Improving the Minimum Free Energy Principle to the Maximum Information Efficiency Principle

Chenguang Lu *

Posted Date: 30 April 2025

doi: 10.20944/preprints202504.2525.v1

Keywords: variational Bayes; free energy principle; entropy; Shannon mutual information; semantic mutual information; information rate-distortion; EM algorithm; active inference; free energy; Boltzmann distribution

 doi:10.20944/preprints202504.2525.v1

Article

Improving the Minimum Free Energy Principle to the Maximum Information Efficiency Principle

Chenguang Lu [1,2]

[1] Intelligence Engineering and Mathematics Institute, Liaoning Technical University, Fuxin 123000, China; survival99@gmail.com

[2] School of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022, China

Abstract: Friston proposed the Minimum Free Energy (MFE) principle based on the Variational Bayesian (VB) method. This principle inherits the basic idea of Evolutionary System Theory. Still, it emphasizes the coordination between the brain, behavior, and the environment, promoting self-organization. However, there are two issues with VB and the MFE principle: 1) When using VB to optimize latent variables, the Variational Free Energy (VFE) F may not always decrease; 2) The concept of VFE is inconsistent with physical free energy. The author proposed the semantic information G theory and $R(G)$ function 30 years ago (R is the minimum mutual information for the given semantic mutual information G). Based on the study of the $R(G)$ function, this paper proposes the Semantic Variational Bayesian (SVB) and the Maximum Information Efficiency (MIE) principle to resolve the two issues with VB and the MFE principle. Theoretic analysis and computing experiments prove that $R–G = F\text{-}H(X|Y)$ instead of F continues to decrease when optimizing latent variables. The experiments show that SVB is reliable and straightforward for latent variables and active inference. This paper also explains the relationship between information, entropy, and free energy in local non-equilibrium and equilibrium systems, concluding that Shannon's information is similar to free energy's increment, semantic information is similar to exergy's increment, and VFE is equivalent to some physical entropy. The MIE principle inherits the basic idea of the MFE principle but is easier to understand and apply.

Keywords: variational Bayes; free energy principle; Shannon mutual information; semantic mutual information; information rate-distortion; EM algorithm; active inference; entropy; free energy; Boltzmann distribution

1. Introduction

In 1993, Hinton et al. [1,2] used the minimum free energy as an optimization criterion to improve neural network learning, which led to significant breakthroughs. This method was later developed into the Variational Bayesian (VB) approach [3,4], which has since been widely and successfully applied in machine learning (including reinforcement learning) [5]. Friston [6,7] extended the application of VB to neuroscience, evolving the Minimum Free Energy (MFE) criterion into the MFE principle, aiming for it to become a unified scientific theory of brain function and biological behavior. This theory integrates concepts such as predictive coding, perceptual prediction, active inference, information, and stochastic dynamical systems [7–9]. Unlike the passive perspective in the book *Entropy: A New World View* [10], Friston's theory inherits the optimistic entropy-based worldview from the existing Evolutionary Systems Theory (EST) [11,12] and emphasizes the ability of bio-organisms to predict, adapt to, and influence their environment, which promotes self-organization and order.

Friston's theory has garnered widespread attention and deep reflection [13]. Many applications have emerged [14]. Information and free energy have been combined to explain self-organization and order [15,16]. However, some criticisms have also been raised. Some argue that free energy means negative entropy and is essential for life; thus, minimizing free energy would imply death [17]. Others

 doi:10.20944/preprints202504.2525.v1

contend that while the MFE principle is valuable as a tool, its validity as a universal principle or falsifiable scientific law remains debatable.

As early as 1990, Silverstein and Pimbley [18] used "minimum free energy method" in the title of their article. Their objective function is defined as a linear combination of the mean-square error energy expression and the signal entropy expression. In the paper entitled *Two Kinds of Free Energy and the Bayesian Revolution*, Gottwald and Braun [19] reviewed many similar studies and argued that Friston's free energy is just one of the two kinds. The other kind uses various objective functions in Maximum Entropy (ME) methods. These objective functions either equal the reward function plus the entropy function (which needs to be maximized) or the average loss minus the entropy function (which needs to be minimized). Both kinds of free energy methods are very significant. But in my opinion, Friston's MFE principle is different from the ME principle [20,21] for the following reasons:

- Subjective prediction and objective reality (or objective reality and subjective goals) approach each other bidirectionally and dynamically.
- Loss functions are expressed in terms of information or entropy measures, transforming the active inference issue into the inverse issue of sample learning.
- Multi-task coordination and tradeoffs exist, requiring solving latent variables.

It seems that the MFE principle can better explain the subjective initiative of bio-organisms than the ME principle. However, from the author's perspective, the MFE principle is still imperfect because it has two issues. Below, we refer to free energy in VB and the MFE principle as Variational Free Energy (VFE).

Thirty years ago, the author extended Shannon's information theory to a semantic information theory [22–24]. The formula for semantic Mutual Information (MI) is: $I(X; Y_\theta)=H(X)–H(X|Y_\theta)$, where $H(X|Y_\theta)$ is the semantic posterior entropy or posterior cross-entropy. Roughly speaking, $H(X|Y_\theta)$ equals VFE; minimizing VFE is equivalent to maximizing semantic MI. Later, the author called the generalized theory the Semantic Information G Theory [25,26] (or simply the G Theory, where "G" stands for "generalization") for machine learning. The author recognized early that $H(X|Y_\theta)$ does not necessarily decrease monotonically when a Shannon channel matches a semantic channel. The author also studied mixture models with the Expectation-Maximization (EM) algorithm [25]. Experimental observations indicated that $H(X|Y_\theta)$ and VFE do not continually decrease as the mixture model converges. Although VB ensures the mixture model's convergence [2,3], its theoretical justification is flawed. This is the first issue with VB and the MFE principle.

The second issue concerns the relationship between VFE and physical entropy and free energy. There is an apparent contradiction. In physics, free energy is the energy available to perform work and is usually maximized [27]. If we actively minimize free energy, aren't we simply following the trend of increasing entropy [17]? Thus, we must clarify what VFE or $H(X|Y_\theta)$ truly represents in thermodynamic systems and how information, entropy, and free energy mutually relate in such systems.

As early as 1993, the author [23] analyzed information between temperature and a molecule's energy in a local no-equilibrium and equilibrium system. The conclusions include that Shannon MI is similar to the increment of free energy in a local no-equilibrium system; semantic MI is similar to the increment of free energy in a local equilibrium system. In addition, the author extended Shannon's rate-distortion function $R(D)$ to obtain the information rate-fidelity function $R(G)$ [23,25], where R represents the minimum Shannon MI for given semantic MI G (representing fidelity). The method for solving $R(G)$ can be called the Semantic Variational Bayesian (SVB) method [28]. SVB can solve issues VB addresses but does not always minimize VFE. Instead, it minimizes $R−G=F−H(X|Y)$. Minimizing $R−G$ is equivalent to maximizing information efficiency: G/R. Thus, the author proposes the Maximum Information Efficiency (MIE) principle. This principle can overcome the two issues with VB and the MFE principle. The applications of the G Theory in machine learning include multi-label learning, maximum MI classification for unseen instances, mixture models [25], Bayesian confirmation [26], and semantic compression [29]. The success of these applications strengthens the validity of the G Theory.

 doi:10.20944/preprints202504.2525.v1

The motivation of this paper is to clarify the above two issues existing in VB and the MFE principle and to improve them. This paper aims to provide an improved version of the MFE principle.

The contributions of this paper:

- Mathematically clarifying the theoretical and practical inconsistencies in VB and the MFE principle.
- Explaining the relationships between Shannon MI, semantic MI, VFE, and physical entropy with free energy from the perspectives of the G Theory and statistical physics.
- Providing experimental evidence by mixture model examples (including one used by Neal and Hinton [2]) to demonstrate that VFE may increase during the convergence of mixture models and explain why this occurs.

The iteration method for minimum information difference in SVB is inspired by the iterative approach used by Shannon et al. in solving the rate-distortion function [30–32].

Abbreviations and explanations can be found in Appendix A. Information about Python source codes for producing most figures in this paper can be found in Appendix B.

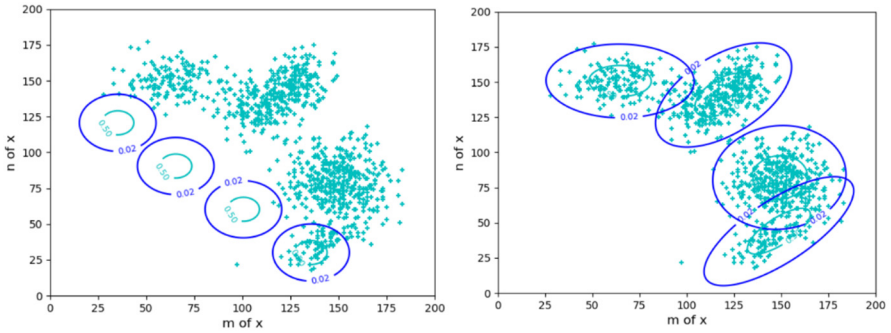
2. Two Typical Tasks of Machine Learning

2.1. Sheep Clustering: Mixture Models

To explain the tasks to be completed by VB. Let's take sheep clustering and sheep herding as examples. This section describes the sheep clustering issue (see Figure 1).

Suppose several sheep flocks are distributed on a grassland with fuzzy boundaries. We can observe that the density distribution (the proportion of the number of sheep per unit area) is $P(x)$. We also know there are n flocks of sheep, and the distributions have a certain regularity, such as the Gaussian distribution. We can establish a mixture model: $P_\theta(x) = P(y_1)P(x|\theta_1) + P(y_2)P(x|\theta_2) + \ldots$ Then, we use the maximum likelihood criterion or the minimum cross entropy criterion to optimize the mixture ratios $P(y_1)$, $P(y_2)$, … and model parameters.

The Expectation-Maximization (EM) algorithm [33,34] is usually used to solve the mixture model. This algorithm is very clever. It can automatically adjust the different components, namely the likelihood function (circled in Figure 1), so that each component covers a group of instances (a flock) and can provide the appropriate ratios $P(y_j)$, j=1,2,3,4, of the mixture model's components, also known as the probability distribution of the latent variable. We sometimes call $P(y)$ the latent variable.



(**a**) (**b**)

Figure 1. Explaining the Gaussian mixture model using sheep clustering as an example. (a) The iteration starts; (b) The iteration converges. The x=(m,n) is two-dimensional.

The EM algorithm is not ideal in two aspects: 1) There have been problems with its convergence proof [25], which has led to blind improvements; 2) $P(y)$ sometimes converges very slowly, and it is challenging to solve $P(y)$ when the likelihood functions remains unchanged. Researchers use VB not only to improve the EM algorithm [2,3] but also to solve latent variables [5].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 April 2025 doi:10.20944/preprints202504.2525.v1

The mixture model belongs to unsupervised learning in machine learning and is very representative. Similar methods can be found in Restricted Boltzmann Machines and deep learning pre-training tasks.

2.2. Driving the Sheep to Pastures: Constrained Control and Active Inference

Driving sheep to pastures is the constraint control issue of random events and also the active inference issue involving reinforcement learning. In this case, Shannon's MI reflects the control complexity. We need to maximize the purposefulness (or utility) and minimize the Shannon MI, the control cost.

The circles in the figure represent the control targets; the points in Figure 2a reflect the initial flock density distribution $P(x)$. There are usually two types of control objectives or constraints:

- The objectives are expressed by the probability distributions $P(x|\theta_j)(j=1,2,3,4)$. Given $P(x)$ and $P(x|\theta_j)$, we solve the Shannon channel $P(y|x)$ and the herd ratio $P(y)$. It is required that $P(x|y_j)$ is close to $P(x|\theta_j)$ and $P(y)$ can minimize the control cost.
- The objectives are expressed by the fuzzy ranges. $P(x|\theta_j)$ ($j=1,2,\ldots$) can be obtained from $P(x)$ and the fuzzy ranges, and the others are the same.

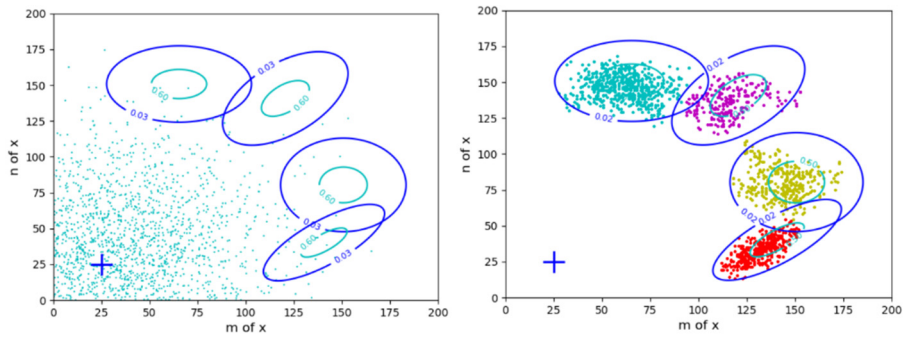


Figure 2. Taking herding sheep as an example to illustrate the constraint control of uncertain events (the constraint condition is some fuzzy ranges). (a) Control starts; (b) Control ends.

Circles in Figure 2 represent constraint ranges. The difference between clustering in mind (see Figure 1) and herding in reality (see Figure 2): for clustering, $P(x)$ is fixed, and the herd ratios are objective; for herding, $P(x)$ is transferred to the target areas, and the herd ratios are adjusted according to the minimum control cost criterion. For example, when clustering, the proportions of the two groups of sheep in the middle are larger; when herding, the proportions of the two groups on the sides are larger because it is easier to drive sheep there. In addition, the centers of four groups of sheep in Figure 2b deviate from the target centers, which is also for the sake of control cost. We can also increase the constraint strength to drive the sheep to the ideal position. However, the control cost must also be considered. Both sheep clustering and herding require solving latent variables.

3. The Semantic Information G Theory and the Maximum Information Efficiency Principle

3.1. The P-T Probability Framework

Why do we need the P-T probability framework? The reasons are:

1. The P-T probability framework [26] allows us to use truth, membership, similarity, and distortion functions as constraints in addition to the likelihood function to solve the latent variables.
2. A hypothesis or label, such as "adult", has two probabilities: statistical probability, defined by Mises [35], and logical probability, defined by Kolmogorov [36]. The former is normalized, whereas the latter is not.

The P-T probability framework is denoted by a five-tuple ($\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{B}$, P, T), where

 doi:10.20944/preprints202504.2525.v1

- $\boldsymbol{U}$=$\{x_1, x_2, \ldots\}$ is a set of instances, $X \in \boldsymbol{U}$ is a random variable;
- $\boldsymbol{V}$=$\{y_1, y_2, \ldots\}$ is a set of labels or hypotheses, $Y \in \boldsymbol{V}$ is a random variable;
- $\boldsymbol{B}$=$\{\theta_1, \theta_2, \ldots\}$ is the set of subsets of $\boldsymbol{U}$; every subset θ_j has a label $y_j \in \boldsymbol{V}$.
- P is the probability of an element in $\boldsymbol{U}$ or $\boldsymbol{V}$, i.e., the statistical probability; it is defined with "=", such as $P(x_i)=P(X=x_i)$.
- T is the probability of a subset of V or an element in $\boldsymbol{B}$, i.e., the logical probability; it is defined with "∈", such as $T(y_j)=P(X \in \theta_j)$.

In addition, we assume θ_j is a fuzzy set and also a model parameter.

The truth value of y_j for given x is the membership grade of x in θ_j, which is also the conditional logic probability of y_j, namely:

$$T(y_j \mid x) \equiv T(\theta_j \mid x) \equiv m_{\theta j}(x). \tag{1}$$

According to Davidson's truth-conditional semantics [39], $T(y_j \mid x)$ reflects the semantics of y_j. The logical and statistical probabilities of a label are often not equal. For example, the logical probability of a tautology is 1, while its statistical probability is close to 0. We have $P(y_1) + P(y_2) + \ldots + P(y_n) = 1$, but it is possible that $T(y_1) + T(y_2) + \ldots + T(y_n) > 1$.

According to the above definition, we have:

$$T(y_j) \equiv T(\theta_j) \equiv P(X \in \theta_j) = \sum_i P(x_i) T(\theta_j \mid x_i) \cdot \tag{2}$$

As we will see later, $T(y_j)$ is the statistical physics' partition function and machine learning's regularization term. We can put $T(\theta_j \mid x)$ and $P(x)$ into Bayes' formula to obtain the semantic probability prediction formula [25]:

$$P(x \mid \theta_j) = \frac{T(\theta_j \mid x) P(x)}{T(\theta_j)}, \quad T(\theta_j) = \sum_i T(\theta_j \mid x_i) P(x_i). \tag{3}$$

$P(x \mid \theta_j)$ is the likelihood function $P(x \mid y_j, \theta)$ in the popular method. We call the above formula the semantic Bayes' formula.

Just as a set of transition probability functions $P(y_j \mid x)$(j=1, 2, …) constitutes a Shannon channel, a set of truth functions $T(\theta_j \mid x)$ (j=1,2,…) constitutes a semantic channel.

The relationship between the truth function and the distortion function is [29]:

$$T(y_j \mid x) \equiv \exp[-d(x, y_j)], \quad d(x, y_j) \equiv -\log T(y_j \mid x). \tag{4}$$

3.2. The Semantic Information Measure and Information Rate-Fidelity Function

Shannon MI can be expressed as.

$$I(X;Y) = \sum_j \sum_i P(x_i) P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)} = H(X) - H(X \mid Y), \tag{5}$$

We replace $P(x_i \mid y_j)$ on the right side of the log with the likelihood function $P(x_i \mid \theta_j)$, leaving the left $P(x \mid y_j)$ unchanged. Hence, we get the semantic MI:

$$\begin{aligned} I(X;Y_\theta) &= \sum_j \sum_i P(x_i) P(x_i \mid y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)} \\ &= \sum_j \sum_i P(x_i) P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \\ &= H(X) - H(X \mid Y_\theta) = H(Y_\theta) - H(Y_\theta \mid X) = H(Y_\theta) - \overline{d}. \end{aligned} \tag{6}$$

The Explanation of the last line can be found in Appendix C, which also includes the Logical Bayes' Inference [25] for solving $T(\theta_j \mid x)$ from $P(y_j \mid x)$ or $P(x \mid y_j)$ and $P(x)$.

 doi:10.20944/preprints202504.2525.v1

Roughly speaking, semantic posterior entropy $H(X|Y_\theta)$ is VFE F in VB and the MFE principle. The smaller it is, the greater the amount of semantic information.

Semantic MI is less than or equal to the Shannon MI and reflects the average code length saved due to the semantic prediction. It is easy to see that the maximum semantic MI criterion is equivalent to the maximum likelihood criterion and is similar to the Regularized Least Squares (RLS) criterion. Semantic entropy $H_\theta(Y)$ is the regularization term. Fuzzy entropy $H(Y_\theta|X)$ is a more general average distortion than the average square error.

Suppose the truth function becomes a similarity function. In that case, the semantic MI becomes the estimated MI [25]. For example, when calculating the information of the GPS pointer or the color perception, the truth function becomes the similarity function, which can be represented by a Gaussian function [25]. The estimated MI has been used by deep learning researchers for Mutual Information Neural Estimation (MINE) [39] and Information Noise Contrast Estimation (InfoNCE) [40].

Shannon [30] defines that given a source $P(x)$, a distortion function $d(y|x)$, and the upper limit D of the average distortion $\bar{d}$, we change the channel $P(y|x)$ to find the minimum MI, $R(D)$. $R(D)$ is the information rate-distortion function.

Now, we replace $d(y_j|x_i)$ with $I(x_i; \theta_j)=\log[T(\theta_j|x_i)/T(\theta_j)$, replace $\bar{d}$ with $I(X; Y_\theta)$, and replace D with the lower limit G of the semantic MI to find the minimum Shannon MI, $R(G)$. $R(G)$ is the information rate-fidelity function. Because G reflects the average code length saved due to semantic prediction, using G as the constraint is more consistent in shortening the code length, and G/R can better represent information efficiency.

The $R(G)$ function is defined as

$$R(G)=\min_{P(Y|X):I(X;\theta)\geq G} I(X;Y). \tag{7}$$

We use the Lagrange multiplier method to find the minimum MI. The constraint conditions include $I(X; Y_\theta)\geq G$ and

$$\sum_j P(y_j|x_i)=1,\ i=1,2,\ldots; \qquad \sum_j P(y_j)=1. \tag{8}$$

The Lagrangian function is:

$$L(P(y|x),P(y)) = I(X;Y) - sI(X;Y_\theta) - \mu_i\sum_j P(y_j|x_i) - \alpha\sum_j P(y_j)$$

Using $P(y|x)$ as a variation, we let $\partial L/\partial P(y_j|x_i)=0$. Then, we obtain:

$$P^*(y_j|x_i)=P(y_j)m_{ij}^s \Big/ \lambda_i,\ \lambda_i=\sum_j P(y_j)m_{ij}^s,\ i=1,2,\ldots;\ j=1,2,\ldots \tag{10}$$

where $m_{ij}=P(x_i|\theta_j)/P(x_i)=T(\theta_j|x_i)/T(\theta_j)$. Using $P(y)$ as a variation, we let $\partial L/\partial P(y_j)=0$. Then, we obtain:

$$P^*(y_j)=\sum_i P(x_i)P(y_j|x_i), \tag{11}$$

where $P^*(y_j)$ means the next $P(y_j)$. Because $P^*(y|x)$ and $P^*(y)$ are interdependent, we can first assume a $P(y)$ and then repeat the above two formulas to obtain convergent $P^*(y)$ and $P^*(y|x)$ (see [36] (P. 326)). We call this method the Minimum Information Difference (MID) iteration. Someone may ask: Why do we obtain Equation (11) through variational methods instead of directly using Equation (11)? If we use Equation (11) directly, we still need to prove that $P^*(y)$ reduces R–G.

The parameter solution of the $R(G)$ function (as illustrated in Figure 3) is:

 doi:10.20944/preprints202504.2525.v1

$$G(s)=\sum_i\sum_j P(x_i)P(y_j\mid x_i)I_{ij}=\sum_i\sum_j I_{ij}P(x_i)P(y_j)m_{ij}^{\ s}/Z_i,$$
$$R(s)=sG(s)-\sum_i P(x_i)\log Z_i,\quad Z_i=\sum_k P(y_k)m_{ij}^{\ s}. \tag{12}$$

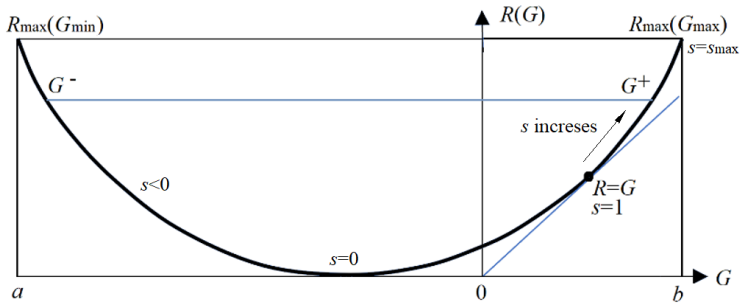


Figure 3. The information rate-fidelity function $R(G)$ for binary communication. Any $R(G)$ function is a bowl-like function. There is a point at which $R(G) = G$ ($s = 1$). For given R, two anti-functions exist: $G^-(R)$ and $G^+(R)$.

Any $R(G)$ function is bowl-shaped (possibly not symmetrical) [24], with the second derivative greater than 0. The $s = dR/dG$ is positive on the right. When $s = 1$, G equals R, meaning the semantic channel matches the Shannon channel. G/R represents information efficiency; its maximum is 1. G has a maximum value, G^+, and a minimum value, G^-, for given R. G^- means how small the semantic information the receiver receives can be when the sender intentionally lies.

3.3. Semantic Variational Bayes and the Maximum Information Efficiency Principle

Using $P(x\mid\theta_i)$ as the variation and letting $\partial L/\partial P(x_i\mid\theta_j)=0$, we get $P(x\mid\theta_j)=P(x\mid y_j)$ or $T(\theta_j\mid x)\propto P(y_j\mid x)$, which is the result of LBI. Therefore, the MID iteration plus LBI plus equals SVB. When using SVB to solve the latent variables, the constraint functions can be likelihood functions, truth (or membership) functions, similarity functions, and distortion functions (or loss functions). When the constraint is a set of likelihood functions, the MID iteration formulas are:

$$P^*(y_j\mid x_i)=P(y_j)[P(x_i\mid\theta_j)]^s\Big/Z_i,\ Z_i=\sum_j P(y_j)[P(x_i\mid\theta_j)]^s,$$
$$P^*(y_j)=\sum_i P(x_i)P(y_j\mid x_i). \tag{13}$$

When the constraint is a set of truth or similarity functions, the formulas become:

$$P^*(y\mid x_i)=P(y)\left[\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\right]^s\Big/Z_i,\ Z_i=\sum_k P(y_k)\left[\frac{T(\theta_k\mid x_i)}{T(\theta_k)}\right]^s,$$
$$P^*(y_j)=\sum_i P(x_i)P(y_j\mid x_i). \tag{14}$$

The hyperparameter s allows us to strengthen the constraint to reduce the ambiguity of boundaries. Therefore, the Shannon channel can also be regarded as a fuzzy classification function. Ambiguity reduces semantic MI but saves Shannon MI.

Under the premise that G is large enough, using the MID criterion means using the MIE criterion. Applying this criterion to various fields is to use the MIE principle. Note that the MIE principle does not only maximize G/R; it maximizes G/R under the condition that G meets the requirement. This requires us to balance between maximizing G and maximizing G/R. As a constraint condition, G is not only about limiting the amount of semantic information but also about specifying what kind of semantic information is needed. It is related to human purposes and needs. Therefore, the MIE principle is associated with information values.

 doi:10.20944/preprints202504.2525.v1

4. The Maximum Information Efficiency Principle for Mixture Models and Constrained Control

4.1. New Explanation of the EM Algorithm for Mixture Models (About Sheep Clustering)

The EM algorithm [41,42] is usually used for the mixture model, an unsupervised learning (clustering) method. We know that $P(x) = \sum_j P(y_j)\, P(x|y_j)$. Given a sample distribution $P(x)$, we use $P_\theta(x) = \sum_j P(x|\theta_j)P(y_j)$ to approximate $P(x)$ so that the relative entropy or KL divergence $KL(P\|P_\theta)$ is close to 0.

The EM algorithm first presets $P(x|\theta_j)$ and $P(y_j)$. The E-step obtains:

$$P(y_j|x) = P(y_j)P(x|\theta_j)/P_\theta(x),\ P_\theta(x) = \sum_k P(y_k)P(x|\theta_k). \tag{15}$$

In the M-step, the log-likelihood of the complete data (usually represented by Q) is maximized. The M-step can be divided into two steps: the M1-step for

$$P^{+1}(y_j) = \sum_i P(x_i)P(y_j|x_i), \tag{16}$$

and the M2-step for

$$P(x|\theta_j^{+1}) = P(x)P(y_j|x)/P^{+1}(y_j). \tag{17}$$

For Gaussian mixture models, we can use the expectation and standard deviation of $P(x)P(y_j|x)/P^{+1}(y_j)$ as those of $P(x|\theta_j^{+1})$. $P^{+1}(y)$ is the above $P^*(y)$.

From the perspective of the G theory, the M2-step is to make the semantic channel match the Shannon channel, the E-step is to make the Shannon channel match the semantic channel, and the M1-step is to make the destination $P(y)$ match the source $P(x)$. Repeating the above three steps can make the mixture model converge.

However, there are two problems with the EM algorithm: 1) $P(y)$ may converge slowly; 2) If the likelihood functions are also fixed, how do we solve $P(y)$?

Based on the $R(G)$ function analysis, the authors improved the EM algorithm to the EnM algorithm [25,31]. The E-step in the EnM algorithm remains unchanged, and the M-step is the M2-step in the EM algorithm. In addition, The n-step is added after the E-step. It repeats Equations (15) and (16) to calculate $P(y)$ n times so that $P^{+1}(y) \approx P(y)$. The EnM algorithm also uses the MIE criterion. The n-step can speed up $P(y)$ matching $P(x)$. The M-step only optimizes likelihood functions. Because after n-step, $P(y_j)/P^{+1}(y_j)$ is close to 1, we can use the following formula to optimize the model parameters:

$$P(x|\theta_j^{+1}) = P(x)P(x|\theta_j)/P_\theta(x). \tag{18}$$

Without the n-step, there will be $P(y_j) \neq P^{+1}(y_j)$, and $\sum_i P(x_i)P(x|\theta_j)/P\theta(x_i) \neq 1$.

When solving the mixture model, we can choose a smaller n, such as n=3. When solving $P(y)$ specifically, we can select a larger n until $P(y)$ converges. When n=1, the EnM algorithm becomes the EM algorithm.

We can deduce that after the E-step, there is (see Appendix D for the proof):

$$KL(P\|P_\theta) = R - G + KL(P_Y^{+1}\|P_Y), \tag{19}$$

where $KL(P||P_\theta)$ is the relative entropy or KL divergence between $P(x)$ and $P_\theta(x)$; $KL(P_Y^{+1}||P_Y)$ is $KL(P^{+1}(y))||P(y))$, which is close to 0 after the M1 step or the n-step.

Equation (19) can be used to prove the convergence of mixture models because the M2-step maximizes G, and the E-step and n-step minimize R–G and $KL(P_Y^{+1}||P_Y)$, $H(P\|P_\theta)$ can be close to 0. The MIE principle is used. Experiments have shown that as long as the sample was large enough, Gaussian mixture models of n=2 would converge globally.

 doi:10.20944/preprints202504.2525.v1

If the likelihood functions are fixed, the EnM algorithm becomes the En algorithm. The En algorithm can be used to find the latent variable $P(y)$ for fixed constraint functions.

4.2. Goal-Oriented Information and Active Inference (About Sheep Herding)

In the above, we use the G measure to measure semantic information in communication systems, requiring that the prediction $P(x|\theta_j)$ conforms to the fact $P(x|y_j)$. Goal-oriented information is the opposite, requiring the fact to conform to the goal. We also call this information control information or purposeful information [28].

An imperative sentence can be regarded as a control instruction. We need to know whether the control result conforms to the control goal. The more consistent the result is, the more information there is. A likelihood function or a truth function can represent a control goal. The following goals can be expressed as truth functions:

- "The grain production should be close to or exceeds 7500 kg/hectare";
- "The age of death of the population should preferably exceed 80 years old";
- "The cruising range of electric vehicles should preferably exceed 500 kilometers";
- "The error of train arrival time should preferably not exceed 1 minute".

Semantic KL information can be used to measure purposeful information:

$$I(X;a_j/\theta_j)=\sum_i P(x_i|a_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)},$$

In this formula, θ_j indicates that the control target is a fuzzy range, and a_j is an action selected for task y_j. In the above formula, y_j is replaced with a_j. One reason is that we may select different a_j for the same task y_j; another reason is that a_j is used in the popular active inference method for the same purpose.

If there are several control targets y_1, y_2, … we can use the semantic MI formula to express the purposeful information:

$$I(X;A/\theta)=\sum_j P(a_j)\sum_i P(x_i|a_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}, \tag{21}$$

where A is a random variable taking a value a_j. Using SVB, the control ratio $P(a)$ can be optimized to minimize the control complexity (i.e., Shannon MI) for given $I(X; A/\theta)$.

Goal-oriented information can be regarded as the cumulative reward function in reinforcement learning. However, the goal here is a fuzzy range representing a plan, command, or imperative sentence. The optimization task is similar to the active inference task using the MFE principle. For a multi-target task, the objective function to be minimized is:

$$f = I(X; A) - sI(X;A/\theta). \tag{22}$$

When the actual distribution $P(x|a_j)$ is close to $P(x|\theta_j)$, the information efficiency reaches its maximum value,1. To further increase both information, we can use the MID iteration with s to get $P(a_j|x)$ and $P(a)$ and use Bayes' formula to obtain

$$P^*(x_i|a_j)=P(a_j|x_i)P(x_i)/P(a_j)=P(x_i)m_{ij}^s\Big/\sum_k P(x_k)m_{kj}^s. \tag{23}$$

We may use the likelihood function $P(x|\theta_j)$ or the truth function $T(\theta_j|x)$ for the sheep-herding task to represent goals, respectively. We may change Equation (23) for different constraint functions, referring to Equations (13) and (14). Compared with VB, the above method is simpler and can change the constraint strength by s.

For both communication and constraint control (or active inference), G stands for the effect, R for the cost, and G/R for the efficiency. We collectively refer to G/R in both cases as information efficiency.

 doi:10.20944/preprints202504.2525.v1

5. The Relationship Between Information and Physical Entropy and Free Energy

5.1. Entropy, Information, and Semantic Information in Local Non-equilibrium and Equilibrium Thermodynamic Systems

Gibbs set up the relationship between thermodynamic entropy and Shannon's entropy; Jaynes [20,21] proved that according to Stirling's formula, $\ln N! = N\ln N - N$ (when $N\to\infty$), there is a simple connection between Boltzmann's microscopic state number Ω of N molecules and Shannon entropy:

$$S = \mathrm{k}\ln\Omega = \mathrm{k}\ln\frac{N!}{\prod_{i=1}^{G_m} N_i!} = -\mathrm{k}N\sum_{i=1}^{G_m} P(x_i \mid T_0)\ln P(x_i \mid T_0) = \mathrm{k}NH(X \mid T_0), \tag{24}$$

where S is entropy, k is the Boltzmann constant, x_i is the i-th microscopic state (i =1,2,…, G_m; G_m is the microscopic state number of one molecule), N is the number of molecules that are mutually independent, N_i is the number of molecules with x_i, and T_0 is the absolute temperature. $P(x_i \mid T_0) = N_i/N$ represents the probability of a molecule in a state x_i at temperature T_0. The Boltzmann distribution for a given energy constraint is:

$$P(x_i \mid T_0) = \exp(-\frac{e_i}{\mathrm{k}T_0}) / Z', \quad Z' = \sum_i \exp(-\frac{e_i}{\mathrm{k}T_0}), \tag{25}$$

where Z' is the partition function.

Information and entropy in Thermodynamic Systems have been discussed by researchers [15,41], but the following methods and conclusions are different.

Considering the information between temperature and molecular energy, we use Maxwell-Boltzmann statistics [42] (refer to Figure 4). Now, x_i becomes energy e_i (i=1,2,…,m); g_i stands for the microscopic state number of a molecule with energy e_i (i.e., degeneracy), and N_i for the number of molecules with energy e_i.

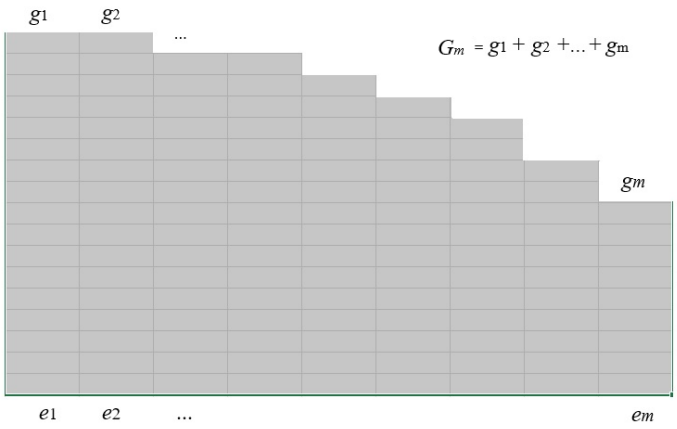


Figure 4. Degeneracy g_i is the microstate number of a molecule with energy e_i.

According to the classical probability definition, the prior probability of each microscopic state of a molecule is $P(x_i)$=1/G_m; the prior probability of a molecule with energy e_i is $P(e_i)$= g_i/G_m. The posterior probability is $P(e_i \mid T_0)$= N_i/N. So, Equation (24) becomes:

$$\begin{aligned} S &= \mathrm{k}\ln(N!\prod_{i=1}^{m}\frac{g_i^{N_i}}{N_i!}) = -\mathrm{k}N\sum_{i=1}^{m} P(e_i \mid T_0)\ln\frac{P(e_i \mid T_0)}{g_i} \\ &= -\mathrm{k}N\sum_i P(e_i \mid T_0)\ln\frac{P(e_i \mid T_0)}{P(e_i)} + \mathrm{k}N\ln G = \mathrm{k}N[\ln G - KL(P(e \mid T_0) \,||\, P(e)]. \end{aligned} \tag{26}$$

 doi:10.20944/preprints202504.2525.v1

Under the energy constraint, when the system reaches equilibrium, Equation (25) becomes:

$$P(e_i \mid T) = P(e_i)\exp(-\frac{e_i}{kT}) / Z, \quad Z = \sum_i P(e_i)\exp(-\frac{e_i}{kT}). \tag{27}$$

Consider a local non-equilibrium system. Different regions y_j (j = 1, 2, …) of the system have different temperatures T_j (j = 1, 2, …). Hence, we have $P(y_j)=P(T_j)=N_j/N$ and $P(x \mid y_j)=P(x \mid T_j)$. From Equation (26), we obtain

$$\begin{aligned} I(E;Y) &= \sum_j P(y_j) KL(P(e_i \mid y_j) \| P(e_i) = \sum_j P(y_j)[\ln G_m - S_j / (\mathrm{k}N_j)] \\ &= \ln G_m - S(kN), \end{aligned} \tag{28}$$

where $\ln G_m$ is the prior entropy $H(X)$ of X. Let E be a random variable taking a value e. Since e is certain for a given X, there are $H(E \mid X)=0$ and $H(X, E)=H(X)=\ln G_m$. From Equation (28), we derive

$$I(E;Y) = \ln G_m - S/(kN) = H(X) - H(X \mid Y) = I(X;Y). \tag{29}$$

This formula indicates that the information about energy E provided by Y is equal to the information about microscopic state X, and $S/(kN)=H(X \mid Y)$.

According to formulas (27-29), when local equilibrium is reached, there is

$$\begin{aligned} I(X;Y) = I(E;Y) &= \sum_j \sum_i P(e_i, y_j) \ln \frac{\exp[-e_i / (\mathrm{k}T_j)]}{Z_j} \\ &= -\sum_j P(y_j) \log Z_j - \mathrm{E}(e/T)/(\mathrm{k}N) = H(Y_\theta) - H(Y_\theta \mid X) = I(X;Y_\theta), \end{aligned} \tag{30}$$

where $\mathrm{E}(e/T)$ is the average of e/T, which is similar to relative square error. It can be seen that in local equilibrium systems, minimum Shannon MI can be expressed by the semantic MI formula. Since $H(X \mid T)$ becomes $H(X \mid T_\theta)$, there is

$$S = \mathrm{k}NH(X \mid T_\theta) = \mathrm{k}NF, \tag{31}$$

which means that VFE, F, is proportional to thermodynamic entropy.

Why do physics and the G theory have the same forms of entropy and information? It turns out that the entropy and information in both are under some constraints. In physics, it is the energy constraint, while in the G theory, it is the extension constraint. There is a simple connection between the two: $T(\theta_j \mid x_i) = \exp[-e_i/(\mathrm{k}T_j)]$.

5.2. Information, Free Energy, Work Efficiency, and Information Efficiency

Helmholtz's free energy formula is:

$$F'' = U - T_0 S, \tag{32}$$

where F'' is free energy, and U is the system's internal energy. In an open system, the free energy may increase when the system changes from an equilibrium state to a non-equilibrium state. When U remains constant, there is

$$\Delta F'' = -\Delta(TS) = T_0 S - \sum_j T_j S_j = \mathrm{k}NT_0 H(X \mid T_0) - \mathrm{k}N\sum_j T_j H(X \mid Y). \tag{33}$$

If T_0 approaches ∞, $H(X \mid T_0)$ is close to $H(X)$. Comparing the above equation with the Shannon MI formula $I(X;Y)=H(X)-H(X \mid Y)$, we can find that Shannon MI is like the increment of free energy in a local non-equilibrium system.

In thermodynamics, exergy is the energy that can do work [43]. When a system's state is changed, exergy is defined as

$$Exergy = (E-E_0)+p_0(V-V_0)-T_0(S-S_0), \tag{34}$$

 doi:10.20944/preprints202504.2525.v1

where $E - E_0$ is the increment of the system's internal energy, p_0 and T_0 are the pressure and temperature of the environment, and $V–V_0$ is the increment of volume. In local non-equilibrium systems, if the volume and temperature of each local region are constant,

$$\Delta Exergy < \Delta F''. \tag{35}$$

When the system reaches local equilibrium,

$$\Delta Exergy = \Delta F''. \tag{36}$$

It can be seen that semantic MI is equivalent to the increment of free energy, that is, the increment of exergy, in a local equilibrium system. Semantic MI is less than or equal to Shannon MI, just as exergy is less than or equal to free energy F''.

We can also regard kNT_0 and kNT_j as the unit information values [24], so $\Delta Exergy$ is equivalent to the increase in information value.

Generally speaking, the larger the free energy, the better. Only when free energy is used to do work do we want to consume less free energy. Similarly, only when Shannon information is consumed to transmit semantic information do we want to consume less Shannon information. Semantic information G parallels work W. The ratio $W/\Delta F''$ reflects work efficiency; similarly, the ratio G/R reflects information efficiency.

6. The MFE Principle and Inconsistency between Theory and Practice

6.1. VFE as the Objective Function and VB for Solving Latent Variables

Hinton and Camp [1] provided the following formula:

$$F = \sum_j r_j E_j - \sum_j r_j \log \frac{1}{r_j}, \tag{37}$$

where r_j is the above $P(y_j)$, and E_j is the encoding cost of x according to y_j, which is also called the reconstruction cost. F is called "free energy" because the formula is similar in form to the free energy formula in physics. In thermodynamics, minimizing free energy can obtain the Boltzmann distribution. Similarly, we can get

$$\begin{aligned} r_j &= \exp(-E_j) \Big/ \sum_j \exp(-E_j), \\ E_j &= H(X,\theta_j) = -\sum_i P(x_i \mid y_j) \log P(x_i, y_j \mid \theta). \end{aligned} \tag{38}$$

Hinton and Camp's variation method was developed into a more general VB method. The objective function of VB is usually expressed as [5]:

$$F = \sum_y g(y) \log \frac{g(y)}{P(x,y \mid \theta)} = -\sum_y g(y) \log P(x \mid y,\theta) + KL(g(y) \,||\, P(y)). \tag{39}$$

where $g(y)$ is $P^{+1}(y)$. Negative F is usually called the evidence lower bound, denoted by $L(g)$. In the above formula, x should be a vector and related to y. Using the semantic information method, we express F as

$$\begin{aligned} F &= \sum_i P(x_i) \sum_i P(y_j \mid x_i) \log \frac{P^{+1}(y_j)}{P(x_i, y_j \mid \theta)} \\ &= \sum_j P^{+1}(y_j) \sum_i P(x_i \mid y_j) \log \frac{P^{+1}(y_j)}{P(x_i \mid \theta_j) P(y_j)} = H(X \mid Y_\theta) + KL(P_Y^{+1} \,||\, P_Y). \end{aligned} \tag{40}$$

In the EM algorithm, after the M1-step or after iteration convergence, $P^{+1}(y)$ equals $P(y)$, and hence F equals $H(X|Y_\theta)$. So, the author said "roughly speaking, F=$H(X|Y_\theta)$." Hereafter, we assume

 doi:10.20944/preprints202504.2525.v1

that F is calculated after the M1-step, so $F = H(X|Y_\theta)$. Hence, the relationship between the information difference $R–G$ and F is:

$$R - G = I(X; Y) - I(X; Y_\theta) = H(X|Y_\theta) - H(X|Y) = F - H(X|Y). \tag{41}$$

To optimize $P(y)$, mean-field approximation [5] is often used to optimize $P(y|x)$ first and then get $P(y)$ from $P(y|x)$ and $P(x)$. This is to use $P(y|x)$ as the variation to minimize

$$F^{\#} = \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{P(x, y|\theta)}. \tag{42}$$

Minimizing $F^{\#}$ is equivalent to minimizing the cross entropy:

$$H_\theta(X) = -\sum_i P(x_i) \log P_\theta(x_i). \tag{43}$$

It is also equivalent to minimizing $KL(P||P_\theta)$ and $R - G$, which can make the mixture model converge. Therefore, the MIE criterion is used.

Neal and Hinton used the VB to improve the EM algorithm [2] to solve the mixture model. They defined

$$F(P(y),\theta) = \mathrm{E}_{P(x,y)} \log P(x, y|\theta) + H(Y) \tag{44}$$

as negative VFE. For convenience, we let $F' = F(P(y),\theta) = -F$.

Neal and Hinton showed that using the Incremental Algorithm (see Equation (7) in [2]) to maximize F' in both the E-step and the M-step can make mixture models converge faster. Their M-step is the same as the M-step in the EnM algorithm, but the E-step only updates one $P(y_j|x)$ each time, leaving $P(y_k|x)(k \neq j)$ unchanged, similar to the mean filed approximation used by Beal [3]. They actually also minimized $H_\theta(X)$, $KL(P||P_\theta)$, and $R–G$.

Experiments show that the EM algorithm, the EnM algorithm, the Incremental Algorithm, and the VB-EM algorithm [3] can all make mixture models converge, but unfortunately, during the convergence of mixture models, F' may not continue to increase, or F may not continue to decrease, which means that the calculation results of VB are correct, but the theory is imperfect.

Some people use the continuous increase of the complete data log-likelihood $Q = -H(X, Y_\theta)$ to explain or prove the convergence of the EM algorithm [33,34]. However, Q, like F', may also decrease during the convergence of mixture models.

6.2. The Minimum Free Energy Principle

Friston et al. first applied VB to brain science and later extended it to biobehavioral science, thus developing the MFE criterion into the MFE principle [6,7].

The MFE principle uses μ, a, s, and η to represent four states, respectively:

- μ: internal state; in SVB, it is θ in the likelihood function $P(x|\theta_j)$.
- a: subjective action; in SVB, it is y (for prediction) and a (for constraint control).
- s: perception; that is, the above observed datum x.
- η: external state, which is the previous y or $P(x|y)$; when SVB is used for constraint control, it is the external response $P(x|\beta_j)$, which is expected to be equal to $P(x|\theta_j)$.

According to the MFE principle, there are two types of optimization [6]:

$$\begin{aligned} \mu^* &= \arg\min_\mu \{F(\mu, a; s)\}, \\ a^* &= \arg\min_a \{F(\mu^*, a; s)\}. \end{aligned} \tag{45}$$

The first equation optimizes the likelihood function $P(s|\mu)$. The second equation optimizes the action selection: $P(a)$ and $P(a|s)$. The latter is also called active inference. This is similar to finding $P(a)$ and $P(a|x)$ when using SVB for constraint control.

 doi:10.20944/preprints202504.2525.v1

Friston sometimes interprets F as unexpectedness, surprise, and uncertainty and sometimes as error. Reducing F means reducing surprise and error. This is easy to understand from the perspective of the G theory. Because F is the semantic posterior entropy, reflecting the average code length of residual coding after the prediction. When the event conforms to the prediction or purpose, uncertainty and surprise are reduced, and semantic information increases. Reducing F means reducing the error because $G = H_\theta(Y) - \overline{d} = H(X) - F$, when $H_\theta(Y)$ and $H(X)$ are fixed, F and $\overline{d}$ increase or decrease at the same time.

6.3. The Progress from the ME Principle to the MFE Principle

In 1957, Jaynes proposed the ME principle [20,21]. He regards physical entropy as a special case of information entropy and provides a method for solving maximum entropy distributions. This method can be used to predict the probability distribution of a system state under certain constraints and to optimize the constraint control of random events. Compared with the Entropy Increase Law, the ME principle uses active constraint control, enabling its application to human intervention in nature. However, bio-organisms have purposes, and the ME principle cannot evaluate whether their prediction and control conform to the fact or purpose.

According to the MFE principle, the smaller the F, the more the subjective prediction $P(x|\theta_j)$ conforms to the objective fact $P(x|y_j)$. On the other hand, the active inference with the MFE principle makes the objective fact $P(x|y_j)$ closer to the subjective purpose $P(x|\theta_j)$. Both principles maximize the Shannon posterior entropy $H(X|Y)$. However, the MFE principle also optimizes the objective function using the maximum likelihood criterion commonly used in machine learning.

6.4. Why May VFE Increase during the Convergence of Mixture Models?

The author considered the properties of cross-entropy

$$H(X|\theta_j) = -\sum_i P(x_i|y_j)\log P(x_i|\theta_j)\,. \tag{46}$$

30 years ago [23]. When $P(x|\theta_j)$ approaches a fixed $P(x|y_j)$, $H(X|\theta_j)$ decreases. But conversely, when $P(x|y_j)$ approaches a fixed $P(x|\theta_j)$, will $I(X;\theta_j)$ increase or decrease? The conclusion is that it may increase or decrease. What is certain is that $KL(P(x|y_j)||P(x|\theta_j) = I(X;y_j) - I(X;\theta_j)$ will decrease. For example, $P(x|y_j)$ and $P(x|\theta_j)$ are two Gaussian distributions with the same expectation. And the standard deviation d_1 of $P(x|y_j)$ is smaller than the standard deviation d_2 of $P(x|\theta_j)$. The d_1 will increase while $P(x|y_j)$ approaches $P(x|\theta_j)$, and the cross-entropy will also increase.

Table 1 shows a simpler example, where x has four possible values. When $P(x|y_j)$ changes from the concentrated to the dispersed, $H(X|\theta_j)$ will increase.

Table 1. $H(X|\theta_j)$ increases when $P(x|y_j)$ is close to $P(x|\theta_j)$.

	x_1	x_2	x_3	x_4	$H(X\|\theta_j)$ (bits)
$P(x\|\theta_j)$	0.1	0.4	0.4	0.1	
$P(x\|y_j)$	0	0.5	0.5	0	log(10/4)= 1.32
$P(x\|y_j)$=$P(x\|\theta_j)$	0.1	0.4	0.4	0.1	0.2log(10)+0.8log(10/4)=1.72

This conclusion can be extended to the semantic MI formula, concluding that when the Shannon channel matches the semantic channel, $I(X;Y_\theta)$ and $H(X|Y_\theta) = F$ are uncertain to increase or decrease; what is certain is that $R–G$ must decrease. VB also makes the Shannon channel match the semantic channel, so during the matching process, $R - G = F - H(X|Y)$ instead of F will continue to decrease.

During the iteration of Gaussian mixture models, two causes affect the increase or decrease of F and $H(X|Y_\theta)$:

1) At the beginning of the iteration, the distribution range of $P(x|\theta_j)$ and $P(x|y_j)$ is quite different. After $P(x|\theta_j)$ approaches $P(x|y_j)$, F and $H(X|Y_\theta)$ will decrease.

 doi:10.20944/preprints202504.2525.v1

2) The true model's VFE $F=H(X|Y)$ (i.e., Shannon conditional entropy) is very large. During the iteration, F and $H(X|Y_\theta)$ may increase.

When reason 1) is dominant, F decreases; when reason 2) is dominant, F increases.

Suppose a Gaussian mixture model has two components; only two initial standard deviations are smaller than the true models' two standard deviations (see Figure 6a). During the iteration, F will continue to increase (see Section 7.1.2).

Asymmetric standard deviations and mixing ratios can also cause F to increase sometimes (see Section 7.1.1). In addition, the initial parameters μ_1 and μ_2 are biased to one side, probably making F increase (see Section 7.1.3).

7. Experimental Results

7.1. Proving Information Difference Monotonically Decreases During the Convergence of Mixture Models rather than VFE

7.1.1. Neal and Hinton's Example: Mixture Ratios Causes F′ and Q to Decrease

According to the popular view, during the convergence of mixture models, $F' = -H(X|Y_\theta) = -F$ and $Q = -H(X, Y_\theta)$ continue to increase. However, counterexamples are often seen in experiments. First, let's look at the example of Neal and Hinton [2] (see Table 2 and Figure 5). Table 2 shows the true and initial model parameters and the mixture ratios (the values of x below are magnified, and the magnified formula is $x = 20(x'-50)$ (x' is the original value in [2]).

Table 2. Neal and Hinton's mixture model example.

	True model's Parameters			Initial parameters		
	μ^*	σ^*	$P^*(y)$	μ	σ	$P(y)$
y_1	46	2	0.7	30	20	0.5
y_2	50	20	0.3	70	20	0.5

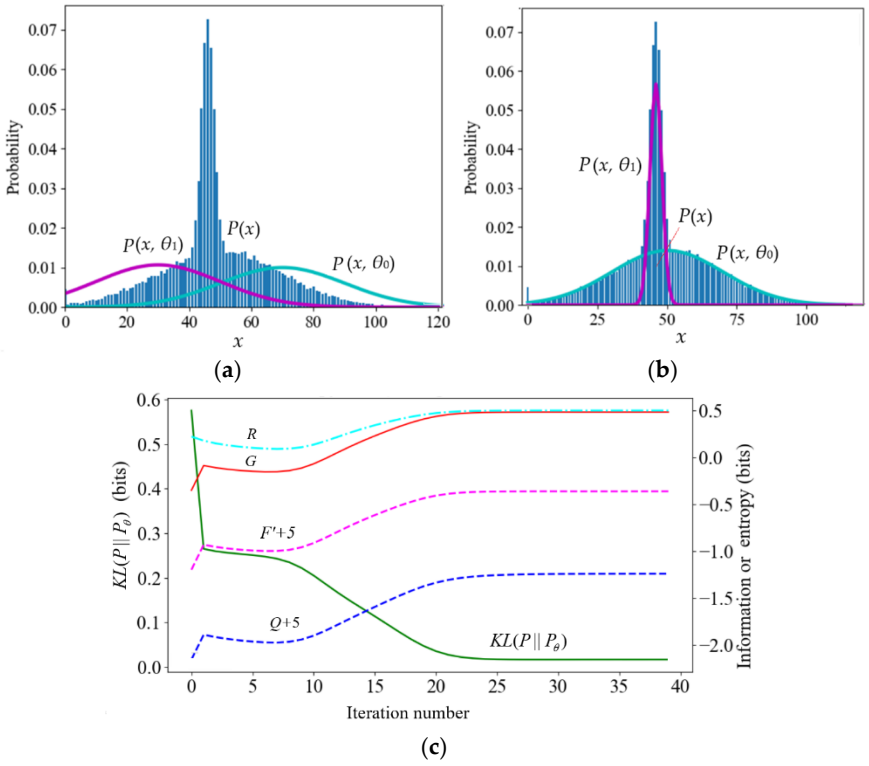


Figure 5. The convergence process of the mixture model used by Neal and Hinton. After the true model's mixture ratio was changed from 0.7:0.3 to 0.3:0.7, F' and Q decreased in some steps. (a) The iteration starts; (b) iteration converges; (c) R, G, F', and Q change in the iteration process.

 doi:10.20944/preprints202504.2525.v1

After the true model's mixture ratio was changed from 0.7:0.3 to 0.3:0.7, F' did not always increase. The reason was that the cross-entropy $H(X|y_2)$ of the second component of the true model was relatively large. So $H(X|Y_\theta)$ increased with $P(y_2)$. Later, F' eventually increased because the cross entropy decreased after $P(x|\theta_j)$ approached $P(x|y_j)$.

7.1.2. A Typical Counterexample against VB and the MFE Principle

Table 3 and Figure 6 show a mixture model where the initial two standard deviations are smaller than the two standard deviations of the true model. During the iteration process, F' and Q continued to decrease (except at the beginning).

Table 3. A mixture model whose F' and Q decreased in the convergent process.

	The true model's Parameters			Initial parameters		
	μ^*	σ^*	$P^*(Y)$	μ	σ	$P(Y)$
y_1	40	15	0.5	40	5	0.5
y_2	75	15	0.5	80	5	0.5

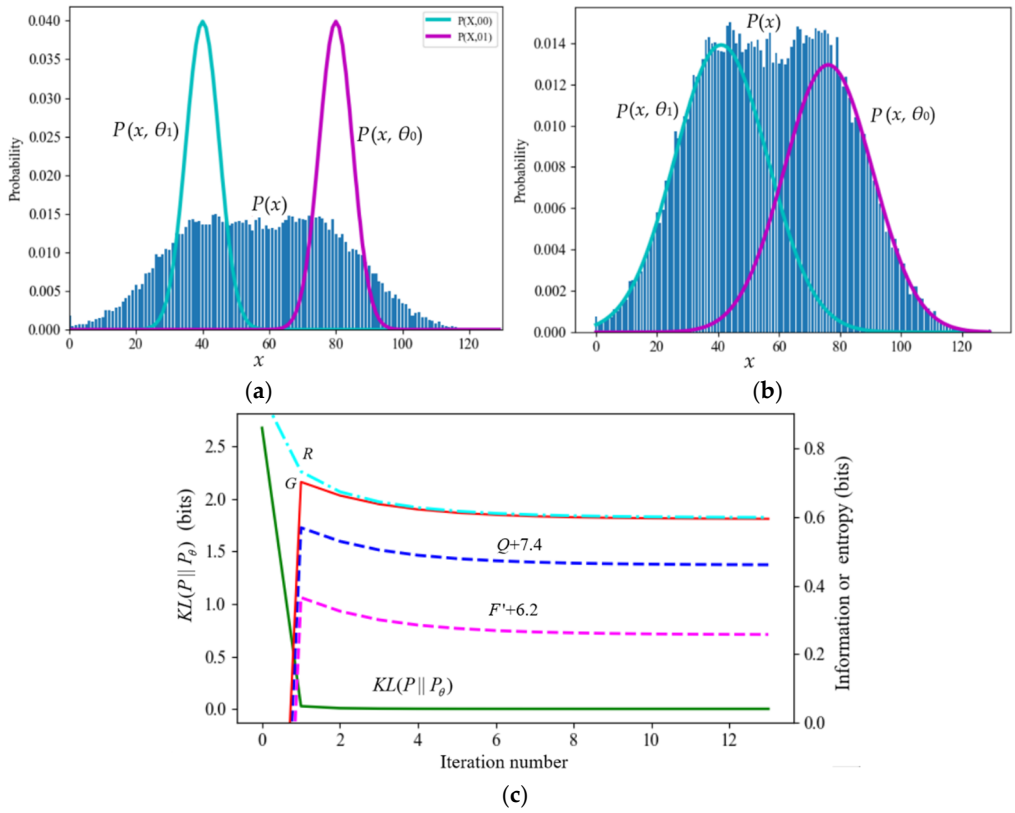


Figure 6. A typical mixture model against the MFE principle. (a) The iteration starts; (b) Iteration converges; (c) R, G, F', and Q change in the iteration process.

7.1.3. A Mixture Model Hard to Converges

Figure 7 shows an example from [34] that is hard to converge. The true model parameters are $(\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*, P^*(y_1)) = (100, 125, 10, 10, 0.7)$. To make convergence more difficult, we set the initial model parameters to $(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1)) = (80, 95, 5, 5, 0.5)$. Experiments showed that as long as the sample was large enough, the EM algorithm, the EnM algorithm, the Incremental algorithm [2], and the VBEM algorithm [3] could all converge. However, during the convergence process, only R–G and $KL(P||P_\theta)$ continued to decrease, while F' and Q did not continue to increase. This example indicates

 doi:10.20944/preprints202504.2525.v1

that if the initialization of μ_1 and μ_2 is inappropriate, F' and Q may also decrease during the iteration. The decrease in Q is more obvious.

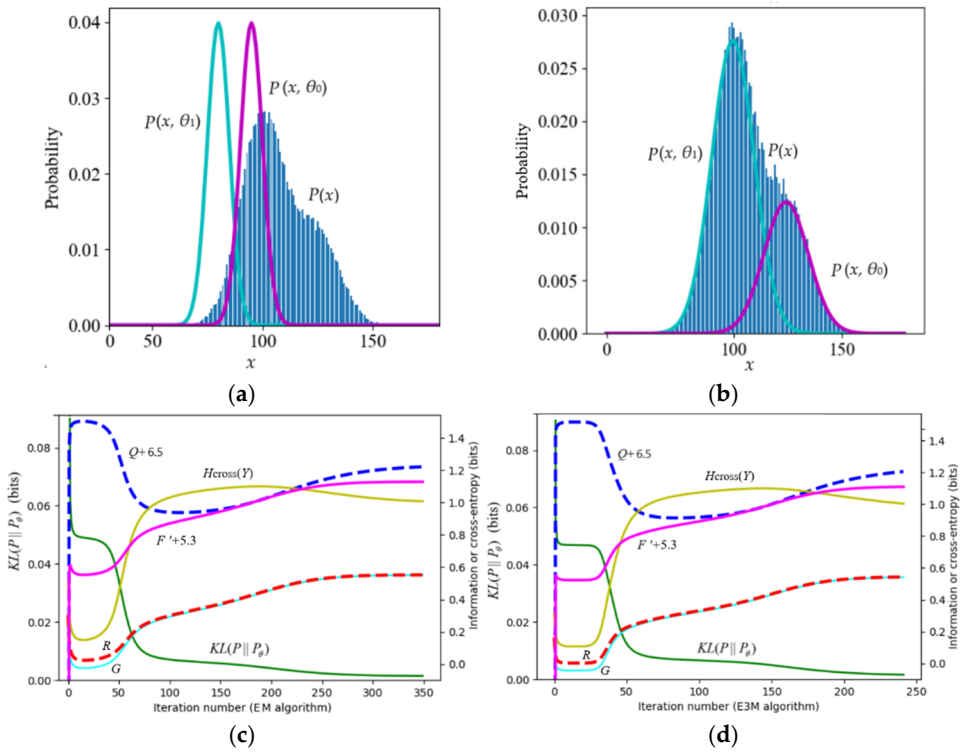
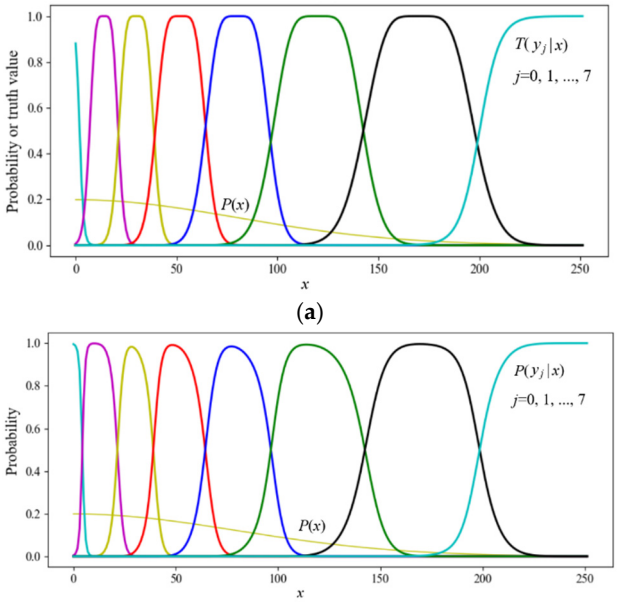


Figure 7. The mixture model that is hard to converge. $H_{cross}(Y) = -\sum_j P^{t+1}(y_j)\log P(y_j)$ is the cross entropy. (a) The iteration starts; (b) the iteration converges; (3) the iteration process of the EM algorithm; (d) the iterative process of the E3M algorithm.

This example also shows that the E3M algorithm requires fewer iterations (240 iterations) than the EM algorithm (350 iterations).

7.2. Simplified SVB (the En Algorithm) for Data Compression

The task was that 8-bit grayscale pixels (256 gray levels) were compressed into 3-bit pixels (8 gray levels). Considering that the eye grayscale discrimination is higher when the brightness is low, we used eight truth functions shown in Figure 8a as the constraint functions. Given $P(x)$ and $T(y|x)$, we found the Shannon channel $P(y|x)$ for the MIE.



 doi:10.20944/preprints202504.2525.v1

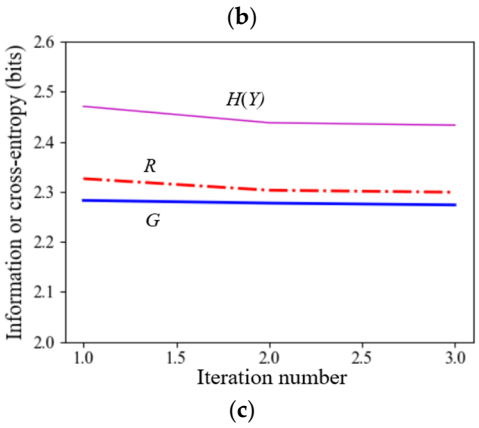


Figure 8. Using the En algorithm to optimize the Shannon channel $P(y|x)$. (a) Eight truth functions as the constraint; (b) the optimized $P(y|x)$; (c) R, G and $H(Y)$ change with the MID iteration.

$P(y)$ converged after repeating the MID iteration three times. At the beginning of the iteration, it was assumed that $P(y_j) = 1/8$ (j=0,1,…), and the entropy $H(Y)$ was 3 bits. When the iteration converged, R was 2.299 bits, G was 2.274 bits, and G/R was 0.989. These results mean that a 3-bit pixel can be transmitted with about 2.3 bits.

7.3. Experimental Results of Constraint Control (Active Inference)

We simplified the sheep-herding space into a one-dimensional space with only two pastures (see Figure 9) to show the relationship between the control results and the goals (the constraint ranges), and how the control results changed with s. We needed to solve the latent variable $P(a)$ according to $P(x)$ and the two truth functions.

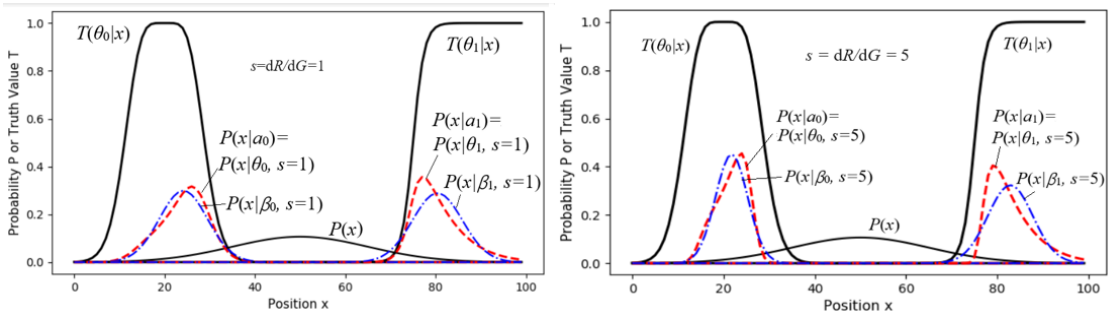


Figure 9. A two-objective control task. (a) For the case with s=1; (b) for the case with s=5. $P(x|\beta_j, s)$ is a normal distribution produced by the action a_j.

For different s, we set the initial ratios: $P(a_0)$=$P(a_1)$=0.5. Then, we used the MID iteration to obtain optimal $P(a_j|x)$ (j=0,1). Then, we got $P(x|a_j)$=$P(x|\theta_j, s)$ by

$$P(x_i \mid a_j, s) = P(a_j \mid x_i, s)P(x_i) / P(a_j) = P(x_i)m_{ij}^s \Big/ \sum_k P(x_k)m_{kj}^s . \quad (48)$$

Then, we used the parameter solution of the $R(G)$ function to obtain $G(s)$, $R(s)$, and $R(G(s))$. Figure 9a,b show $P(x|\theta_j, s)$ and $P(x|\beta_j, s)$ for s=1 and s=5, respectively. When s=5, the constraints are stricter, and some sheep at fuzzy boundaries are moved to more ideal positions. Figure 10 shows that when s > 5, G changes very little, indicating that we need to balance maximum purposeful information G and the MIE G/R. A larger s will reduce information efficiency and is unnecessary.

 doi:10.20944/preprints202504.2525.v1

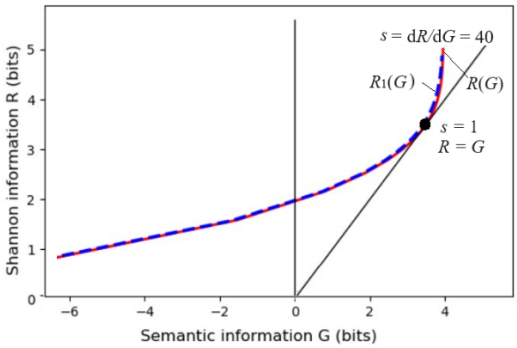


Figure 10. The $R(G)$ for constraint control. G slightly increases when s increases from 5 to 40, meaning s=5 is good enough.

The dashed line for $R_1(G)$ indicates that if we replace $P(x|a_j)$=$P(x|\theta_j, s)$ with a normal distribution, $P(x|\beta_j, s)$, G and G/R_1 do not obviously become worse.

For the constraint control of one task (Y=y_j), such as for the age control of death of adults by medical conditions, we can set $R = I(X; y_j)$, G=$I(X; \theta_j)$. The optimization method is similar, but there is no need to find latent variables. For details, see [28].

8. Discussion

8.1. Two Issues with VB and the MFE Principle and Their Solutions

Based on the previous analysis and experimental results, we identify two key issues with VB and the MFE principle.

The first issue concerns the inconsistency between theory and practice in VB. Although the computed results are correct, VFE does not always decrease monotonically during the convergence of mixture models. As demonstrated in Section 7.1 (see Figures 5–7), only the information difference $R-G$ consistently decreases.

The second issue relates to the interpretation of "free energy" in VB, which contradicts its physical meaning. As discussed in Section 5.1, $H(X|Y_\theta)$=F is proportional to physical entropy in a local equilibrium system. Consequently, interpreting $H(X|Y_\theta)$ as free energy introduces conceptual confusion.

To resolve the two issues with VB, Sections 3.3 and 3.4 introduce SVB, which is theoretically and practically consistent. The experiments in Section 7.1 demonstrate that throughout the iterative process of estimating latent variables, the information difference $R-G$ continuously decreases, or the information efficiency G/R steadily increases. According to the analysis in Section 5, information is like the increment of free energy. We can interpret that bio-organisms, particularly humans, can promote the Earth's order by acquiring more information and preserving more free energy.

8.2. Similarities and Differences Between SVB and VB

Both SVB and VB aim to solve two fundamental tasks:

1) Optimizing model parameters or likelihood functions.

2) Using variational methods to solve latent variables $P(y)$ according to observed data and constraints.

However, they differ in several key aspects:

- **Optimization Criteria**: Both VB and SVB optimize model parameters using the maximum likelihood criterion. When optimizing $P(y)$, VB nominally follows the MFE criterion but, in practice, employs the minimum KL divergence criterion (i.e., minimizing $KL(P||P_\theta)$ to make mixture models converge). This criterion is equivalent to the MIE criterion used in SVB.

 doi:10.20944/preprints202504.2525.v1

- **Variational Methods**: VB uses either $P(y)$ or $P(y|x)$ as the variation, whereas SVB alternatively uses $P(y|x)$ and $P(y)$ as the variation.
- **Computational Complexity**: VB relies on logarithmic and exponential functions to compute $P(y|x)$ [3,5], leading to relatively high computational complexity. In contrast, SVB offers a simpler approach to calculate $P(y|x)$ and $P(y)$ for the same task (s=1).
- **Constraint Functions**: VB can only use likelihood functions as the constraint. In contrast, SVB allows for various functions, including likelihood, truth, membership, similarity, and distortion functions. In addition, the constraints in SVB can be enhanced by the parameter s (see Figures 8 and 9).

SVB is potentially more suitable for various machine learning applications. However, because SVB does not consider the probability of the parameters, it may not be as applicable as VB on some occasions.

8.3. Optimizing Shannon Channel with The MFE or MIE Criterion

Shannon's information theory uses the distortion criterion instead of the information criterion when optimizing the Shannon channel for data compression. Hinton and Camp [1] initially used the MFE criterion to compress data, which is consistent with the purpose of reducing the residual coding length. VB's success in the field of machine learning reveals that VFE is more suitable for optimizing Shannon channels than distortion as an optimization criterion.

SVB uses the MID or MIE criterion, which is essentially the same as the MFE criterion. The MIE criterion is easier to understand and apply. In addition, SVB allows us to use truth, membership, similarity, and distortion functions as constraint functions for optimizing the Shannon channel (see section 7.2).

9. Conclusion

The MFE principle inherits the positive insight from EST that living systems increase the Earth's order by self-organization. Unlike the ME principle, the MFE principle implicitly combines the maximum likelihood criterion (for optimizing model parameters) and the ME principle (for maximizing conditional entropy $H(X|Y)$). This theoretical framework explains how bio-organisms predict and adapt to (even change) their environments. Furthermore, the optimization technique may be applied to promote the sustainable development of ecological systems.

However, the MFE principle faces two issues originating from VB. One issue is that the authors claim to minimize F in both optimization processes. However, in practice, what is minimized is $F - H(X|Y)$. The practice is correct, but the theory is incomplete. The reason is that in some cases, when $F - H(X|Y)$ decreases, F may increase. Another issue is that the MFE principle contradicts the concept of free energy in physics. In physics, the larger the free energy, the better. Only in a closed system will free energy passively decrease due to increased entropy. Actively reducing free energy contradicts the fundamental goal of increasing the Earth's order.

This paper proposes SVB and the MIE principle as the improved versions of VB and the MFE principle. SVB minimizes the difference between Shannon and semantic MI, i.e., $R - G = F - H(X|Y)$, rather than only minimizing F. This ensures the theoretical framework aligns with practical optimization methods. Additionally, SVB can simplify the algorithm for solving latent variables, and the MIE principle is easy to understand.

According to the analysis in Section 5, Shannon's information corresponds to the increment of free energy in local non-equilibrium systems; semantic information corresponds to the increment of exergy or the increment of free energy in local equilibrium systems; VFE F is equivalent to physical entropy in local non-equilibrium systems. We may say that acquiring and increasing information is similar to acquiring and increasing free energy; maximizing information efficiency parallels maximizing work efficiency. We will minimize Shannon's mutual information or consume free energy only when considering information or work efficiency.

 doi:10.20944/preprints202504.2525.v1

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author thanks the reviewers for their comments. The author also thanks Dr. Chuyu Xiong for reminding him of Friston's minimum free energy principle 5 years ago.

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

Appendix A. Abbreviations

Abbreviation	Original text
EM	Expectation-Maximization
En	Expectation-n
EnM	Expectation-n-Maximization
EST	Evolutionary System Theory
G theory	Semantic information G theory (G means generalization)
KL	Kullback–Leibler
LBI	Logical Bayes' Inference
ME	Maximum Entropy
MFE	Minimum Free Energy
MI	Mutual Information
MID	Minimum Information Difference
MIE	Maximum Information Efficiency
SVB	Semantic Variational Bayes
VFE	Variational Free Energy
VB	Variational Bayes

Appendix B. Python Source Codes Download Address

Python 3.6 source codes for eight figures in this paper can be downloaded from http://www.survivor99.com/lcg/Lu-py2025-2.zip.

Appendix C. Pome Generalized Entropies and Logical Bayes' Inference

In Equation (6), fuzzy entropy $H(Y_\theta|X)$, semantic entropy $H(Y_\theta)$, and semantic posterior entropy $H(X|Y_\theta)$ are

$$H(Y_\theta \mid X) = -\sum_j \sum_i P(x_i, y_j) \log T(\theta_j \mid x_i) = \bar{d} \,. \tag{a}$$

$$H(Y_\theta) = -\sum_i P(y_j) \log T(\theta_j) \,. \tag{b}$$

$$H(X \mid Y_\theta) = -\sum_j \sum_i P(x_i, y_j) \log P(x_i \mid \theta_j) = F \,. \tag{c}$$

$H(Y_\theta|X)$ equals $\bar{d}$ according to Equation (4). See Section 6.1 for $H(X|Y_\theta)= F$.

When $Y = y_j$, the semantic MI becomes semantic KL information:

$$I(X;\theta_j) = \sum_i P(x_i \mid y_j) \log \frac{P(x_i|\theta_j)}{P(x_i)} = \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \,. \tag{d}$$

When $P(x|\theta_j) = P(x|y_j)$, the semantic KL information reaches its maximum. Letting the maximum value of $T^*(\theta_j|x)$ be 1 and bringing $P(x|\theta_j)$ in Equation (3) into $P(x|\theta_j) = P(x|y_j)$, we can get the optimized truth function from the sample distribution:

 doi:10.20944/preprints202504.2525.v1

$$T^*(\theta_j \mid x) = \frac{P(x \mid y_j)}{P(x)} \Big/ \max_x \left(\frac{P(x \mid y)}{P(x)} \right) = \frac{P(y_j \mid x)}{\max_x (P(y_j \mid x))}. \tag{e}$$

Solving $T^*(\theta_j \mid x)$ with the above formula requires that the sample distribution is continuous and smooth. Otherwise, we need to use the following formula to get $T^*(\theta_j \mid x)$:

$$T^*(\theta_j \mid x) = \arg\max_{\theta_j} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \tag{f}$$

The above method for solving $T^*(\theta_j \mid x)$ is called Logical Bayes' Inference (LBI) [25].

Appendix D. The Proof of Equation (19)

After the M-step, the Shannon MI becomes:

$$R = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \frac{P(y_j \mid x_i)}{P^{+1}(y_j)}, \tag{g}$$

We define:

$$R'' = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)}. \tag{h}$$

Hence, $KL(P \| P_\theta) = R'' - G = R - G + KL(P_Y^{+1} \| P_Y)$.

References

1 Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. *In Proceedings of COLT*, pp. 5–13, 1993.

2 Neal, R.; Hinton, G. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*. Michael, I.J. Ed. *MIT Press: Cambridge, MA, USA, 1999; pp. 355–368*.

3 Beal, M.J. Variational algorithms for approximate Bayesian inference. Doctoral thesis (Ph.D), University College London, 2003.

4 Tran, M.; Nguyen, T.; Dao, V. A practical tutorial on Variational Bayes. Available online: https://arxiv.org/pdf/2103.01327. (accessed on 20 January 2025).

5 Wikipedia, Variational Bayesian methods, Available online: https://en.wikipedia.org/wiki/Variational_Bayesian_methods (accesses on 8 Feb. 2025).

6 Friston, K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **2010**, **11**, 127–138. https://doi.org/10.1038/nrn2787.

7 Friston, K.J., Parr, T., and de Vries, B. The graphical brain: Belief propagation and active inference. Network Neuroscience, **2017**, *1*:381–414.

8 Parr, T.; Pezzulo, G.; Friston, K.J. Active Inference: The Free Energy Principle in Mind, Brain, and Behavior, The MIT Press, 2022. https://doi.org/10.7551/mitpress/12441.001.0001.

9 Thestrup Waade, P.; Lundbak Olesen, C.; Ehrenreich Laursen, J.; Nehrer, S.W.; Heins, C.; Friston, K.; Mathys, C. As One and Many: Relating Individual and Emergent Group-Level Generative Models in Active Inference. *Entropy* **2025**, *27*, 143. https://doi.org/10.3390/e27020143.

10 Rifkin, T; Howard, T. *Thermodynamics and Society: Entropy. A New World View*. Viking, New York, 1980.

11 Ramstead, M.J.D.; Badcock, P.B.; Friston, K.J. Answering Schrödinger's question: A free-energy formulation. Phys Life Rev. **2018** *24*, 1-16. doi: 10.1016/j.plrev.2017.09.001.

12 Schrödinger, E. What Is Life? Cambridge: Cambridge University Press, *Cambridge*, UK, 1944

 doi:10.20944/preprints202504.2525.v1

13 Huang, G.T. Is this a unified theory of the brain? 28 May 2008 From NewScientist Print Edition. Available online: https://www.fil.ion.ucl.ac.uk/~karl/Is%20this%20a%20unified%20theory%20of%20the%20brain.pdf. (accessed on 10 January 2025).

14 Portugali, J. Schrödinger's What is Life?—Complexity, cognition and the city. Entropy **2023,** *25*, 872. https://doi.org/10.3390/e25060872.

15 Haken, H.; Portugali, J. Information and Selforganization: A Unifying Approach and Applications. *Entropy* **2016**, *18*, 197. https://doi.org/10.3390/e18060197

16 Haken, H.; Portugali, J. Information and Self-Organization II: Steady State and Phase Transition. *Entropy* **2021**, *23*, 707. https://doi.org/10.3390/e23060707

17 Martyushev, L.M. Living systems do not minimize free energy: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Dèsormeau Ramstead et al., Physics of Life Reviews, Volume 24, 2018, Pages 40-41, https://doi.org/10.1016/j.plrev.2017.11.010.

18 Gottwald S.; Braun1, D.A. The two kinds of free energy and the Bayesian revolution. Available online: https://arxiv.org/abs/2004.11763. (accessed on 20 January 2025).

19 Silverstein, S.D.; Pimbley, J.M. Minimum-free-energy method of spectral estimation: autocorrelation-sequence approach, *J. Opt. Soc. Am*. **1990**, *3*, 356-372.

20 Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620.

21 Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev. II* **1957**, *108*, 171.

22 Lu, C. Shannon equations reform and applications. *BUSEFAL* **1990**, *44*, 45–52. Available online: https://www.listic.univ-smb.fr/production-scientifique/revue-busefal/version-electronique/ebusefal-44/ (accessed on 5 March 2019).

23 Lu, C. *A Generalized Information Theory*; China Science and Technology University Press: Hefei, China, 1993; ISBN 7-312-00501-2. (in Chinese)

24 Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* **1999**, *28*, 453–490.

25 Lu, C. Semantic Information G Theory and Logical Bayesian Inference for Machine Learning. *Information,* **2019**, *10*, 261.

26 Lu, C. The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning. *Philosophies* **2020**, *5*, 25.

27 Kolchinsky, A.; Marvian, I.; Gokler, C.; Liu, Z.-W.; Shor, P.; Shtanko, O.; Thompson, K.; Wolpert, D.; Lloyd, S. Maximizing Free Energy Gain. *Entropy* **2025**, *27*, 91. https://doi.org/10.3390/e27010091.

28 Lu C. Semantic Variational Bayes Based on a Semantic Information Theory for Solving Latent Variables, Available online: https://doi.org/10.48550/arXiv.2408.13122. (accessed on 1 January 2025)

29 Lu, C. Using the Semantic Information G Measure to Explain and Extend Rate-Distortion Functions and Maximum Entropy Distributions. *Entropy* **2021**, *23*, 1050. https://doi.org/10.3390/e23081050.

30 Shannon, C.E. Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 1959, 4, 142–163.

31 Berger, T. *Rate Distortion Theory*; Prentice-Hall: Enklewood Cliffs, NJ, USA, 1971.

32 Zhou, J.P. Fundamentals of information theory, Beijing, China: People's Posts and Telecommunications Press, 1983. (in Chinese).

33 Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1997**, *39*, 1–38.

34 Ueda N.; Nakano, R. Deterministic annealing EM algorithm, *Neural Networks*, **1998**, *11*, 271-282, 1998.

35 Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitrechnung*; Ergebnisse Der Mathematik (1933); translated as *Foundations of Probability*; Dover Publications: New York, NY, USA, 1950.

36 von Mises, R. *Probability, Statistics and Truth*, 2nd ed.; George Allen and Unwin Ltd.: London, UK, 1957.

37 Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353.

38 Davidson, D. Truth and meaning. *Synthese* **1967**, *17*, 3, 304-323.

39 Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. MINE: Mutual information neural estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1–44. https://doi.org/10.48550/arXiv.1801.04062.

 doi:10.20944/preprints202504.2525.v1

40 Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with Contrastive Predictive Coding. Available online: https://arxiv.org/abs/1807.03748 (accessed on 10 January 2025).
41 Wikipedia, Maxwell-Boltzmann statistics , Available online: https://en.wikipedia.org/wiki/Maxwell%E2%80%93Boltzmann_statistics (accessed on 20 March 2025).
42 *Ben-Naim, A. Can entropy be defined for and the Second Law applied to the entire universe?* Available online: https://arxiv.org/abs/1705.01100 (accessed on 20 March 2025).
43 Bahrani, M. Exergy, Available online: https://www.sfu.ca/~mbahrami/ENSC%20461/Notes/Exergy.pdf (accessed on 23 March 2025).