how a resource distributes over the space (X, Σ) relatively with respect to the Lebesgue measure $\mu, \mathrm{d}P/\mathrm{d}\mu$ represents the density of P measure on a piece $\mathrm{d}\mu, H_\mu(P||P)$ describes the concentration or compactness of this distribution. Thus, information is the negation of the compactness of this distribution configuration.

The last but not least, for future directions of studies, we list ten further topics as follows:

1) Extensive studies have been made on the KL divergence $\mathrm{KL}(p||q)$ from the perspective of geometry [109] and then differential geometry with its metrics defined by Fisher information matrix [22]. $H(p||q)$ can be regarded as an evolution from the concepts about intersection and inner product but actually not an inner product because symmetrical nature disappeared. As argued in Ref. [49], it still can be regarded as a geometry concept about projection from one manifold to the other. It would be interested to study the natures of differential flow of this projection as discussed in Fig. 10, especially on how two manifolds shrink towards appropriate volumes such that automatic model selection emerges. It would be also interested to develop a new convergence analyzing tool for a quite unique feature. As illustrated in Fig. 5(b), the manifold shrinking will make $H(p||q)$ tend infinity. However, this diverging should not be regarded as a bad thing but as indicators that some related dimensions should be discarded.

2) For the family of harmony functional by Eq. (21), only the one with $f(r) = \ln r$ has the favorable separable nature. Explorations can be made on further justifications and also on comparison with the harmony functional by Eq. (A.1) (see Appendix A).

3) As discussed in Sect. 4.2, a Yang machine is designed under a variety preservation principle by Eq. (27), with a scale parameter $\rho \geqslant 0$. One extreme $\rho = 0$ leads to $p(R|X) = \delta(R - R^*)$, while the other extreme with a large enough ρ leads to $p(R|X) = q(R|X)$ for every R. There is a spectrum of choices between the two extremes. Further explorations can be made on this spectrum and also on what is an appropriate ρ.

4) As discussed in Sect. 2.2 and Fig. 5(a), the maximum likelihood learning has an invariant nature with respect to extra number of hidden factors and their parameterization due to marginalization. This nature is a special case of a bottom-up hierarchical decoupling nature owned by $\mathrm{KL}(p||q)$ based BYY best matching as discussed in Sect. 5.2, but not shared by the BYY harmony functional. Instead, the latter is sensitive to extra number of hidden factors, which is favorable to model selection. It deserves to reexamine those models with different parameterizations that are usually regarded as equivalent from a maximum likelihood sense. E.g., a study has found the conventional parameterization for factor analysis [82], though widely adopted in the literature, is inferior to the parameterization in Fig. 1(c).

5) The approximation of $H(p||q, \boldsymbol{k}, \Xi)$ in Fig. 5(b) is made in help of a peak convex analysis by Eq. (35) around Θ^* that is obtained with $q(\Theta^a)$ included. Also, we get $d_k(\Theta|\Xi)$ in Eq. (36) with $\Omega(\Theta^*)$ around Θ^*. Alternatively, a peak convex analysis by Eq. (35) may also be made around the following Θ^m obtained without $q(\Theta^a)$ included.

$$\Theta^m = \arg\max_{\Theta} \ln \left[q\left(X \left| Y, \theta_{X|Y} \right.\right) q\left(Y \left| \theta_{\boldsymbol{k}_Y}, \boldsymbol{k}_Y \right.\right) \right].$$

Further explorations can be made on whether two different ways makes differences, and whether different partitions of $\Theta = \Theta^a \cup \Theta^b$ makes differences, as well as on what are favorable features of the induced bias cancelation priors by Eqs. (32) and (33) in comparison with Jeffreys prior [97].

6) To implement automatic model selection by detecting Eq. (4), further explorations can be made on a statistical testing $\mathrm{H}_0 : \rho(\theta_l) = 0$ and $\mathrm{H}_1 : \rho(\theta_l) \neq 0$ via appropriate probability distributions, e.g., a Dirichlet distribution for α_l and a gamma distribution for λ_l. Moreover, investigations are also needed on how the performances are affected by local optimum problems.

7) In the cases that the inner representation consists of $Y = \{y\}$ of real vectors, we removed the integral over y in help of Eq. (35) and get Eq. (37) that

includes a term $\mathrm{Tr}\big[\Gamma_L^{Y|X}\Pi_L^{Y|X}\big] = \mathrm{Tr}\big[I\big] + \mathrm{Tr}\big[\rho^2\Pi_L^{Y|X}\big]$ for $\Gamma_L^{Y|X} = \Pi_L^{Y|X\,-1} + \rho^2$. The term $\mathrm{Tr}[I]$ becomes an integer for the dimension of y that however has no help on automatic model selection via a gradient based updating, e.g., in the algorithms in Figs. 8 and 9. Alternatively, we may investigate whether the information about this term can be better used by switching the order, i.e., making the operation ∇_Θ first and then approximating the integral over Y.

8) To tackle the problems of updating α_l that should be nonnegative and Σ_l that should be nonnegative definite, one way is making the constraints satisfied in help of techniques introduced in Refs. [49,72,80] with extra computing costs. Alternatively, in this paper they are updated by the Ying step (e.g., in Figs. 7 and 8) in a batch way similar to the EM algorithm without ensuring the constraints. When $\Delta_{i,t} = 0$, as discussed after Eq. (7), it degenerates back to the EM algorithm that guarantees the constraints. However, there lacks such a guarantee when $\Delta_{i,t} \neq 0$, which needs a further investigated.

9) In help of the BYY system, semi-unsupervised learning can be simply embedded into the Yang machine, typically via the linear combination in Eq. (42) and Eq. (43). Further studies may be conducted on which types of a priori $q(\varrho)$ is appropriate to this purpose. Also, we may consider other combining rules used for combining classifiers and learning mixture-of-experts [102].

10) For temporal BYY harmony learning, the design of Yang machine encounters the three choices in Eq. (64), we have two different perspectives on the choices. One prefers the choice (a) based on Eq. (37) that comes from Eq. (34) with the integral over X approximately removed first, while the second prefers the choice (c) based on Eq. (58) that comes from Eq. (34) with the Markovian temporal dependence considered first. A further comparative study may be explored. Also, a further study may be made on how to select an appropriate apex approximation with the κ best candidates for $l_{t-1} \to l_t$.

Appendix A Harmony functional versus Kullback-Leibler divergence

Shown in Fig. A1(a) are two streams of the evolutions on the two related concepts:

1) The harmony functional $H(p||q)$ by Eq. (24) can be regarded as evolution from the concept of inner product to a probability space. The inner product in the Hilbert space has been widely used in the literature of statistics and signal processing for measuring correlation and thus also called correlation function, while the inner product in an Euclidean space of normalized vectors has been widely used as a measure of similarity. Moreover, the inner product is involved from the intersection $P \cap Q$ of two sets P and Q for expressing the concept of common or agreement. The concepts of maximizing harmony, correlation, and similarity all join together to form one evolution stream originated from the concept of seeking common points or mutual agreement.

2) The Kullback-Leibler (KL) divergence $\mathrm{KL}(p||q) = \min$ and also $\mathrm{KL}(p||q) + \mathrm{KL}(q||p) = \min$ is involved from the orthogonal nature and the least square error in the Euclidean and Hilbert space, respectively, and further from the least difference concepts about two sets. We have $\mathrm{KL}(p||q) + \mathrm{KL}(q||p) = 0$ if and only if $\mathrm{KL}(p||q) = 0$ and the square error is least if and only if x, y are orthogonal, while $\#\ P\Delta Q = 0$ if both $\#(Q - P) = 0$ and $\#(P - Q) = 0$.

The first stream seeks maximizing agreements while the second stream seeks minimizing disagreements. Moreover, the second stream is actually defined from the concept of the first stream with more restrictions. In other words, the first stream can be regarded as more fundamental than the second stream, which concurs with the discussions made after Eq. (24). In a view of geometry [49(Sect. III)], the first one seeks a maximum projection of the manifold q onto the manifold p, while the second one minimizes the residual that can not be projected. As listed in Figs. A1(b) and A1(c), the original concepts of two steams are equivalent in the set theory. After evolutions they become different and lead us to the following three scenarios:

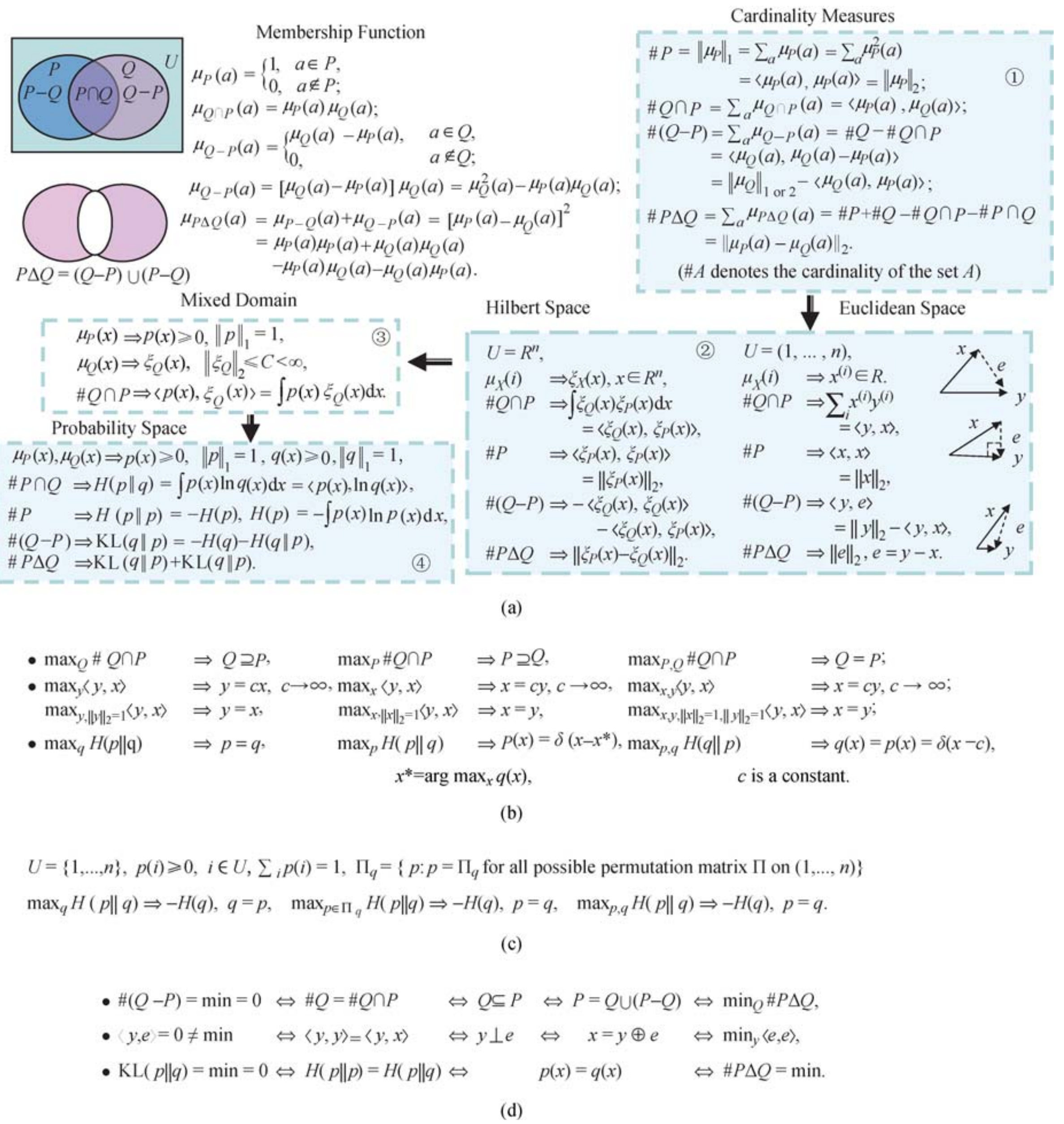
1) The equivalence is kept under certain constraints. In the Euclidean and Hilbert space, the equivalence holds with the L_2 norm $\|\cdot\|_2 = 1$. While in the probability space, one scenario is that the optimization is made over q with p fixed, the other scenario is shown in Fig. A1(c), i.e., the optimization is made over q, p within their permutation set Π_q.

2) The first stream leads to a relaxed relation than the one by the second stream. In the Euclidean and Hilbert space, as shown in Fig. A1(b), minimizing the squared residuals makes two become equal is relaxed to that maximizing the inner product makes two become equal up to a unknown scale.

3) The first stream leads to a more strict relation than the one by the second stream. In the probability space, minimizing $\mathrm{KL}(p||q)$ for $p = q$ is strengthen into maximizing $H(p||q)$ that makes not only $p = q$ and but also both towards a most compacted form, in other words, towards a deterministic agreement as possible. Such a situation may also occur by introducing certain constrain in the Euclidean and Hilbert space. As in the Box-③, we have

$$H\,(p||q) = \langle p(x),\ \xi_Q(x)\rangle = \int p(x)\xi_Q(x)\mathrm{d}x,$$

$$\text{subject to } p(x) \geqslant 0, \int p(x)\mathrm{d}x = 1, \text{ and}$$


Membership Function
P Q U
$P-Q$ $P\cap Q$ $Q-P$
$\mu_P(a)=\begin{cases}1, & a\in P,\\ 0, & a\notin P;\end{cases}$
$\mu_{Q\cap P}(a)=\mu_P(a)\mu_Q(a);$
$\mu_{Q-P}(a)=\begin{cases}\mu_Q(a)-\mu_P(a), & a\in Q,\\ 0, & a\notin Q;\end{cases}$
$\mu_{Q-P}(a)=[\mu_Q(a)-\mu_P(a)]\,\mu_Q(a)=\mu_Q^2(a)-\mu_P(a)\mu_Q(a);$
$\mu_{P\Delta Q}(a)=\mu_{P-Q}(a)+\mu_{Q-P}(a)=[\mu_P(a)-\mu_Q(a)]^2=\mu_P(a)\mu_P(a)+\mu_Q(a)\mu_Q(a)-\mu_P(a)\mu_Q(a)-\mu_Q(a)\mu_P(a).$
$P\Delta Q=(Q-P)\cup(P-Q)$
Cardinality Measures
$\#P=\|\mu_P\|_1=\sum_a\mu_P(a)=\sum_a\mu_P^2(a)=\langle\mu_P(a),\mu_P(a)\rangle=\|\mu_P\|_2;$ ①
$\#Q\cap P=\sum_a\mu_{Q\cap P}(a)=\langle\mu_P(a),\mu_Q(a)\rangle;$
$\#(Q-P)=\sum_a\mu_{Q-P}(a)=\#Q-\#Q\cap P=\langle\mu_Q(a),\mu_Q(a)-\mu_P(a)\rangle=\|\mu_Q\|_{1\text{ or }2}-\langle\mu_Q(a),\mu_P(a)\rangle;$
$\#P\Delta Q=\sum_a\mu_{P\Delta Q}(a)=\#P+\#Q-\#Q\cap P-\#P\cap Q=\|\mu_P(a)-\mu_Q(a)\|_2.$
($\#A$ denotes the cardinality of the set A)
Mixed Domain
$\mu_P(x)\Rightarrow p(x)\geqslant 0,\ \|p\|_1=1,$ ③
$\mu_Q(x)\Rightarrow\xi_Q(x),\ \|\xi_Q\|_2\leqslant C<\infty,$
$\#Q\cap P\Rightarrow\langle p(x),\xi_Q(x)\rangle=\int p(x)\,\xi_Q(x)\mathrm{d}x.$
Hilbert Space
Euclidean Space
$U=R^n,$ ② $U=(1,\ldots,n),$
$\mu_X(i)\Rightarrow\xi_X(x),\ x\in R^n,$ $\mu_X(i)\Rightarrow x^{(i)}\in R.$
$\#Q\cap P\Rightarrow\int\xi_Q(x)\xi_P(x)\mathrm{d}x=\langle\xi_Q(x),\xi_P(x)\rangle,$ $\#Q\cap P\Rightarrow\sum_i x^{(i)}y^{(i)}=\langle y,x\rangle,$
$\#P\Rightarrow\langle\xi_P(x),\xi_P(x)\rangle=\|\xi_P(x)\|_2,$ $\#P\Rightarrow\langle x,x\rangle=\|x\|_2,$
$\#(Q-P)\Rightarrow-\langle\xi_Q(x),\xi_Q(x)\rangle-\langle\xi_Q(x),\xi_P(x)\rangle,$ $\#(Q-P)\Rightarrow\langle y,e\rangle=\|y\|_2-\langle y,x\rangle,$
$\#P\Delta Q\Rightarrow\|\xi_P(x)-\xi_Q(x)\|_2.$ $\#P\Delta Q\Rightarrow\|e\|_2,\ e=y-x.$
x e y
Probability Space
$\mu_P(x),\mu_Q(x)\Rightarrow p(x)\geqslant 0,\ \|p\|_1=1,\ q(x)\geqslant 0,\|q\|_1=1,$
$\#P\cap Q\Rightarrow H(p\|q)=\int p(x)\ln q(x)\mathrm{d}x=\langle p(x),\ln q(x)\rangle,$
$\#P\Rightarrow H(p\|p)=-H(p),\ H(p)=-\int p(x)\ln p(x)\mathrm{d}x,$
$\#(Q-P)\Rightarrow\mathrm{KL}(q\|p)=-H(q)-H(q\|p),$
$\#P\Delta Q\Rightarrow\mathrm{KL}(q\|P)+\mathrm{KL}(q\|p).$ ④

(a)

- $\max_Q\#Q\cap P\Rightarrow Q\supseteq P,$ $\max_P\#Q\cap P\Rightarrow P\supseteq Q,$ $\max_{P,Q}\#Q\cap P\Rightarrow Q=P;$
- $\max_y\langle y,x\rangle\Rightarrow y=cx,\ c\to\infty,$ $\max_x\langle y,x\rangle\Rightarrow x=cy,\ c\to\infty,$ $\max_{x,y}\langle y,x\rangle\Rightarrow x=cy,\ c\to\infty;$
  $\max_{y,\|y\|_2=1}\langle y,x\rangle\Rightarrow y=x,$ $\max_{x,\|x\|_2=1}\langle y,x\rangle\Rightarrow x=y,$ $\max_{x,y,\|x\|_2=1,\|y\|_2=1}\langle y,x\rangle\Rightarrow x=y;$
- $\max_q H(p\|q)\Rightarrow p=q,$ $\max_p H(p\|q)\Rightarrow P(x)=\delta(x-x^*),$ $\max_{p,q}H(q\|p)\Rightarrow q(x)=p(x)=\delta(x-c),$
  $x^*=\arg\max_x q(x),$ c is a constant.

(b)

$U=\{1,\ldots,n\},\ p(i)\geqslant 0,\ i\in U,\ \sum_i p(i)=1,\ \Pi_q=\{p: p=\Pi_q$ for all possible permutation matrix Π on $(1,\ldots,n)\}$

$\max_q H(p\|q)\Rightarrow -H(q),\ q=p,\quad \max_{p\in\Pi_q}H(p\|q)\Rightarrow -H(q),\ p=q,\quad \max_{p,q}H(p\|q)\Rightarrow -H(q),\ p=q.$

(c)

- $\#(Q-P)=\min=0\Leftrightarrow\#Q=\#Q\cap P\Leftrightarrow Q\subseteq P\Leftrightarrow P=Q\cup(P-Q)\Leftrightarrow\min_Q\#P\Delta Q,$
- $\langle y,e\rangle=0\neq\min\Leftrightarrow\langle y,y\rangle=\langle y,x\rangle\Leftrightarrow y\perp e\Leftrightarrow x=y\oplus e\Leftrightarrow\min_y\langle e,e\rangle,$
- $\mathrm{KL}(p\|q)=\min=0\Leftrightarrow H(p\|p)=H(p\|q)\Leftrightarrow p(x)=q(x)\Leftrightarrow\#P\Delta Q=\min.$

(d)

Fig. A1 Evolutions of harmony functional versus KL divergence. (a) Two streams of the evolutions; (b) maximization of intersection, inner product, and harmony measure; (c) best harmony within their permutation set Π_q; (d) least difference and orthogonal property

$$\int\xi_Q^2(x)\mathrm{d}x\leqslant C,\quad 1<C<\infty. \tag{A.1}$$

Maximizing it with respect a free $\xi_Q(x)$ leads to $\xi_Q(x)=Cp(x)$ and $H(p||q)=C||p(x)||_2$. Therefore, we can get $q(x)=\xi_Q(x)/\int\xi_Q(x)\mathrm{d}x=p(x)$, and a further maximization of $C||p(x)||_2$ pushes $p(x)$ into a most compacted form.

Applied to the BYY system by Eq. (1), it follows from Eqs. (21)–(24) that $H(p||q)$ and $\mathrm{KL}(p||q)$ are two typical cases of $H_\mu(P||Q)$ by Eq. (21), namely BYY best harmony by Eq. (2) versus BYY best matching by Eq. (24), respectively. In sequel, we further discuss their relations to existing typical approaches, summarized on the roadmap shown in Fig. A2.

We start at the BYY best matching by Eq. (24) that is rewritten as follows:

$$\mathrm{KL}(p||q)=H(p(X)\|p(X))+\int p(X)\,\mathrm{KL}(p(R|X)\|q(X|R)q(R))\,\mathrm{d}X. \tag{A.2}$$

For $p(X)=p(X|X_N,h)$ by Eq. (6) at $h=0$, we have $\mathrm{KL}(p||q)$ tends ∞ due to $H(p(X)||p(X))=-\mathrm{d}_X\ln\varepsilon$ with $\varepsilon>0\to 0$, where d_X is the dimension of the input data. However, this part $d_X\ln\varepsilon$ is irrelevant to what to be learned, and thus its effect can be removed by merely considering $\mathrm{KL}(p(R|X_N)||q(X_N|R)q(R))$, from which we are lead to two situations as follows:

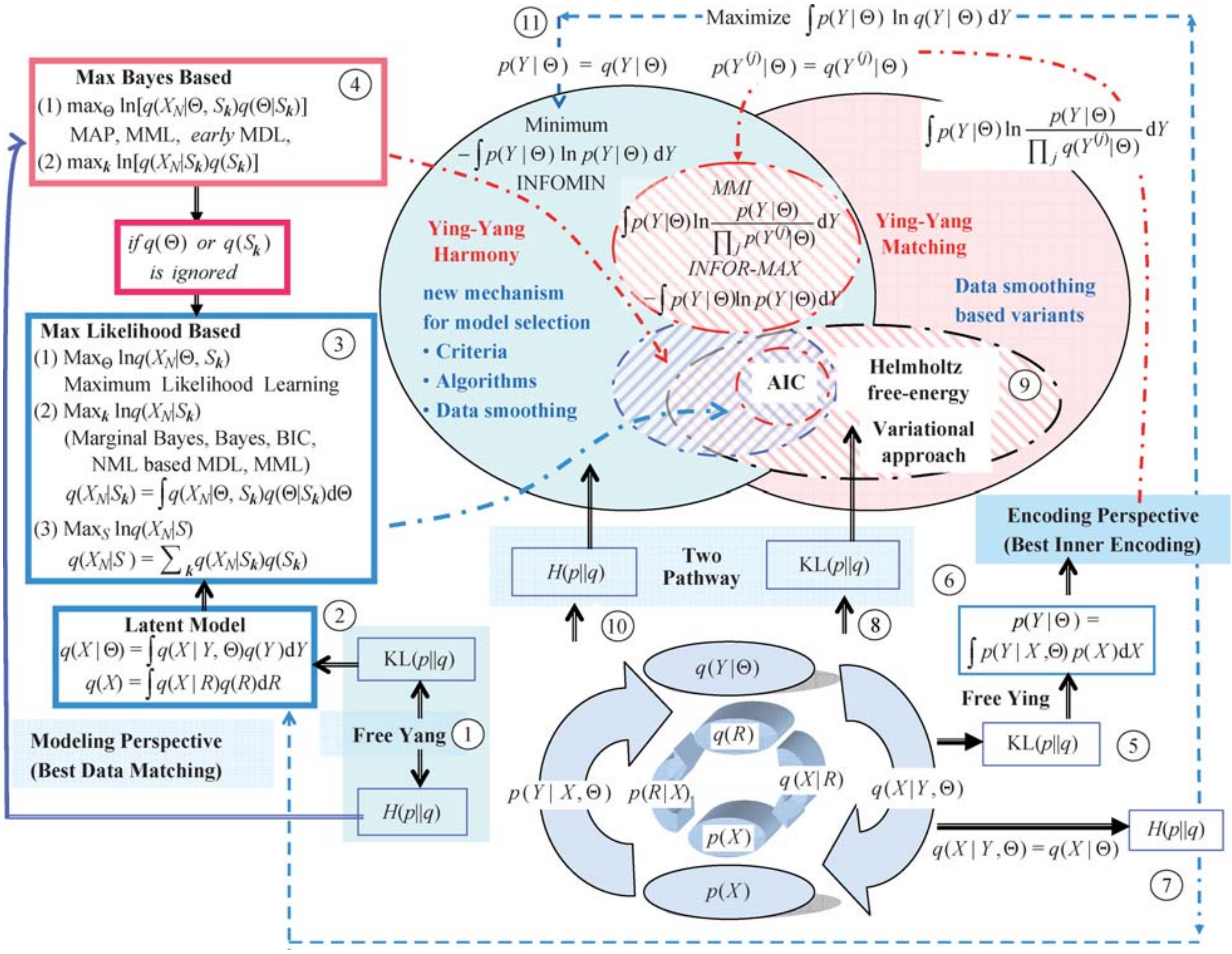

⑪ Maximize $\int p(Y|\Theta)\ln q(Y|\Theta)\mathrm{d}Y$
$p(Y|\Theta) = q(Y|\Theta)$
$p(Y^{(j)}|\Theta) = q(Y^{(j)}|\Theta)$
Max Bayes Based ④
(1) $\max_\Theta \ln[q(X_N|\Theta, S_k)q(\Theta|S_k)]$
MAP, MML, early MDL,
(2) $\max_k \ln[q(X_N|S_k)q(S_k)]$
if $q(\Theta)$ or $q(S_k)$ is ignored
Max Likelihood Based ③
(1) $\mathrm{Max}_\Theta \ln q(X_N|\Theta, S_k)$
Maximum Likelihood Learning
(2) $\mathrm{Max}_k \ln q(X_N|S_k)$
(Marginal Bayes, Bayes, BIC, NML based MDL, MML)
$q(X_N|S_k) = \int q(X_N|\Theta, S_k)q(\Theta|S_k)\mathrm{d}\Theta$
(3) $\mathrm{Max}_S \ln q(X_N|S)$
$q(X_N|S) = \sum_k q(X_N|S_k)q(S_k)$
Latent Model ②
$q(X|\Theta) = \int q(X|Y,\Theta)q(Y)\mathrm{d}Y$
$q(X) = \int q(X|R)q(R)\mathrm{d}R$
Modeling Perspective (Best Data Matching)
Minimum $-\int p(Y|\Theta)\ln p(Y|\Theta)\mathrm{d}Y$ INFOMIN
Ying-Yang Harmony
new mechanism for model selection
• Criteria
• Algorithms
• Data smoothing
MMI
$\int p(Y|\Theta)\ln\frac{p(Y|\Theta)}{\prod_j p(Y^{(j)}|\Theta)}\mathrm{d}Y$
INFOR-MAX
$-\int p(Y|\Theta)\ln p(Y|\Theta)\mathrm{d}Y$
$\int p(Y|\Theta)\ln\frac{p(Y|\Theta)}{\prod_j q(Y^{(j)}|\Theta)}\mathrm{d}Y$
Ying-Yang Matching
Data smoothing based variants
AIC
Helmholtz free-energy ⑨
Variational approach
Encoding Perspective (Best Inner Encoding)
$H(p||q)$
Two Pathway
$\mathrm{KL}(p||q)$ ⑥
⑩ ⑧
$p(Y|\Theta) = \int p(Y|X,\Theta)p(X)\mathrm{d}X$
Free Ying
KL(p||q)
Free Yang ①
H(p||q)
$q(Y|\Theta)$
$q(R)$
$p(Y|X,\Theta)$ $p(R|X)$
$q(X|R)$ $q(X|Y,\Theta)$
KL(p||q) ⑤
$p(X)$
$p(X)$
$q(X|Y,\Theta) = q(X|\Theta)$
H(p||q) ⑦

Fig. A2 BYY best harmony versus BYY best matching: A roadmap to other approaches

1) Its minimization with respect to a free structure $p(R|X)$, equivalently the case of $D^*_{\rho=0}(X)$ in Eq. (27), leads to $p(R|X_N) = q(X_N|R)q(R)/q(X_N)$ and the maximization of $q(X_N) = \int q(X_N|R)q(R)\mathrm{d}R$, from which we are lead to the canonical stream of Bayesian inference based studies, as indicated by the path of Box-① → Box-② → Box-③ on Fig. A2. Also, it relates to Akaike information criterion (AIC) and extensions [57–59].

2) Its minimization with respect to a parametric structure $p(R|X)$ provides a general formulae that becomes equivalent to the Helmholtz free energy or variational function, as indicated by the path of Box-⑧ → Box-⑨ on Fig. A2.

For $p(X) = p(X|X_N, h)$ by Eq. (6) at $h \neq 0$, we have $H(p(X)||p(X)) \neq \infty$. Along both the above paths, minimizing $\mathrm{KL}(p||q)$ by Eq. (24) leads to further extensions in help of data smoothing regularization by Eq. (32). Details are referred to Sects. 3.3 and 4.1 in Ref. [2] and the last section in Ref. [3]. Also, minimizing $\mathrm{KL}(p||q)$ with respect to a free structure $q(X|R)$ leads to

$$q(X|R) = \frac{p(R|X)\,p(X)}{p(R)} \text{ and } \min \mathrm{KL}(p(R)\,||q(R)),$$

$$q(R) = \int p(R|X)\,p(X)\mathrm{d}X, \tag{A.3}$$

as indicated by the path of Box-⑤ → Box-⑥ on Fig. A2. Putting $p(X) = p(X|X_N, h)$ by Eq. (6) at $h = 0$ into the above integral for $q(R)$, we are lead to the MMI [15], the INFOR-MAX [16], and their applications to ICA [17,18], as discussed in the beginning of Sect. 1.2.

On the other hand, it follows from the discussions around Eq. (26) at the end of Sect. 4.1 that the BYY best harmony via maximizing $H(p||q)$ by Eq. (2) with respect to a free structure $p(R|X)$ leads to

$$R^* = \arg\max_R [q(X|R)\,q(R)],$$
$$H(p||q) = \int p(X)\ln[q(X|R^*)\,q(R^*)]\mathrm{d}X, \tag{A.4}$$

which further leads to those maximum Bayesian posteriori based studies, as indicated by the path of Box-① → Box-④ on Fig. A2. It also leads to Box-④ → Box-③ to share some common parts with BYY best matching.

Moreover, in another special case $q(X|Y,\theta) = q(X|\theta)$, the BYY best harmony by Eq. (2) degenerates into two separated paths. One is the path of Box-⑦ → Box-② → Box-③ again, while the other is the path of Box-⑦ → Box-⑪ for minimizing a counterpart of INFOR-MAX, namely the minimum information transfer (INFO-MIN), which includes those studies under the name of minor component or subspace analysis (MCA and MSA) [110,111], and extensions to minor ICA (M-ICA) [17]. Readers are referred to a recent review [18].

Even importantly, maximizing $H(p||q)$ by Eq. (2) with a parametric structure $p(R|X)$, i.e., the case indicated by the Box-⑩ on Fig. A2, leads to a framework with a new mechanism for model selection, as previously discussed in Sect. 4.1.

Appendix B A modern perspective on Yin-Yang and WuXing

The Chinese ancient Yin-Yang (or preferably Ying-Yang) and WuXing philosophy came from more than 3000 years ago [53,87–89] (also see references in Fig. B1(b), including why the spelling "Ying" is preferred over "Yin"). Being very different from western sciences, evolutions of this famous ancient theory are featured by a large volume of documents and interpretations by various peoples, with a diversified coverage. Even so, there are a number of gradually converged viewpoints by serious Chinese scholars from generations to generations, though often buried among various superficial statements, applications, and variations. From a modern science perspective and based on the present author's understanding, this paper concentrates on regarding it as a Meta theory of system sciences, especially for information processing, knowledge discovering, and decision making in an intelligent system. As shown in Fig. B1, this Meta theory is believed to consist of the following three major ingredients.

1) Complementary composition of Ying-Yang system

A system that survives or interacts with its world is able to be functionally divided into two different but complement parts. One is called Yang that inputs from its external world called Yang domain and transforms what gathered via a Yang pathway into an inner domain; while the other is Ying that consists of this inner domain called Ying domain and a Ying pathway. The Ying domain accumulates, integrates, digests, and condenses whatever came from Yang, and the Ying pathway selects among the Ying domain the best ones to produce the reconstructions back to the Yang domain. Taking the Bayesian Ying-Yang system at the center of Fig. B1(a) as an example, the Yang consists of a Yang domain $P(X)$ and a Yang pathway $P(R|X)$, while the Ying consists of a Ying domain $q(R)$ and a Ying pathway $P(X|R)$.

Instead of simply regarding Ying-Yang as two opposite parts, which was frequently misunderstood by westerns, the major natures of a Ying-Yang pair are described by the following two propositions.

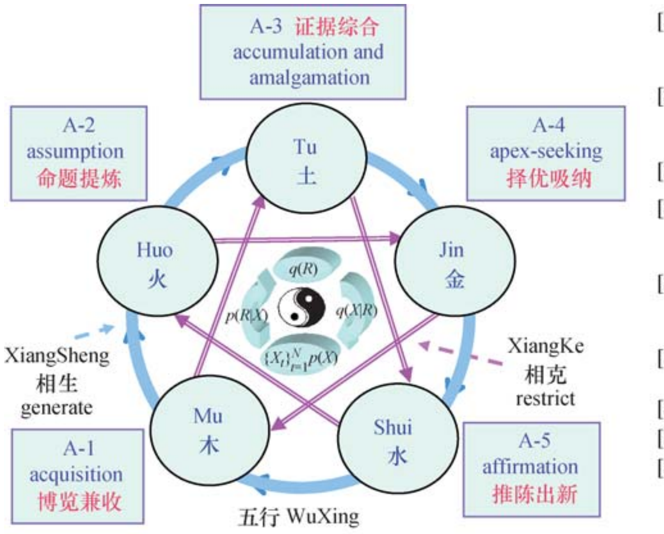
Proposition 1 Ying is primary, while Yang is secondary and comes from Ying. Yang should be vigor and capable in adapting to not only variety of external world but also the needs of Ying. In contrast, Ying should has a capacity of accommodating and accumulating, and a good ability of integrating and digesting whatever came from Yang. Yang is featured by its dynamism and vitality, while Ying is featured by its solid and compactness. Readers are referred to Refs. [S1–S3] in Fig. B1(b) for the sources of these viewpoints. Taking the Bayesian Ying-Yang system at the center of Fig. B1(a) as an example, its Yang uses $P(X)$ to get external samples while the part $P(R|X)$ is built up based on the Ying for mapping the samples into candidate assumptions that are further supplied to the Ying. On the other hand, the Ying $P(X|R)q(R)$ should have an enough capacity and a compact expression.

Proposition 2 Ying and Yang are not exclusive each other, though they were sometimes misunderstood by ones from a logical perspective. Illustrated by the well-known Yin-Yang sign at the center of Fig. B1(a), either of the Ying domain and Yang domain has a common area (called fish eye) for interacting and transferring between Ying-Yang. Also, a Ying-Yang system usually consists of multiple layers of Ying-Yang subsystems with one nested within the other. Moreover, the Ying-Yang pair keeps dynamic changes to seek a best harmony (e.g., growing and falling, expanding and shrinking) in a cyclic and balanced manner. Readers are referred to Ref. [S4] in Fig. B1(b) about its Chinese meanings.

2) WuXing circling of system operation

A Ying-Yang system survives or operates in a WuXing circling, as descried by the famous ancient Chinese WuXing theory [53]. Together with the Yin-Yang philosophy, the WuXing theory lays the foundation of TCM, by which WuXing is regarded as five states of the flow about Qi, namely, an abstract concept that is not observable but believed to be hidden causes of various observable matters, phenomena, and events. Accordingly, each of the five states is also an abstract concept named by one ancient Chinese character with its meanings being most close to the nature that the state represents, as shown in Fig. B1(a). Readers are referred to Ref. [S5] in Fig. B1(b). Be different from a spelling language, every Chinese character has a rich meaning, and each of five characters evolves itself in Chinese language for thousands of years, which acts as one source that may incur for confusions to this classic theory. Interestingly, in an intelligent system as we are studying in Fig. 6 and re-illustrated in Fig. B1(a), the five actions of the problem solving paradigm A5 [77] functionally coincides well with the classic five states as follows:

Mu → **A-1 Acquisition**: Instead of understanding Mu simply as plant, we understand it as a process that samples (e.g., plants and trees) of the world are observed,



This diagram is rotated from the classic WuXing diagram clockwise by 2π/5

(a)

[S1] 老子《道德经》“万物负阴而抱阳，冲气以为和”

（A）阴阳：系统构成互补观

[S2] 周易《系辞上》“乾以易知，坤以简能”

《周易集解》虞翻曰“阳见称易，阴藏为简，简，阅也”

[S3] 宋邵雍《观物外篇》“阳以阴为基”“阴以阳为唱”

[S4] 阴阳交错，流转，消长，转化

The sign ☯ becomes ☯ if the Yang domain is on the top.

（B）五行：系统运行节奏律

[S5] 《尚书·洪范》《黄帝内经》

阴阳二气之流动而存在的五种“存在形式”

宋张载《正蒙·乾称篇》“凡象皆气”

[S6] 生克乘诲 Sheng-Ke-Cheng-Hui：相生，相克，相乘，相诲

（C）和谐：系统寻优根本道

[S7] 《易·象传》“保合太和，乃利贞”《正蒙》“太和所谓道”

[S8] 和谐指阴阳以最默契或交换信息最小的方式达到最大共识。

[S9] ‘Ying’ should be ‘Yin’ by the current Chinese Pin Yin system that could be backtracked to over 400 years evolution from the initiatives by Westerns (e.g., M. Ricci, N. Trigault). But, the length of ‘Yin’ lost its harmony with Yang, thus ‘Ying’ is preferred by the present author since 1995.

(b)

Fig. B1 A Meta theory of intelligent system and information processing. (a) Ying-Yang, WuXing, and BYY system; (b) Chinese ancient references

which is functionally same as A-1 acquisition for getting samples as inputs to the Ying-Yang system.

Huo → **A-2 Assumption**: Instead of understanding Huo as fire, we understand it as a process that turns samples into another form or expression, which is functionally same as A-2 Assumption for making candidate assumptions based on observed samples.

Tu → **A-3 Accumulation and Amalgamation**: Instead of understanding Tu as earth, we understand it as a process of accumulation and amalgamation.

Jin → **A-4 Apex-seeking**: Instead of understanding Jin as metal, we understand it as a process of refining or selecting among possible candidates, which is functionally same as searching optimum or decision making.

Shui → **A-5 Affirmation**: Instead of understanding Shui as water, we understand it as a process that watering makes seeds rebirth new generations, which is functionally same as reconstruction from selected inner causes to fit observations.

As a result, the counterpart of the WuXing circling becomes the A5 circling in an intelligent system as shown in Fig. 6. One most practically useful part of the WuXing theory in TCM is the following Sheng-Ke-Cheng-Hui law for keeping five states well balanced, referred to Ref. [S6] in Fig. B1(b) for its Chinese concept.

Proposition 3 Sheng or XiangSheng specifies one generating or promoting the other and thus specifies the circling order, as shown in Fig. B1(a), while Ke or XiangKe tells that one state jumps to restrict the one after the next state, e.g., if Jin is too strong, we should enhance Huo to let Jin to return back a balance. Moreover, a too strong Huo will make Jin deviated from its balance towards to the other side, which is said over-restricted or Cheng; while a too weak Huo is unable to bring Jin back to its balance, which will reversely cause Huo deviated from its balance (it may also be understood from the path Jin → Mu → Huo alternatively). This is said Hui. In other words, Ke-Cheng-Hui jointly states that reversing an unbalance of one state should be made from the one before the preceding state, in a correct direction and with an appropriate strength.

Moreover, a Ying-Yang system usually consists of multiple layers of Ying-Yang subsystems with one nested within the other, and the XiangSheng also applies across layers as follows:

Proposition 4 The A5 circling in a Ying-Yang system that consists of a series of smaller circles of Ying-Yang pairs nested within its inner layers, e.g., featured by three or more levels of A5 circling as described in Fig. 6. A completion of a smaller circle assists with one moving step in its upper circle, and a jamming in an upper layer circling can be resolved by a series of lower layer circling. Therefore, a well balanced circling means a well balance within each layer and across layers.

3) Ying-Yang best harmony principle

By the WuXing theory in TCM, the circling flow is regarded as the flow about Qi that confused many westerns and attracted many Chinese scholars to find it out materially. In a BYY system, one manifestation of this flow is a mixed flow from $X_N \to R$ (or more specifically $Y, L \to \Theta \to \boldsymbol{k}) \to X_N$ featured by three nested levels of five action circling, as shown in Fig. 6. Though it is unclear how this flow relates to the flow of Qi, we may have an alternative insight that suggests to figure out what drives this flow. Referring to Refs. [S1,S7] in Fig. B1(b), this flow and all the parts in a Ying-Yang system are governed by the highest principle of seeking a Ying-Yang best harmony, called TaiHe. Though there is no definition of a western science style on what is best harmony, there were several descriptive interpretations, involving the concept of Qi and various dynamic behaviors. Here, we merely focus on one simplified understanding that includes two natures. First, it means a best dynamic agreement or match between Ying and Yang. Second, it follows from the above Proposition 1 that Ying is in a most compact form in a sense that either it uses a given capacity to accommodate as many as possible or it uses a capacity as least complexity as possible to accommodate what obtained from a given set of external observations. In turn, a compact Ying makes Yang in a compact form. Shortly, an alternative understanding is that Ying and Yang seeks a best agreement in a most tacit manner with a least amount of information transferred from data via Yang to Ying. Readers are referred to Ref. [S8] in Fig. B1(b) about its Chinese meanings.

Proposition 5 Ying-Yang seeks a best harmony in a sense of a best dynamic agreement or match by a Ying-Yang system in a most compact form with a least complexity. For the Bayesian Ying-Yang system by Eq. (1), i.e., the one at the center of Fig. B1(a), the best harmony principle is mathematically implemented by maximizing a harmony functional by Eq. (21) in general and by Eq. (2) in particular.

The last but not the least, we may also get an alternative interpretation about Qi in TCM. It follows from Ref. [S5] in Fig. B1(b) that Qi is not observable but believed to be causes of observable matters, phenomena, and events, for which we simply denote them by $f_1, f_2, \ldots, f_m$. The Ying-Yang resources are jointly distributed over the space (X, Σ) of its system, described by a Yang σ-finite measure and Ying σ-finite measure, respectively. The two resources interact on every small volume $\mathrm{d}\mu$ to seek a best harmony that maximizes $H_\mu(P||Q)$ by Eq. (21). The flow of Qi consists of the changing flows of $\mathrm{d}H_\mu/\mathrm{d}f_1, \mathrm{d}H_\mu/\mathrm{d}f_2, \ldots, \mathrm{d}H_\mu/\mathrm{d}f_m$, as well as $\mathrm{d}P/\mathrm{d}f_1, \mathrm{d}P/\mathrm{d}f_2, \ldots, \mathrm{d}P/\mathrm{d}f_m$ and $\mathrm{d}Q/\mathrm{d}f_1, \mathrm{d}Q/\mathrm{d}f_2, \ldots, \mathrm{d}Q/\mathrm{d}f_m$ that drive the Ying-Yang system towards a best harmony. In other words, Qi could be a relative concept between H_μ, P, Q and $f_1, f_2, \ldots, f_m$ on how the Ying-Yang resources dynamically change everywhere.

Appendix C   Typical notations and symbols

1) Common mathematical terms

a) $||x||^2 = xx^{\mathrm{T}}, x^{\mathrm{T}}y$ is an inner product, xy^{T} is an outer product. $x \sim y$ means $x = cy$ for a scalar c, e.g., $x^{\mathrm{new}} - x \sim y$ means $x^{\mathrm{new}} = x + cy$.

b) Given a matrix A, $|A|$ is its determinant, $\kappa[A]$ is its conditional number, and $\mathrm{Tr}[A]$ is its trace, i.e., the sum of its diagonal elements.

c) $\mathrm{vec}[A]$ is a vector from the columns of a matrix A stacked one by one. $A \geqslant 0$ means nonnegative definite, i.e., $u^{\mathrm{T}}Au \geqslant 0$ for any u.

d) either $\mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_m]$ or $\mathrm{diag}[y]$ denotes a diagonal matrix with diagonal elements being either $\lambda_1, \lambda_2, \ldots, \lambda_m$ or the elements of y.

e) For $f(x)$, $\nabla_x f(x) = \dfrac{\partial f(x)}{\partial x}$, $\nabla_{xx^{\mathrm{T}}} f(x) = \dfrac{\partial^2 f(x)}{\partial x \partial x^{\mathrm{T}}}$, and $\arg\max_x f(x)$ denotes x^* at which $f(x^*)$ is its maximum.

f) $E_{p(x)}(x) = \int xp(x)\mathrm{d}x$,
$\mathrm{Var}_{p(x)}(x) = \int \left(x - E_{p(x)}(x)\right)\left(x - E_{p(x)}(x)\right)^{\mathrm{T}} p(x)\mathrm{d}x$.
Kronecker delta $\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$
Dirac delta $\delta(x) = \begin{cases} \infty, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases}$

g) $G(x|\mu, \Sigma)$ denotes a Gaussian with a mean vector μ and a covariance matrix Σ. $B(y|v) = \prod_{\mathrm{j}} (v^{(j)})^{y^{(j)}} (1 - v^{(j)})^{1-y^{(j)}}$ is a multivariate Bernoulli.

h) $\boldsymbol{f}(u)$ is a quasi-linear vector function $\boldsymbol{f}(u) = [f(u^{(1)}), f(u^{(2)}), \ldots, f(u^{(m)})]^{\mathrm{T}}$ for a scalar function $f(r)$, e.g., $s(r) = \dfrac{1}{1 + \mathrm{e}^{-r}}$.

2) Symbols for variables and parameters

a) x is an observation vector, x_t is a sample of x at time $t, X = \{x\}$ is a set of one or a number of observation vectors.

b) $X_N = \{x_t\}_{t=1}^N$ denotes a set of N samples, or a set of N sequential observation vectors whenever there is no confusion.

c) y is a vector, either real or binary, as a part of inner representation of $x; \ell$ is a label, as a part of inner representation of x.

d) $Y = \{y\}$, $L = \{\ell\}$, $Y_N = \{y_t\}_{t=1}^N$, $L_N = \{\ell_t\}_{t=1}^N$ are counterparts of X, X_N. Simply $YL = \{y, \ell\}, YL_N = \{y_t, \ell_t\}_{t=1}^N$, even Y for YL, Y_N for YL_N.

e) $\boldsymbol{k}$ consists of one or a set of integers. $\boldsymbol{k}_Y$ is the dimension of the STM domain, and STM denotes the short term memory domain YL.

f) Θ denotes a set of all the unknown parameters, usually θ denotes a subset of Θ. Ξ denotes a set of unknown hyper-parameters.

g) $S_{\boldsymbol{k}}$ denotes a structure resulted from $\boldsymbol{k}$ element structures in a simple combination. For different $\boldsymbol{k}, S_{\boldsymbol{k}}$ shares a same configuration S but in a different scale. S denotes a type of configuration, featured by element structures and the rule for combining them.

h) R denotes a set $\{Y, L, \Theta, \boldsymbol{k}, \Xi\}$ or one of its subsets, e.g., $\{Y, L, \Theta, \boldsymbol{k}\}, \{Y, \Theta, \boldsymbol{k}\}, \{\Theta, \boldsymbol{k}, \Xi\}, \{\Theta, \boldsymbol{k}\}$, etc.

3) Custom symbols in BYY systems

YING machine $\dfrac{\mathrm{d}Q(X,R)}{\mathrm{d}\mu} = q(X,R) = q(X|R)q(R)$

a) μ is a σ-finite measure that describes a volume or capacity (e.g., a Lebesgue measure) about the measure space (X,Σ).

b) For $q(X|R)$, we have $q(x|y,\theta_{x|y})$, $q(X|Y,\theta_{x|y})$, $q(X|Y, L,\theta_{X|YL})$, etc.

c) For $q(R)$ and $\Theta = \Theta^a \cup \Theta^b$, we have $q(y|\theta_y)$, $q(Y|L,\Theta_{Y|L})$, $q(L)$, $\alpha_\ell = q(\ell)$, $\prod_L q(\Theta^a_L)$, $\prod_L q(\Theta^b_L|\Xi)$, $q(h|X_N) \propto 1/\sum_{t=1}^N p_h(x_t)$, etc.

d) $q(\Theta^a_L)$ is a noninformative priori (usually improper), $q(\Theta^b_L|\Xi)$ is a priori with hyper-parameter.

YANG machine $\dfrac{dP(X,R)}{d\mu} = p(X,R) = p(R|X)p(X)$

For $p(X) = p(X|X_N,h)$ and $p(R|X) = p(\Theta^a|\,X)$ $p(\Theta^b\big|\,X,\Xi)p(Y|X,L)p(L|X)$, we consider

a) $p(Y|X,L)$ via $\mu_L(X) = E_{p(Y|X,L)}Y$ and $\Gamma^{Y|X}_L = \mathrm{Var}_{p(Y|X,L)}[\mathrm{vec}(Y)]$. Typically, we have $\mu_L(X) = \mu_L(X,W_L)$ in a parametric structure and $\Gamma^{Y|X}_L = \Pi^{Y|X\,-1} + \rho I,\ \rho \geqslant 0,\ \Pi^{Y|X}_L = -\dfrac{\partial^2 \pi_L(X,Y,\Theta)}{\partial \mathrm{vec}[Y]\partial \mathrm{vec}[Y]^{\mathrm{T}}}$, $\pi_L(X,Y,\Theta) = \ln\left[q(X|Y,L,\Theta_{X|YL})q(Y,L\,|\Theta_{YL}\,)\right]$, including the detailed form $\pi_t(\theta_\ell)$, etc.
$Y^*_L = \arg\max\limits_Y \pi_L(X,Y,\Theta)$.

b) $q(L|X) = \dfrac{q(X|L,\Theta_L)q(L)}{\sum\limits_{L\in C_\kappa(X_N)} q(X|L,\Theta_L)q(L)}$,

$$q(X|L,\Theta_L) = \int q(X|Y,L,\Theta_{X|YL})q(Y,L|\Theta_{YL})\mathrm{d}Y$$
$$= q(X|Y^*_L,L,\Theta_{X|YL})q(Y^*_L,L|\Theta_{YL})\frac{(2\pi)^{d_Y/2}}{\left|\Pi^{Y|X}_L\right|^{1/2}}.$$

$p(L|X) = \chi_\kappa(L)q(L|X)$,
$\chi_\kappa(L) = \begin{cases} 1, & \text{for } L \in C_\kappa(X_N), \\ 0, & \text{for } L \notin C_\kappa(X_N); \end{cases}$ including the detailed forms $\chi_{\kappa,t}(i), \chi_{\kappa|i,t}(\ell), \chi_{\kappa|i\ell,t}(j)$, etc.
$C_\kappa(X_N) = \{L$: for the first κ largest ones of $H_L(\Theta) = \pi_L(X_N,Y^*_L,\Theta) + R_L(X_N,Y^*_L,\Theta)\}$, including $C^\kappa_{J,t}$, $J = i, \ell|i, j|\ell i$, etc.

$$R_L(X_N,Y^*_L,\Theta)$$
$$= -\frac{1}{2}\mathrm{Tr}[\{\Gamma^{Y|X}_L + \varepsilon_L(X_N)\varepsilon^{\mathrm{T}}_L(X_N)\}\Pi^{Y|X}_L]$$
$$+\ln[q(h)q(\Theta^a_L)] - \frac{1}{2}h^2\mathrm{Tr}[\Sigma_L(X_N)],$$

including $R(\theta_\ell,h)$.
$\Sigma_L(X) = -\dfrac{\partial^2\pi_L(X,Y,\Theta)}{\partial\mathrm{vec}[X]\partial\mathrm{vec}[X]^{\mathrm{T}}}$,
$\varepsilon_L(X) = \mathrm{vec}[\mu_L(X) - Y^*_L]$, including
$\varepsilon_t = y_t - \mu(x_t,W)$.

c) $p(\Theta^b\big|\,X,\Xi)$ is a conjugate of $q(\Theta^b_L|\Xi)$ while $p(\Theta^a|X)$ and $\mu_L(\Theta) = E_{p(\Theta^a|X)}\Theta = \Theta^{(t)}$ that is the last estimate at time t.
$\Gamma^\Theta = \mathrm{Var}_{p(R|X)}[\mathrm{vec}(\Theta)] = \Pi^{\Theta\,-1}$ and $\Pi^\Theta = -\dfrac{\partial^2\ln[q(X|R)q(R)]}{\partial\mathrm{vec}[\Theta]\partial\mathrm{vec}[\Theta]^{\mathrm{T}}}$.

4) Specific notations in BYY harmony learning

a) $\psi(\theta_{SR})$ is an indicator on a set θ_{SR} of scale representative parameters, $\psi(\theta_{SR}) \to 0$ leads to automatic model selection.

b) Apex approximation
$\displaystyle\int p(u)Q(u)\mathrm{d}u \approx Q(u^*) + \frac{1}{2}\mathrm{Tr}\left[(\Gamma^u + \varepsilon_u\varepsilon^{\mathrm{T}}_u)\frac{\partial^2 Q(u^*)}{\partial u\partial u^{\mathrm{T}}}\right]$,
$u^* = \arg\max_u Q(u), \varepsilon_u = u^\mu - u^*$.

c) $H_\mu(P||Q)$ is the general harmony functional that covers two typical cases:

- $\mathrm{d}\mu = \mathrm{d}x$, we get harmony functional $H(p||q)$, including its detailed forms $H(p||q,\boldsymbol{k},\Xi)$ and $H(p||q,\Theta,\boldsymbol{k},\Xi)$;
- $\mathrm{d}\mu = \mathrm{d}P$, we get f-divergence $-H_P(P||Q)$, including Kullback-Leibler divergence $\mathrm{KL}(p||q)$ and its detailed forms;
- $P = Q$, we get f-entropy $-H_\mu(P||P)$, including entropy $-H(p||p)$.

d) Gradient flow

$$\nabla_{\Theta_L} H(p||q,\Theta,\boldsymbol{k})$$
$$= [p(L|X_N) + \Delta\pi_L(X_N,Y^*_L)]\nabla_{\Theta_L}\pi_L(X_N,Y^*_L,\Theta)$$
$$+\ p(L|X_N)\nabla_{\Theta_L}R_L(X_N,Y^*_L,\Theta)$$
$$-\ \frac{1}{2}\Delta\pi_L(X_N,Y^*_L)\nabla_{\Theta_L}\ln|\Pi^{Y|X}_L|,$$
$$\Delta\pi_L(X,Y)$$
$$= q(L|X_N)$$
$$\times\left[\chi_\kappa(L)H_L(\Theta) - \sum_L q(L|X_N)\chi_\kappa(L)H_L(\Theta)\right]$$

and $p(L|X) = \chi_\kappa(L)q(L|X)$,

including the detailed forms:

$$\Delta_{J,t}$$
$$= q(J|x_t)$$
$$\times\left[\chi_{\kappa,t}(J)H_t(\theta_J) - \sum_J q(J|x_t)\chi_{\kappa,t}(J)H_t(\theta_J)\right],$$

$p(J|x_t) = \chi_{\kappa,t}(J)q(J|x_t),\ p_{J,t} = p(J|x_t) + \Delta_{J,t}$,
$J = i, \ell|i, j|\ell i$.

e) For learning HMM in Fig. 14, we have $p^\delta(\,\ell_t,\ell_{t-1}|\,\theta) = p(\,\ell_t,\ell_{t-1}|\,\theta) + \Delta_{\ell_t,\ell_{t-1}}$ in Box-③ and Δq^δ_{ij} in Box-④ that is obtained from

$$\sum_{\ell_t,\ell_{t-1}} \Delta_{\ell_t,\ell_{t-1}}\nabla_\phi \ln q_{\ell_{t-1}}$$
$$= \tau^{\mathrm{T}}\nabla_\phi q$$
$$= \mathrm{Tr}(\{(I - [q_{i|j}]^{\mathrm{T}})^{-1}\tau q^{\mathrm{T}}\}^{\mathrm{T}}\nabla_\phi[q_{i|j}])$$
$$= \sum_{ij}\Delta q^\delta_{ij}\nabla_\phi \ln q_{i|j},$$
$$\Delta q^\delta_{ij} = q_j q_{i|j}\sum_\ell (I - [q_{i|j}]^{\mathrm{T}})^{-1}_{i\ell}\tau_\ell,$$
$$\tau_j = \sum_{\ell_t}\frac{\Delta_{\ell_t,\ell_{t-1}=j}}{q_j},$$

which comes from $\nabla_\phi q = (I - [q_{i|j}])^{-1}\nabla_\phi[q_{i|j}]q$, where $\tau = \mathrm{vec}\{\tau_j\}$, $q = \mathrm{vec}\{q_j\}$, $[U]_{ij}$ is the ijth element of U.

Acknowledgements The work described in this paper was supported by a grant of the General Research Fund (GRF) from the Research Grant Council of the Hong Kong SAR (Project No: CUHK4180/10E). The work also came from four years of summer lectures taught by the present author in Peking University, supported by Chang Jiang Scholars Program by Ministry of Education of China for Chang Jiang Chair Professorship.

References

1. Duda R O, Hart P E, Stork D G. Pattern Classification. 2nd ed. New York: John Wiley & Sons, 2001
2. Xu L. Machine learning problems from optimization perspective. Journal of Global Optimization, 2010, 47: 369–401
3. Xu L. Bayesian Ying Yang learning. Scholarpedia, 2007, 2(3): 1809 http://scholarpedia.org/article/Bayesian_Ying_Yang_learning
4. Aster R, Borchers B, Thurber C. Parameter Estimation and Inverse Problems. New York: Elsevier Academic Press, 2004
5. Brown R G, Hwang P Y C. Introduction to Random Signals and Applied Kalman Filtering. 3rd ed. New York: John Wiley & Sons, 1997
6. Narendra K S, Parthasarathy K. Identification and control of dynamical systems using neural networks. IEEE Transactions on Neural Networks, 1990, 1(1): 4–27
7. Redner R A, Walker H F. Mixture densities, maximum likelihood, and the EM algorithm. SIAM Review, 1984, 26(2): 195–239
8. Xu L, Jordan M I. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation, 1996, 8(1): 129–151
9. Anderson T W, Rubin H. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. 1956, 5: 111–150
10. Rubi D, Thayer D. EM algorithm for ML factor analysis. Psychometrika, 1976, 57: 69–76
11. Bozdogan H, Ramirez D E. FACAIC: Model selection algorithm for the orthogonal factor model using AIC and FACAIC. Psychometrika, 1988, 53(3): 407–415
12. Burnham K P, Anderson D R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer, 2002
13. Tikhonov A N, Arsenin V Y. Solutions of Ill-Posed Problems. Washington: Winston and Sons, 1977
14. Poggio T, Girosi F. Networks for approximation and learning. Proceedings of the IEEE, 1990, 78(9): 1481–1497
15. Amari S I, Cichocki A, Yang H. A new learning algorithm for blind separation of sources. In: Touretzky D S, Mozer M C, Hasselmo M E, eds. Advances in Neural Information Processing System 8. Cambridge: MIT Press, 1996, 757–763
16. Bell A J, Sejnowski T J. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 1995, 7(6): 1129–1159
17. Xu L. Independent component analysis and extensions with noise and time: A Bayesian Ying-Yang learning perspective. Neural Information Processing — Letters and Reviews, 2003, 1(1): 1–52
18. Xu L. Independent subspaces. In: Ramón J, Dopico R, Dorado J, Pazos A, eds. Encyclopedia of Artificial Intelligence, Hershey(PA): IGI Global. 2008, 903–912
19. Xu L. Least mean square error reconstruction principle for self-organizing neural-nets. Neural Networks, 1993, 6(5): 627–648
20. McLachlan G J, Krishnan T. The EM Algorithms and Extensions. New York: John Wiley & Sons, 1997
21. 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
22. Amari S. Information geometry of the EM and EM algorithms for neural networks. Neural Networks, 1995, 8(9): 1379–1408
23. Grenander U, Miller M. Pattern theory: From representation to inference. Oxford: Oxford University Press, 2007
24. Mumford D. On the computational architecture of the neocortex II: The role of cortico-cortical loops. Biological Cybernetics, 1992, 66(3): 241–251
25. Friston K. A theory of cortical responses. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2005, 360(1456): 815–836
26. Yuille A L, Kersten D. Vision as Bayesian inference: Analysis by synthesis? Trends in Cognitive Sciences, 2006, 10(7): 301–308
27. Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2): 461–464
28. Rissanen J. Modeling by shortest data description. Automatica, 1978, 14: 465–471
29. Rissanen J. Information and Complexity in Statistical Modeling. New York: Springer, 2007
30. DeGroot M H. Optimal Statistical Decisions. Hooken: Wiley Classics Library, 2004
31. Mackay D J C. A practical Bayesian framework for backpropagation networks. Neural Computation, 1992, 4(3): 448–472
32. MacKay D. Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press, 2003
33. Wallace C S, Boulton D M. An information measure for classification. Computer Journal, 1968, 11(2): 185–194
34. Wallace C S, Dowe D R. Minimum message length and Kolmogorov complexity. Computer Journal, 1999, 42(4): 270–280
35. Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 1988, 59: 291–294
36. Palmieri F, Zhu J, Chang C. Anti-Hebbian learning in topologically constrained linear networks: A tutorial. IEEE Transactions on Neural Networks, 1993, 4(5): 748–761
37. Grossberg S, Carpenter G A. Adaptive resonance theory. In: Arbib M A, ed. The Handbook of Brain Theory and

Neural Networks. 2nd ed. Cambridge: MIT Press, 2002, 87–90

38. Carpenter G A, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing, 1987, 37: 54–115
39. Kawato M. Cerebellum and motor control. In: Arbib M A, ed. The Handbook of Brain Theory and Neural Networks. 2nd ed. Cambridge: MIT Press, 2002, 190–195
40. Shidara M, Kawano K, Gomi H, Kawato M. Inverse-dynamics model eye movement control by Purkinje cells in the cerebellum. Nature, 1993, 365(6441): 50–52
41. Wolpert D, Kawato M. Multiple paired forward and inverse models for motor control. Neural Networks, 1998, 11(7–8): 1317–1329
42. 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
43. Dayan P, Hinton G E, Neal R M, Zemel R S. The Helmholtz machine. Neural Computation, 1995, 7(5): 889–904
44. Jaakkola T S. Tutorial on variational approximation methods. In: Opper M, Saad D, eds. Advanced Mean Field Methods: Theory and Practice. Cambridge: MIT press, 2001, 129–160
45. Jordan M, Ghahramani Z, Jaakkola T, Saul L. Introduction to variational methods for graphical models. Machine Learning, 1999, 37(2): 183–233
46. Corduneanu A, Bishop C M. Variational Bayesian model selection for mixture distributions. In: Jaakkola T, Richardson T, eds. Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics. 2001, 27–34
47. Xu L. Bayesian-Kullback coupled YING-YANG machines: Unified learning and new results on vector quantization. In: Proceedings of the International Conference on Neural Information Processing. 1995, 977–988 (A further version in NIPS8. In: Touretzky D S, et al. eds. Cambridge: MIT Press, 444–450)
48. Xu L. Ying-Yang learning. In: Arbib M A, ed. The Handbook of Brain Theory and Neural Networks. 2nd ed. Cambridge: MIT Press, 2002, 1231–1237
49. Xu L. Advances on BYY harmony learning: Information theoretic perspective, generalized projection geometry, and independent factor auto-determination. IEEE Transactions on Neural Networks, 2004, 15(4): 885–902
50. Xu L. Learning algorithms for RBF functions and subspace based functions. In: Olivas E, et al. eds. Handbook of Research on Machine Learning, Applications and Trends: Algorithms, Methods and Techniques. Hershey(PA): IGI Global, 2009, 60–94
51. Xu L. Bayesian Ying Yang system, best harmony learning, and Gaussian manifold based family. In: Zurada et al. eds. Computational Intelligence: Research Frontiers, WCCI2008 Plenary/Invited Lectures. Lecture Notes in Computer Science, 2008, 5050: 48–78
52. Xu L, Oja E. Randomized Hough transform. In: Ramón J, Dopico R, Dorado J, Pazos A, eds. Encyclopedia of Artificial Intelligence. Hershey(PA): IGI Global, 2008, 1354–1361
53. Veith I. The Yellow Emperor's Classic of Internal Medicine. Berkeley: University of California Press, 1972
54. Vapnik, V. Estimation of Dependences Based on Empirical Data. Springer, 2006
55. Stone M. Cross-validation: A review. Mathematics, Operations and Statistics, 1978, 9(1): 127–140
56. Rivals I, Personnaz L. On cross validation for model selection. Neural Computation, 1999, 11(4): 863–870
57. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6): 714–723
58. Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extension. Psychometrika, 1987, 52(3): 345–370
59. Cavanaugh J E. Unifying the derivations for the Akaike and corrected Akaike information criteria. Statistics & Probability Letters, 1997, 33(2): 201–208
60. Williams P M. Bayesian regularization and pruning using a Laplace prior. Neural Computation, 1995, 7(1): 117–143
61. Tibshirani R, Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267–288
62. MacKay D J C. Bayesian interpolation. Neural Computation, 1992, 4(3): 415–447
63. Salah A A, Alpaydin E. Incremental mixtures of factor analyzers. In: Proceedings the 17th International Conference on Pattern Recognition. 2004, 1: 276–279
64. Xu L, Krzyzak A, Oja E. Rival penalized competitive learning for clustering analysis, RBF net and curve detection. IEEE Transactions on Neural Networks, 1993, 4(4): 636–649
65. Xu L, Krzyzak A, Oja E. Unsupervised and supervised classifications by rival penalized competitive learning. In: Proceedings of the 11th International Conference on Pattern Recognition. 1992, I: 672–675
66. Xu L. Rival penalized competitive learning. Scholarpedia, 2007, 2(8): 1810 http://www.scholarpedia.org/article/Rival_penalized_competitive_learning
67. Corduneanu A, Bishop C M. Variational Bayesian model selection for mixture distributions. In: Richardson T, Jaakkola T, eds. Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics. 2001, 27–34
68. 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
69. Tu S, Xu L. A study of several model selection criteria for determining the number of signals. In: Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. 2010, 1966–1969
70. Xu L. Fundamentals, challenges, and advances of statistical learning for knowledge discovery and problem solving: A BYY harmony perspective, keynote talk. In: Proceedings

of the International Conference on Neural Networks and Brain. 2005, 1: 24–55

71. Hinton G E, Zemel R S. Autoencoders, minimum description length and Helmholtz free energy. In: Cowan J D, Tesauro G, Alspector J, eds. Advances in Neural Information Processing Systems 6. San Mateo: Morgan Kaufmann, 1994, 449–455
72. Xu L. Data smoothing regularization, multi-sets-learning, and problem solving strategies. Neural Networks, 2003, 16(5–6): 817–825
73. Xu L. Bayesian Ying Yang system and theory as a unified statistical learning approach: (I) Unsupervised and semi-unsupervised learning. In: Amari S, Kassabov N, eds. Brain-like Computing and Intelligent Information Systems. Springer-Verlag, 1997, 241–274
74. Xu L. Bayesian Ying Yang system and theory as a unified statistical learning approach: (II) From unsupervised learning to supervised learning and temporal modeling and (III) Models and algorithms for dependence reduction, data dimension reduction, ICA and supervised learning. In: Wong K M, King I, Yeung D Y, eds. Proceedings of Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective. 1997: 25–60
75. Xu L. Bayesian Ying Yang system and theory as a unified statistical learning approach (VII): Data smoothing. In: Proceedings of the International Conference on Neural Information Processing. 1998, 1: 243–248
76. Bishop C M. Training with noise is equivalent to Tikhonov regularization. Neural Computation, 1995, 7(1): 108–116
77. Xu L. A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition, 2007, 40(8): 2129–2153
78. Xu L, Oja E, Kultanen P. A new curve detection method randomized Hough transform (RHT). Pattern Recognition Letters, 1990, 11(5): 331–338
79. Hough P V C. Method and means for recognizing complex patterns. US Patent, 3069654, 1962-12-18
80. Xu L. Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, ME-RBF models and three-layer nets. International Journal of Neural Systems, 2001, 11(1): 3–69
81. Xu L. Bayesian Ying-Yang learning theory for data dimension reduction and determination. Journal of Computational Intelligence in Finance, 1998, 6(5): 6–18
82. Tu S, Xu L. Theoretical analysis and comparison of several criteria on linear model dimension reduction. In: Adali T, Jutten C, Romano J M T, Barros A K, eds. Independent Component Analysis and Signal Separation. Lecture Notes in Computer Science, 2009, 5441: 154–162
83. Xu L. BYY harmony learning, independent state space and generalized APT financial analyses. IEEE Transactions on Neural Networks, 2001, 12(4): 822–849
84. Xu L. Temporal BYY encoding, Markovian state spaces, and space dimension determination. IEEE Transactions on Neural Networks, 2004, 15(5): 1276–1295
85. Kalman R E. A new approach to linear filtering and prediction problems. Transactions of the ASME Journal of Basic Engineering, 1960, 35–45
86. Sun K, Tu S, Gao D Y, Xu L. Canonical dual approach to binary factor analysis. In: Adali T, Jutten C, Romano J M T, Barros A K, eds. Independent Component Analysis and Signal Separation. Lecture Notes in Computer Science, 2009, 5441: 346–353
87. Nathan S. Science and medicine in imperial China — The state of the field. The Journal of Asian Studies, 1988, 47(1): 41–90
88. Wilhelm R, Baynes C. The I Ching or Book of Changes, with Foreword by Carl Jung. 3rd ed. Bollingen Series XIX. Princeton: Princeton University Press, 1967
89. Hansen C. A Daoist Theory of Chinese Thought: A Philosophical Interpretation. New York: Oxford University Press, 2000
90. Shilov G E, Gurevich B L. Integral, Measure, and Derivative: A Unified Approach. Silverman R trans. New York: Dover Publications, 1978
91. Ali S M, Silvey S D. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society: Series B, 1966, 28(1): 131–140
92. Kullback S, Leibler R A. On information and sufficiency. Annals of Mathematical Statistics, 1951, 22(1): 79–86
93. Shore J. Minimum cross-entropy spectral analysis. IEEE Transactions on Acoustics, Speech and Signal Process, 1981, 29(2): 230–237
94. Burg J P, Luenberger D G, Wenger D L. Estimation of structured covariance matrices. Proceedings of the IEEE, 1982, 70(9): 963–974
95. Jaynes E T. Information theory and statistical mechanics. Physical Review, 1957, 106(4): 620–630
96. 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
97. Jeffreys H. An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences, 1946, 186(1007): 453–461
98. Xu L. BYY learning system and theory for parameter estimation, data smoothing based regularization and model selection. Neural, Parallel and Scientific Computations, 2000, 8(1): 55–82
99. Xu L. BYY $\sum$-$\prod$ factor systems and harmony learning. Invited talk. In: Proceedings of International Conference on Neural Information Processing (ICONIP'2000). 2000, 1: 548–558
100. Xu L. Bayesian Ying Yang learning. In: Zhong N, Liu J, eds. Intelligent Technologies for Information Analysis. Berlin: Springer, 2004, 615–706
101. Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory, 1998, 44(6): 2743–2760
102. Xu L, Amari S. Combining classifiers and learning mixture-of-experts. In: Ramón J, Dopico R, Dorado J, Pazos A,

eds. Encyclopedia of Artificial Intelligence. Hershey(PA): IGI Global, 2008, 318–326

103. Xu L. BYY learning, regularized implementation, and model selection on modular networks with one hidden layer of binary units. Neurocomputing, 2003, 51: 277–301 (Errata on Neurocomputing, 2003, 55(1–2): 405–406)
104. Gales M J F, Young S. The application of hidden Markov models in speech recognition. Foundations and Trends in Signal Processing, 2008, 1(3): 195–304
105. Su D, Wu X H, Xu L. GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection. In: Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. 2010, 4890–4893
106. Rosti A V, Gales M. Factor analysed hidden Markov models for speech recognition. Computer Speech and Language, 2004, 18(2): 181–200
107. Gales M J F. Discriminative models for speech recognition. In: Proceedings of Information Theory and Applications Workshop. 2007, 170–176
108. Woodland P C, Povey D. Large scale discriminative training of hidden Markov models for speech recognition. Computer Speech and Language, 2002, 16(1): 25–47
109. Csiszár I, Tusnády G. Information geometry and alternating minimization procedures. Statistics and Decisions, 1984, (Suppl. 1): 205–237
110. Xu L, Oja E, Suen C Y. Modified Hebbian learning for curve and surface fitting. Neural Networks, 1992, 5(3): 441–457
111. Xu L, Krzyzak A, Oja E. A neural net for dual subspace pattern recognition methods. International Journal of Neural Systems, 1991, 2(3): 169–184

Lei Xu is a chair professor of The Chinese University of Hong Kong (CUHK), a Chang Jiang Chair Professor of Peking University, a guest Research Fellow of Institute of Biophysics, Chinese Academy of Sciences, an honorary Professor of Xidian University. He graduated from Harbin Institute of Technology by the end of 1981, and completed his master and Ph.D thesis at Tsinghua University during 1982–1986. Then, he joined Department Mathematics, Peking University in 1987 first as a postdoc and then exceptionally promoted to associate professor in 1988 and to a full professor in 1992. During 1989–1993, he worked at several universities in Finland, Canada and USA, including Harvard and MIT. He joined CUHK in 1993 as senior lecturer, became professor in 1996 and took the current position since 2002. Prof. Xu has published dozens of journal papers and also many papers in conference proceedings and edited books, covering the areas of statistical learning, neural networks, and pattern recognition, with a number of well-cited papers, e.g., his papers got over 3200 citations according to SCI-Expended (SCI-E) and over 5500 citations according to Google Scholar (GS), and over 2000 (SCI-E) and 3600 (GS) for his 10 most frequently cited papers. He served as associate editor for several journals, including Neural Networks (1995–present) and IEEE Transactions on Neural Networks (1994–1998), and as general chair or program committee chair of a number of international conferences. Moreover, Prof. Xu has served on governing board of International Neural Networks Society (INNS) (2001–2003), INNS Award Committee (2002–2003), and Fellow Committee of IEEE Computational Intelligence Society (2006, 2008), chair of Computational Finance Technical Committee of IEEE Computational Intelligence Society (2001–2003), and a past president of Asian-Pacific Neural Networks Assembly (APNNA). He has also served as an engineering panel member of Hong Kong RGC Research Committee (2001–2006), a selection committee member of Chinese NSFC/HK RGC Joint Research Scheme (2002–2005), external expert for Chinese NSFC Information Science (IS) Panel (2004–2006, 2008), external expert for Chinese NSFC IS Panel for distinguished young scholars (2009–2010), and an nominator for the prestigious Kyoto Prize (2003, 2007). Prof. Xu has received several Chinese national academic awards (including 1993 National Nature Science Award) and international awards (including 1995 INNS Leadership Award and the 2006 APNNA Outstanding Achievement Award). He has been elected to an IEEE Fellow (2001–) and a Fellow of International Association for Pattern Recognition (2002–), and a member of European Academy of Sciences (2002–).