Review

An Introduction to the Semantic Information G Theory and Applications

Chenguang Lu *

Posted Date: 12 February 2025

doi: 10.20944/preprints202502.0799.v1

Keywords: semantic information theory; semantic information measure; information rate-distortion; information rate-fidelity; variational Bayes; minimum free energy; maximum information efficiency; portfolio; information value; constraint control

Review

An Introduction to the Semantic Information G Theory and Applications

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: Does semantic communication require a semantic information theory parallel to Shannon's information theory, or can Shannon's work be generalized for semantic communication? This paper advocates for the latter and introduces the semantic information G theory (with "G" denoting generalization). The core approach involves replacing the distortion constraint with the semantic constraint, achieved by utilizing a set of truth functions as a semantic channel. These truth functions enable the expression of semantic distortion, semantic information measures, and semantic information loss. Notably, the maximum semantic information criterion is shown to be equivalent to the maximum likelihood criterion and parallels the Regularized Least Squares criterion. The G theory is compatible with machine learning methodologies, offering enhanced capabilities for handling latent variables, often addressed through Variational Bayes. This paper systematically presents the generalization of Shannon's information theory into the G theory and its wide-ranging applications. The applications involve semantic communication, machine learning, constraint control, Bayesian confirmation, portfolio theory, and information value. Furthermore, insights from statistical physics are discussed: Shannon information is equated to free energy, semantic information to the free energy of local equilibrium systems, and information efficiency to the efficiency of free energy in performing work. The paper also proposes refining Friston's minimum free energy principle into the maximum information efficiency principle. Lastly, it discusses the limitations of the G theory in representing the semantics of complex data.

Keywords: semantic information theory; semantic information measure; information rate-distortion; information rate-fidelity; variational Bayes; minimum free energy; maximum information efficiency; portfolio; information value; constraint control

1. Introduction

Although Shannon's information theory [1] has achieved remarkable success, it faces three significant limitations that restrict its semantic communication and machine learning applications. First, it cannot measure semantic information. Second, it relies on the distortion function to evaluate communication quality, but the distortion function is subjectively defined and lacks an objective standard. Third, it is challenging to incorporate model parameters into entropy formulas. In contrast, machine learning often requires cross-entropy and cross Mutual Information (MI) involving model parameters (Appendix A lists all abbreviations with original texts). Moreover, the minimum distortion criterion resembles the philosophy of "absence of fault is a virtue," whereas a more desirable principle might be "merit outweighing fault is a virtue." Why did Shannon's information theory use the distortion criterion instead of the information criterion? This is intriguing.

The study of semantic information gained attention soon after Shannon's theory emerged. Weaver initiated research on semantic information and information utility [2], and Carnap and Bar-Hillel proposed a semantic information theory [3]. Thirty years ago, the author of this article extended Shannon's theory to a semantic information framework [4–7], now known as the semantic

 doi:10.20944/preprints202502.0799.v1

information G theory (abbreviated as the G theory) [8]. Here, "G" stands for generalization, reflecting the G theory's role as a generalized form of Shannon's information theory. Earlier contributions to semantic information theories include works by Carnap and Bar-Hillel, Dretske [9], Wu [10], and Zhong [11], while more recent contributions after the author's generalization include those by Floridi [12,13] and others [14]. These theories primarily address natural language information and semantic information measures (upstream problems). In contrast, newer approaches have focused on electronic semantic communication over the past decade, particularly semantic compression (downstream problems) [14–17]. These explorations are highly valuable.

Researchers hold two extreme views on semantic information theory. One view argues that Shannon's theory suffices, rendering a dedicated semantic information theory unnecessary; at most, semantic distortion needs consideration; the opposing view advocates for a parallel semantic information theory alongside Shannon's framework. Among parallel approaches, some researchers (e.g., Carnap and Bar-Hillel) use only logical probability, avoiding statistical probability, while others incorporate semantic sources, semantic channels, semantic destinations, and semantic information rate distortion [17].

The G theory offers a compromise between these extremes. It fully inherits Shannon's information theory, including its derived theories. Only the semantic channel composed of truth functions is newly added. Based on Davidson's truth-conditional semantics [18], truth functions represent the extensions and semantics of concepts or labels. By leveraging the semantic channel, the G theory can:

1. derive the likelihood function from the truth function and source, enabling semantic probability predictions, thereby quantifying semantic information, and
2. replace the distortion constraint in Shannon's theory with semantic constraints, which include semantic distortion, semantic information quantity, and semantic information loss constraints.

The semantic information measure does not replace Shannon's information measure but supplants the distortion metric used to evaluate communication quality. Truth functions can be derived from sample distributions using machine learning techniques with the maximum semantic information criterion, addressing the challenges of defining classic distortion functions and optimizing Shannon channels with an information criterion. A key advantage of generalization over reconstruction is that semantic constraint functions can be treated as new or negative distortion functions, allowing the use of existing coding methods without additional electronic semantic communication coding considerations.

In addition to Shannon's ideas, the G theory integrates Popper's views on semantic information, logical probability, and factual testing [19,20]; Fisher's maximum likelihood principle [21]; and Zadeh's fuzzy set theory [22,23]. To unify Popper's logical probability with Zadeh's fuzzy sets, the author proposed the P-T probability framework [8,24], simultaneously accommodating both statistical and logical probabilities.

Thirty years ago, the G theory was applied to image data compression based on visual discrimination [5,7]. In the past decade, the author has introduced model parameters into truth functions, utilized truth functions as learning functions [7,25], and optimized them with sample distributions. The G theory has also been employed to optimize semantic communication for machine learning tasks, including multi-label learning, maximum MI classification, mixture models [8], Bayesian confirmation [26,27], semantic compression [28], constraint control [29], and latent variable solutions [30]. The concept of mutually aligning semantic and Shannon channels aids in understanding decentralized machine learning and reinforcement learning methods.

The main motivations and objectives of this paper are as follows:

Semantic communication urgently requires a semantic compression theory analogous to the information rate-distortion theory [31–33]. The author extends the information rate-distortion function to derive the information rate fidelity function $R(G)$(where R is the minimum Shannon MI for a given semantic MI G), which provides a theoretic foundation for semantic compression theory.

 doi:10.20944/preprints202502.0799.v1

Estimated MI (a specific case of semantic MI) [25,34,35] and Shannon MI minimization [36] have been utilized in deep learning. However, researchers often conflate estimated MI with Shannon MI and remain unclear about which should be maximized or minimized [37–39]. The G theory can clarify these distinctions.

The G theory has undergone continuous refinement, with many results scattered across more than 20 articles by the author. This paper aims to provide a comprehensive overview, helping future researchers avoid redundant efforts.

The remainder of this paper is organized as follows: Section 2 introduces the G theory; Section 3 discusses electronic semantic communication; Section 4 explores goal-oriented information and information value (in conjunction with portfolio theory); and Section 5 examines the G theory's applications to machine learning. The final section provides discussions and conclusions, including comparing the G theory with other semantic information theories, exploring the concept of information, and identifying the G theory's limitations and areas for further research.

2. From Shannon's Information Theory to the Semantic Information G Theory

2.1. Semantics and Semantic Probabilistic Predictions

Popper stated in his 1932 book *The Logic of Scientific Discovery* [19] (p. 102): the significance of scientific hypotheses lies in their predictive power, and predictions provide information; the smaller the logical probability and the more it can withstand testing, the greater the amount of information it provides. He also explicitly emphasized the necessity of distinguishing between two types of probability: statistical probability and logical probability. The G theory incorporates two kinds of probabilities and probability predictions, with T representing logical probability and truth value. Statistical probability predictions are expressed using Bayes' formula, and semantic probability predictions follow a similar approach.

The semantics of a word or label encompass both its connotation and extension. Connotation refers to an object's essential attributes, while extension denotes the range of objects the term refers to. For example, the extension of "adult" includes individuals aged 18 and above, while its connotation is "over 18 years old." Extensions for some concepts, like "adult," may be explicitly defined by regulations, whereas others, such as "elderly," "heavy rain," "excellent grades," or "hot weather," are more subjective and evolve through usage. Connotation and extension are interdependent; one can often infer one from the other.

According to Tarski's truth theory [40] and Davidson's truth-conditional semantics [9], a concept's semantics can be represented by a truth function, which reflects the concept's extension. For a crispy set, the truth function acts as the characteristic function of the set. For example, x is age, and y_1 is the label of the set {adult}, we denote the truth function as $T(y_1|x)$, which is also the characteristic function of the set {adult}.

In 1931, Popper put forward in the book "*The Logic of Scientific Discovery*" [39] (P.96) that the smaller the logical probability of a scientific hypothesis, the greater the amount of (semantic) information if it can stand the test. We can say that Popper is the earliest researcher of semantic information [19]. Later, he proposed a logical probability axiom system. He emphasized that there are two kinds of probabilities, statistical and logical probabilities, at the same time ([39] (pp. 252-258). But he had not established a probability system that includes both.

The truth function serves as the tool for semantic probability predictions (illustrated in Figure 1). The formula is:

$$P(x \mid y_1 \text{ is true}) = P(x)T(y_1 \mid x) / \sum_i P(x_i)T(y_1 \mid x_i) \quad . \tag{1}$$

If "adult" is changed to "elderly", the crispy set becomes a fuzzy set, the truth function is equal to the membership function of the fuzzy set, and the above formula remains unchanged.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

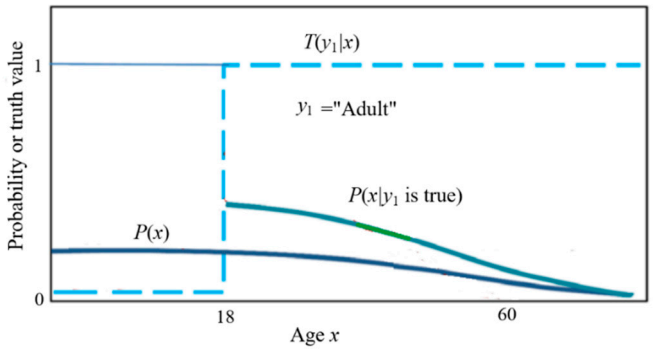


Figure 1. The semantic probability prediction according to that "x is adult" is true.

The extension of a sentence can be regarded as a fuzzy range in a high-dimensional space. For example, an instance described by a sentence with a subject, object, and predicate structure can be regarded as a point in the Cartesian product of three sets, and the extension of a sentence is a fuzzy subset in the Cartesian product. For example, the subject and the predicate are two people in the same group, and the predicate can be selected as one of "bully", "help", etc. The extension of "Tom helps Jone" is an element in the three-dimensional space, and the extension of "Tom helps an old man" is a fuzzy subset in the three-dimensional space. The extension of a weather forecast is a subset in the multidimensional space with time, space, rainfall, temperature, wind speed, etc., as coordinates. The extension of a photo or a compressed photo can be regarded as a fuzzy set, including all things with similar characteristics.

Floridi affirms that all sentences or labels that may be true or false contain semantics and provide semantic information [26]. The author agrees with this view and suggests converting the distortion function and the truth function $T(y_j|x)$ to each other. To this end, we define:

$$T(y_j|x) \equiv \exp[-d(y_j|x)], \quad d(y_j|x) \equiv -\log T(y_j|x). \tag{2}$$

where exp and log are a pair of inverse functions; $d(y_j|x)$ means the distortion when y_j represents x_i. We use $d(y_j|x)$ instead of $d(x, y_j)$ because the distortion may be asymmetrical.

For example, the pointer on a Global Positioning System (GPS) map has relative error or distortion; the distortion function can be converted to a truth function or similarity function:

$$T(y_j|x)=\exp[-d(y_j|x)]=\exp[-(x-x_j)^2/(2\sigma^2)], \tag{3}$$

where σ is the standard deviation; the smaller it is, the higher the precision. Figure 2 shows the mobile phone positioning seen by someone on a train.

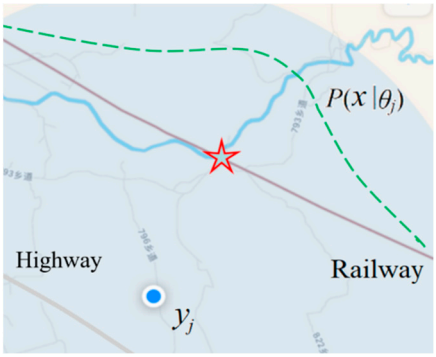


Figure 2. A GPS device's positioning with a deviation. The round point is the pointed position with a deviation, and the place with the star is the most possible.

According to the semantics of the GPS pointer, we can predict that the actual position is an approximate normal distribution on the high-speed rail, and the red five-star indicates the maximum possible position. If a person is on a specific highway, the prior probability distribution $P(x)$ will change, and the maximum possible position is the place closest to the small circle on that highway.

 doi:10.20944/preprints202502.0799.v1

Clocks, scales, thermometers, and various economic indices are similar to the positioning pointers and can all be regarded as estimates ($y_j=\hat{x}_j$), with error ranges or extensions, so they can all be used for semantic probability prediction and provide semantic information. Color perception can also be regarded as an estimate of color or color light. The higher the discrimination of the human eye (similar to the smaller σ), the smaller the extension. A Gaussian function can also express its truth function or discrimination function.

2.2. The P-T Probability Framework

Carnap and Bar-Hillel only use logical probability. We use logical probability, truth value, and statistical probability in the above semantic probability prediction. The truth value is conditional logical probability. The G theory is based on the P-T probability framework.

Why do we need a new probabilistic framework? Because a hypothesis or label, such as "adult", has two probabilities simultaneously. One is the probability of the set represented by the label, which is defined by Kolmogorov [41]; it is not normalized. Another is the probability in which the label is selected. It is defined by Mises [42]; it is normalized. The P-T probability framework attempts to unify the two probabilities and generalize the set to the fuzzy set.

We define:

1. X and Y are two random variables, taking $x \in U=\{x_1, x_2, \ldots\}$ and $y \in V=\{y_1, y_2, \ldots\}$ as their values. For machine learning, x_i is an instance, and y_j is a label or hypothesis; $y_j(x_i)$ is a proposition, and $y_j(x)$ is a proposition function.
2. The θ_j is a fuzzy subset of the domain U, whose elements make y_j true. We have $y_j(x)$ = "$x \in \theta_j$". The θ_j can also be understood as a model or a set of model parameters.
3. Probability defined by "=", such as $P(y_j)\equiv P(Y=y_j)$, is a statistical probability; probability defined by "$\in$", such as $P(X \in \theta_j)$, is a logical probability. To distinguish $P(Y=y_j)$ and $P(X \in \theta_j)$, we define the logical probability of y_j as $T(y_j)\equiv T(\theta_j)\equiv P(X \in \theta_j)$.
4. $T(y_j|x) \equiv T(\theta_j|x)\equiv P(X \in \theta_j|X=x)\in[0,1]$ is the truth function of y_j and also the membership function $m_{\theta j}(x)$ of the fuzzy set θ_j, that is,

$$T(y_j|x) \equiv T(\theta_j|x)\equiv m_{\theta j}(x). \quad (4)$$

The logical probability of a label is generally not equal to its statistical probability. 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 $T(y_1) + T(y_2) + \ldots + T(y_n) > 1$. For example, the age labels include "adult", "non-adult", "child", "youth", "elderly", etc., and the sum of their statistical probabilities is 1, while the sum of their logical probabilities is greater than 1 because the sum of the logical probabilities of "adult" and "non-adult" alone is equal to 1.

According to the above definition, we have:

$$T(y_j) \equiv T(\theta_j) \equiv P(X \in \theta_j) = \sum_i P(x_i)T(\theta_j|x_i) \quad (5)$$

This is the probability of a fuzzy event defined by Zadeh [23].

We can put $T(\theta_j|x)$ and $P(x)$ into the Bayesian formula to obtain the semantic probability prediction formula:

$$P(x|\theta_j)=\frac{T(\theta_j|x)P(x)}{T(\theta_j)},\ T(\theta_j)=\sum_i T(\theta_j|x_i)P(x_i) \quad (6)$$

To $P(x|\theta_j)$ is the likelihood function $P(x|y_j, \theta)$ in the popular method. We use $P(x|\theta_j)$ here because the j-th parameter is bound to y_j. We call the above formula the semantic Bayesian formula:

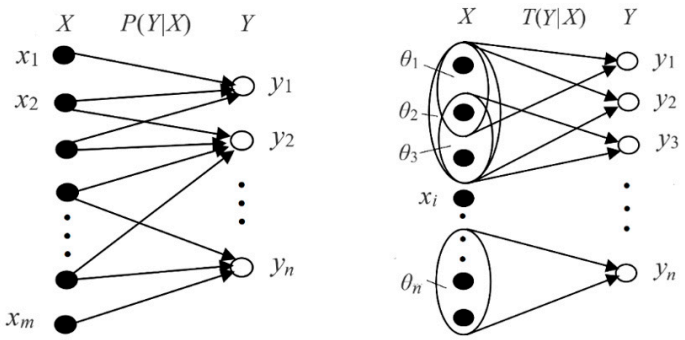
Because the maximum value of $T(y_j|x)$ is 1, from $P(x)$ and $P(x|\theta_j)$, we derive a new formula:

$$T(\theta_j|x)=\frac{P(x|\theta_j)}{P(x)}\Big/\max_x\left[\frac{P(x|\theta)}{P(x)}\right] \quad (7)$$

2.3. Semantic Channel and Semantic Communication Model

 doi:10.20944/preprints202502.0799.v1

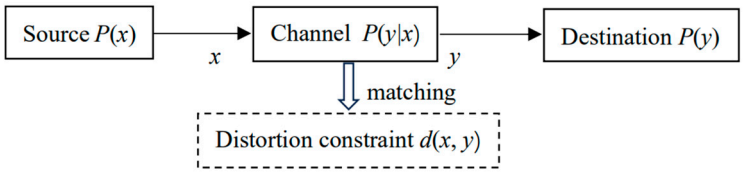
Shannon calls $P(X)$, $P(Y|X)$, and $P(Y)$ the source, the channel, and the destination. Just as a set of transition probability functions $P(y_j|x)$(j=1, 2, …) constitutes a Shannon channel, a set of truth value functions $T(\theta_j|x)$ (j=1,2,…) constitutes a semantic channel. The comparison of the two channels is shown in Figure 3. For convenience, we also call $P(x)$, $P(y|x)$, and $P(y)$ the source, the channel, and the destination, and we call $T(y|x)$ the semantic channel.



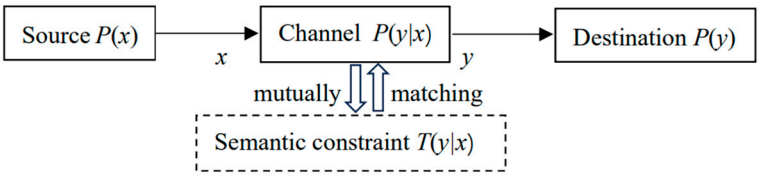
(a) (b)

Figure 3. The Shannon channel (a) and the semantic channel (b).

The semantic channel reflects the semantics or extensions of labels, while the Shannon channel indicates the usage of labels. The comparison between the Shannon and the semantic communication models is shown in Figure 4. The distortion constraint is usually not drawn, but it actually exists.



(a)



(b)

Figure 4. Communication models. (a) The Shannon communication model where the channel needs to match the distortion function. (b) The semantic communication models where two channels need to match mutually.

The semantic channel contains information about the distortion function, and the semantic information represents the communication quality, so there is no need to define a distortion function anymore. Optimizing the model parameters is to make the semantic channel match the Shannon channel, that is, $T(\theta_j|x) \propto P(y_j|x)$ or $P(x|\theta_j) = P(x|y_j)$ (j=1,2,…), so that the semantic MI reaches its maximum value and is equal to the Shannon information. Conversely, when the Shannon channel matches the semantic channel, the information difference reaches the minimum, or the information efficiency reaches the maximum.

2.4. Generalizing Shannon Information Measure to Semantic Information G Measure

Shannon MI can be expressed as.

 doi:10.20944/preprints202502.0799.v1

$$I(X;Y)=\sum_j\sum_i P(x)P(x\mid y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}=H(X)-H(X\mid Y), \quad (8)$$

where $H(X)$ is the entropy of X, reflecting the minimum average code length. $H(X|Y)$ is the posterior entropy of X, reflecting the minimum average code length after predicting x based on Y. Therefore, Shannon MI means the average code length saved due to the prediction.

We replace $P(x_i|y_j)$ on the right side of the log with the likelihood function $P(x_i|\theta j)$. Then we get the semantic MI:

$$\begin{aligned}I(X;Y_\theta)&=\sum_j\sum_i P(x_i)P(x_i\mid y_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}\\&=\sum_j\sum_i P(x_i)P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}\\&=H(X)-H(X\mid Y_\theta)=H(Y_\theta)-H(Y_\theta\mid X)=H(Y_\theta)-\bar{d},\end{aligned} \quad (9)$$

where $H(X|Y_\theta)$ is the semantic posterior entropy of x:

$$H(X\mid Y_\theta)=-\sum_j\sum_i P(x_i,y_j)\log P(x_i\mid\theta_j). \quad (10)$$

$H(X|Y_\theta)$ is the free energy F in the Variational Bayes method (VB) [44,45] and the Minimum Free Energy (MFE) principle [46]. The smaller it is, the greater the amount of semantic information. $H(Y_\theta|X)$ is called the fuzzy entropy, equal to the average distortion $\bar{d}$. Because according to Equation (2), there is:

$$H(Y_\theta\mid X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\mid x_i)=\bar{d}. \quad (11)$$

$H(Y_\theta)$ is the semantic entropy:

$$H(Y_\theta)=-\sum_i P(y_j)\log T(\theta_j). \quad (12)$$

Note that $P(x|y_j)$ on the left side of the log is used for averaging and represents the sample distribution. It can be a relative frequency and may not be smooth or continuous. $P(x|\theta_j)$ and $P(x|y_j)$ may differ, reflecting that obtaining information needs factual testing. 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 is the regularization term. Fuzzy entropy is a more general average distortion than the average square error.

Semantic entropy has a clear coding meaning. Assume that the sets θ_1, θ_2, … are crispy sets; the distortion function is:

$$d(y_j\mid x_i)=\begin{cases}\infty, & x_i\notin\theta_j,\\ 0, & x_i\in\theta_j.\end{cases} \quad (13)$$

We regard $P(Y)$ as the source and $P(X)$ as the destination, then the parameter solution of the information rate-distortion function is [31]:

$$\begin{aligned}&R(D)=sD(s)-\sum_j P(y_j)\log\lambda_j,\\&\lambda_j=\sum_i P(x_i)\exp[-d(x_i,y_j)]=\sum_i P(x_i)T(\theta_j\mid x_i)=T(\theta_j).\end{aligned} \quad (14)$$

It can be seen that the minimum Shannon MI is equal to the semantic entropy, that is, $R(D=0)=H(Y_\theta)$.

The following formula indicates the relationship between Shannon MI and semantic MI and the encoding significance of semantic MI:

 doi:10.20944/preprints202502.0799.v1

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i\mid y_j)}{P(x_i)}=$$
$$\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}+\sum_j\sum_i P(x_i,y_j)\log\frac{P(x_i\mid y_j)}{P(x_i\mid\theta_j)}$$
$$=I(X;Y_\theta)+\sum_j P(y_j)KL(P(x\mid y_j)\,\|\,P(x\mid\theta_j)), \tag{15}$$

where KL(…) is the Kullbak-Leibler (KL) divergence with a likelihood function, which Akaike [43] first used to prove that the minimum KL divergence criterion is equivalent to the maximum likelihood criterion. The last term in the above formula is always greater than 0, reflecting the average code length of residual coding. Therefore, the semantic MI is less than or equal to the Shannon MI; it reflects the average code length saved due to semantic prediction.

From the above formula, the semantic MI reaches its maximum value when the semantic channel matches the Shannon channel. According to Equation (15), letting $P(x\mid\theta_j)= P(x\mid y_j)$, we can obtain the optimized truth function from the sample distribution:

$$T^*(\theta_j\mid x)=\frac{P(x\mid y_j)}{P(x)}\Big/\max_x\left(\frac{P(x\mid y)}{P(x)}\right)=\frac{P(y_j\mid x)}{\max_x(P(y_j\mid x))}. \tag{16}$$

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

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{17}$$

The KL divergence cannot usually be interpreted as information because the smaller it is, the better. But $I(X;\theta_j)$ above can be said to be information because the larger it is, the better.

Solving $T^*(\theta_j\mid x)$ with equation (16) requires that the sample distributions $P(x)$ and $P(x\mid y_j)$ are continuous and smooth. Otherwise, by using Equation (17), we can obtain:

$$T^*(\theta_j\mid x)=\arg\max_{\theta_j}\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{18}$$

The above method for solving $T^*(\theta_j\mid x)$ is called Logical Bayesian Inference (LBI) [8] and can be called the random point falling shadow method. This method inherits Wang's idea of random set falling shadow [47,48].

Suppose the truth function in (10) becomes a similarity function. In that case, the semantic MI becomes the estimated MI [25], which has been used by deep learning researchers for Mutual Information Neural Estimation (MINE) [34] and Information Noise Contrast Estimation (InfoNCE) [35].

In the semantic KL information formula, when $X=x_i$, $I(X;\theta_j)$ becomes the semantic information between a single instance x_i and a single label y_j:

$$I(x_i;\theta_j)=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}. \tag{19}$$

The above formula reflects Popper's idea about factual testing. Figure 5 illustrates the above formula. It shows that the smaller the logical probability, the greater the absolute value of the information; the greater the deviation, the smaller the information; wrong assumptions convey negative information.

 doi:10.20944/preprints202502.0799.v1

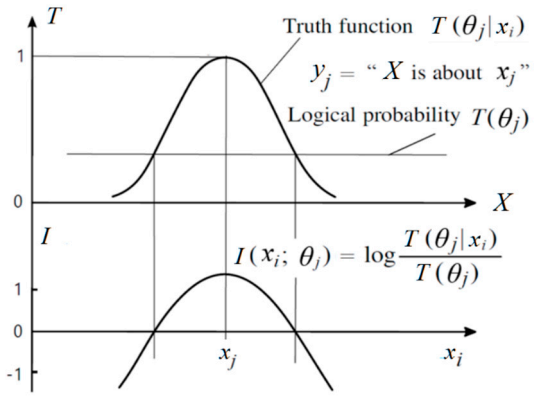


Figure 5. Semantic information y_j conveys about x_i decreases with the deviation or distortion increasing.

Bring Equation (2) into (19), we have

$$I(x_i;\theta_j)=\log[1/T(\theta_j)]-d(y_j|x), \quad (20)$$

which means $I(X;\theta_j)$ equals Carnap and Bar-Hillel's semantic information minus distortion. If $T(\theta_j|x)$ is always 1, the two amounts of information become equal.

2.5. From the Information Rate-distortion Function to the Information Rate-fidelity Function

Shannon 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, which can guide us in using Shannon information economically.

Now, we replace $d(y_j|x_i)$ with $I(x_i;\theta_j)$, 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 reflect information efficiency.

The author uses the word "fidelity" because Shannon originally proposed the information rate-fidelity criterion [12], and later used minimum distortion to express maximum fidelity. The author has previously referred to $R(G)$ as "the information rate of keeping precision" [16] or "information rate-verisimilitude" [25].

The $R(G)$ function is defined as

$$R(G)=\min_{P(Y|X):I(X;\theta)\geq G} I(X;Y) \quad (21)$$

We use the Lagrange multiplier method to find the minimum MI. The Lagrangian function is:

$$L(P(y|x),P(y))=I(X;Y)-sI(X;Y_\theta)-\mu_j\sum_i P(x_i|y_j)-\alpha\sum_j P(y_j) \quad (22)$$

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

$$P(y_j|x_i)=P(y_j)m_{ij}^s\Big/\lambda_i,\ \lambda_i=\sum_j P(y_j)m_{ij}^s,\ i=1,2,...;\ j=1,2,... \quad (23)$$

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

$$P^{+1}(y_j)=\sum_i P(x_i)P(y_j|x_i), \quad (24)$$

where $P^{+1}(y_j)$ 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 [33] (P. 326)). We call this method the Minimum Information Difference (MID) iteration.

The parameter solution of the $R(G)$ function (as illustrated in Figure 6) is:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

$$G(s)=\sum_i\sum_j P(x_i)P(y_j\mid x_i)I_{ij}=\sum_i\sum_j I_{ij}P(x_i)P(y_j)m_{ij}^{\ s}/Z_i,$$
$$R(s)=sG(s)-\sum_i P(x_i)\log Z_i,\quad Z_i=\sum_k P(y_k)m_{ij}^{\ s}. \tag{25}$$

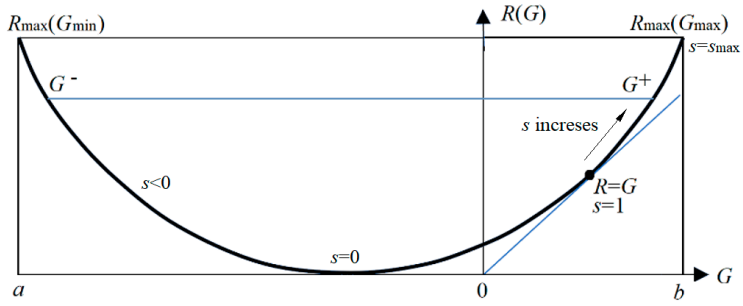


Figure 6. The information rate-fidelity function $R(G)$ for binary communication. Any $R(G)$ function is a bowl-like function. There is a point at which $R(G) = G$ ($s = 1$). For given R, two anti-functions exist: $G^-(R)$ and $G^+(R)$.

Any $R(G)$ function is bowl-shaped (possibly not symmetrical) [6], 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.

We can apply the $R(G)$ function to image compression based on visual discrimination [5,6], maximum MI classification of unseen instances, the convergence proof of mixture models [7], and semantic compression [28].

It is worth noting that, given a semantic channel $T(y\mid x)$, matching the Shannon channel with the semantic channel, i.e., letting $P(y_j\mid x) \propto T(y_j\mid x)$ or $P(x\mid y_j) = P(x\mid\theta_j)$, does not maximize the semantic MI, but minimizes the information difference between R and G or the information efficiency G/R. Then, we can increase G and R simultaneously by increasing s. When s-->∞ in Equation (23), $P(y_j\mid x)$ ($j = 1, 2, \ldots, n$) only takes the value 0 or 1, becoming a classification function.

We can also replace the average distortion with fuzzy entropy $H(Y_\theta\mid X)$ (using semantic distortion constraints) to obtain the information rate truth function $R(\Theta)$ [28]. In situations where information rather than truth is more important, $R(G)$ is more appropriate than $R(D)$ and $R(\Theta)$. $P(y)$ and $P(y\mid x)$ obtained for $R(\Theta)$ are different from those obtained for $R(G)$ because the optimization criteria are different. Under the minimum semantic distortion criterion, $P(y\mid x)$ becomes:

$$P(y_j\mid x_i)=P(y_j)[T(\theta_{xi}\mid y_j)]^s\Big/\sum_j P(y_j)[T(\theta_{xi}\mid y_j)]^s,\ i=1,2,\ldots;\ j=1,2,\ldots \tag{26}$$

where $T(\theta_{xi}\mid y)$ is a constraint function so that the distortion function $d(x_i\mid y)$ =–log $T(\theta_{xi}\mid y)$. $R(\Theta)$ becomes $R(D)$. If $T(\theta_j)$ is small, the $P(y_j)$ required for $R(G)$ will be larger than the $P(y_j)$ required for $R(D)$ or $R(\Theta)$.

2.6. Semantic Channel Capacity

Shannon calls the maximum MI obtained by changing the source $P(x)$ for the given Shannon channel $P(y\mid x)$ the channel capacity. Because the semantic channel is also inseparable from the Shannon channel, we must provide both the semantic and Shannon channels to calculate the semantic MI. Therefore, after the semantic channel is given, there are two cases: 1) the Shannon channel is fixed; 2) we must first optimize the Shannon channel according to a specific criterion.

When the Shannon channel is fixed, the semantic MI is less than the Shannon MI, so the semantic channel capacity is less than or equal to the Shannon channel capacity. The difference between the two is shown in Equation (15).

If the Shannon channel is variable, we can use the MID iteration to find the Shannon channel for $R=G$ after each change of the source $P(x)$, and then use s—>∞ to find the Shannon channel $P(y\mid x)$ that

 doi:10.20944/preprints202502.0799.v1

makes R and G reach their maxima **simultaneously**. At this time, $P(y|x)\in\{0,1\}$ becomes the classification function. Then, we calculate the semantic MI. For different $P(x)$, the maximum semantic MI is the semantic channel capacity. That is

$$C_{T(y|x)}=\underset{P(x);P(y|x)\text{ for }R(G,s\to\infty)}{\arg\max} I(X;Y_\theta)=\underset{P(x)}{\arg\max}\, G_{\max}, \quad (27)$$

where $G_{\max}$ is G^+ when $s\to\infty$ (see Figure 6). Hereafter, the semantic channel capacity only refers to $C_{T(y|x)}$ in the above formula.

Practically, to find $C_{T(Y|X)}$, we can look for $x(1)$, $x(2)$, …, x(j) ∈ U, which are instances under the highest points of $T(y_1|x)$, $T(y_2|x)$, …, $T(y_n|x)$ respectively. Let $P(x(j))=1/n$, $j=1, 2, \ldots, n$, and the probability of any other x equals 0. Then we can choose the Shannon channel: $P(y_j|x(j))=1$, $j=1, 2, \ldots, n$. At this time, $I(X; Y))=H(Y)=\log n$, which is the upper limit of $C_{T(Y|X)}$. If there is x_i among the n $x(j)$s, which makes more than one truth function true, then either $T(y_j)>P(y_j)$ or the fuzzy entropy is not 0. $C_{T(Y|X)}$ will be slightly less than $\log n$ in this case.

According to the above analysis, the encoding method to increase the capacity of the semantic channel is:

1. .Try to choose x that only makes one label's true value 1 (avoid ambiguity and reduce the logical probability of y);
2. Encoding should make $P(y_j|x_j)=1$ as much as possible (to ensure that Y is used correctly).
3. Choose $P(x)$ so that each Y's probability and logical probability are as equal as possible (close to $1/n$, thereby maximizing the semantic entropy).

3. Electronic Semantic Communication Optimization

3.1. Electronic Semantic Communication Model

The previous discussion of semantic communication did not consider conveying semantic information by electronic communication. Assuming that the time and space distance between the sender and the receiver is very far, we must transmit information through cables or disks. At this time, we need to add an electronic channel based on the previous communication model, as shown in Figure 7:

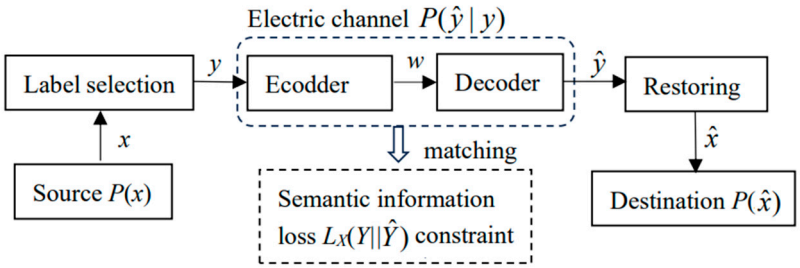


Figure 7. The electronic semantic communication model. The distortion constraint is replaced with the semantic information loss constraint.

Electronic semantic communication is still electronic communication, in essence. The difference is that we need to use semantic information loss instead of distortion as the optimization criterion. The choice of Y includes the optimization of the semantic channel $P(y|x)$ and the Shannon channel $T(y|x)$ between x and y.

3.2. Optimization of Electronic Semantic Communication with Semantic Information Loss as Distortion

Consider electronic semantic communication. If there is no distortion, that is, $\hat{y}_j=y_j$, the semantic information about x transmitted by both is the same, and there is no semantic information loss. If $\hat{y}_j\neq y_j$, there is semantic information loss. Farsad et al. call it the semantic error and propose the

 doi:10.20944/preprints202502.0799.v1

corresponding formula [49,50]. Papineni et al. also proposed a similar formula for translation [51]. For more discussion, see [52].

According to the G theory, the semantic information loss caused by using $\hat{y}_j$ instead of y_j is:

$$L_X(y_j \| \hat{y}_j) = I(X;Y_\theta) - I(X;\hat{Y}_\theta) = \sum_i P(x_i \mid y_j) \log \frac{P(x_i|\theta_j)}{P(x_i \mid \hat{\theta}_j)} . \quad (28)$$

$Lx(y_j||\hat{y}_j)$ is a generalized KL divergence because there are three functions. It indicates the average code length of residual coding.

Since the loss is generally asymmetric, there may be $Lx(y_j||\hat{y}_j) \neq Lx(y_j||\hat{y}_j)$. For example, when "motor vehicle" and "car" are substituted for each other, the information loss is asymmetric. The reason is that there is a logical implication relationship between the two. Using "motor vehicle" to replace "car", although it reduces information, it is not wrong; while using "car" to replace "motor vehicle" may be wrong, because the actual may be a truck or a motorcycle. When an error occurs, the semantic information loss is enormous. An advantage of using the truth function to generate the distortion function is that it reflects concepts' implications or similarity relationships.

Assuming that y_j *is* the correct label used, it comes from sample learning, so $P(x|\theta_j)=P(x|y_j)$, and $Lx(y_j||\hat{y}_j)=KL(P(x|\theta_j)||P(x|\hat{\theta}_j))$. The average semantic information loss is:

$$L_X(Y \| \hat{Y}) = \sum_j \sum_k P(y_j) P(\hat{y}_k \mid y_j) KL(P(x \mid \theta_j) \| P(x \mid \hat{\theta}_j)) . \quad (29)$$

Consider using $P(y)$ as the source and $P(\hat{Y})$ as the destination to encode y. Let $d(\hat{y}_k|y_j)=Lx(y_j||\hat{y}_k)$; we can obtain the information rate-distortion function $R(D)$ for replacing Y with $\hat{Y}$. We can code Y for data compression according to the parameter solution of the $R(D)$ function.

In the electronic communication part (from Y to $\hat{Y}$), other problems can be resolved by classical electronic communication methods, except for using semantic information loss as distortion.

If finding $I(x;\hat{\theta}_j)$ is not too difficult, we can also use $I(x;\hat{\theta}_j)$ as a negative distortion function. Minimizing $I(X;\hat{Y})$ for given when $G= I(X;\hat{Y}_\theta)$, we can get the $R(G)$ function between x and $\hat{Y}$ and compress the data accordingly.

3.3. Experimental Results: Compress Image Data According to Visual Discrimination

The simplest visual discrimination is the discrimination of human eyes to different colors or gray levels. The nest is the spatial discrimination of points. Suppose the movement of a point on the screen is not detected. In that case, the fuzzy movement range can represent the spatial position discrimination, which can be represented by a truth function (such as the Gaussian function). What is more complicated is to distinguish whether two figures are the same person. Advanced image compression needs to extract image features like Autoencoder and use features to represent images. The following methods need to be combined with the feature extraction method in deep learning to get better applications.

The simplest gray-level discrimination is taken as an example to illustrate digital image compression.

1) Measuring Color Information

A color can be represented by a vector (B, G, R). For convenience, we assume that the color is one-dimensional (or we only consider the gray level), expressed in x, and the color sense Y is the estimation of x, similar to the GPS indicator. The universes of x and Y are the same, and y_j="x is about x_j". If the color space is uniform, the distortion function can be defined by distance, that is, $d(y_j|x) = \exp[-(x-x_j)^2/(2\sigma^2)]$. Then there is the average information of color perception, $I(X;Y_\theta)=H(Y_\theta)-\overline{d}$.

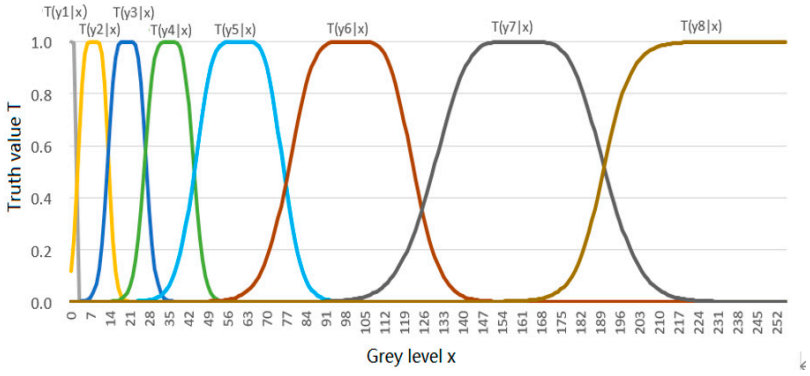
Given the source $P(x)$ and the discrimination function $T(y|x)$, we can solve $P(y|x)$ and $P(y)$ using the SVB method. The Shannon channel is matched with the semantic channel to maximize the information efficiency.

2) Gray Level Compression

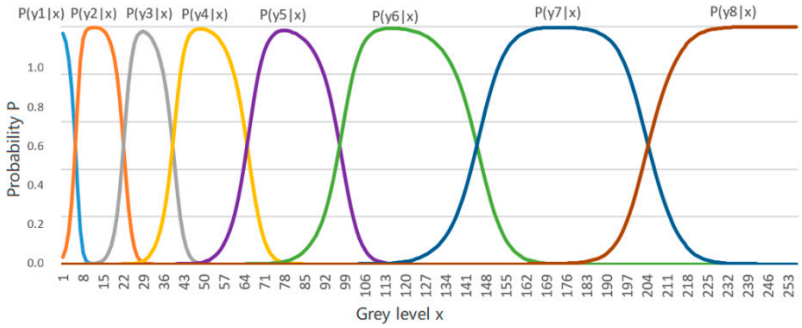
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

We use an example to illustrate color data compression. Assuming that the original gray level is 256 (8-bit pixels) and is now compressed into 8 (3-bit pixels), we can define eight constraint functions, as shown in Figure 8a.

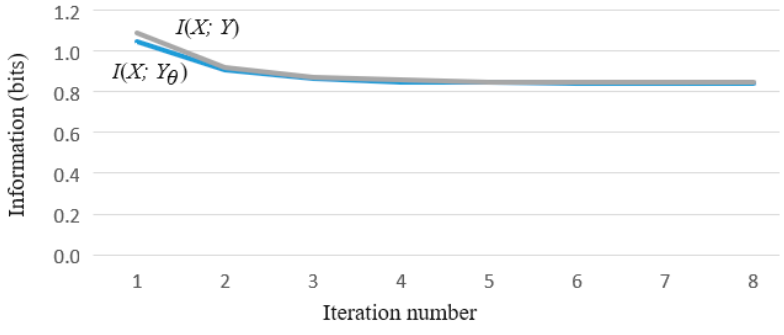
Considering that human visual discrimination varies with the gray level (the higher the gray level, the lower the discrimination), we use the eight truth functions shown in Figure 8a, representing eight fuzzy ranges. Appendix C in Reference [28] shows how these curves are generated. The task is to use the Maximum Information Efficiency (MIE) criterion to find the Shannon channel $P(y|x)$ that makes R close to G (s=1).



(a)



(b)



(c)

Figure 8. The iteration results of Example 2. (a) 8 truth value functions or the semantic channels $T(y|x)$ (see Appendix C in [28] for the Data generation method). (b) Convergent Shannon channel $P(y|x)$. (c) The variation of $I(X; Y_\theta)$ and $I(X; Y)$ during iteration.

The convergent $P(y|x)$ is shown in Figure 8b. Figure 8c shows that Shannon MI and semantic MI gradually approach in the iteration process. Comparing figures 8a and 8b, we find it easy to control

 doi:10.20944/preprints202502.0799.v1

$P(y|x)$ by $T(y|x)$. However, defining the distortion function without using the truth function is difficult. It is also difficult to predict the convergent $P(y|x)$ by $d(y|x)$.

If we use s to strengthen the constraint, we get the parametric solution of the $R(G)$ function. As $s\to\infty$, $P(y_j|x)$ (j=1,2,…) display as rectangles and becomes classification functions.

3) Influence of Discrimination and Quantization Level on the $R(G)$ Function

Consider the semantic information of gray pixels. The discrimination function determines the semantic channel $T(y|x)$, and the source entropy $H(X)$ increases with the quantization level $b=2^n$ (n is the number of quantization bits). Figure 9 shows that when the quantization level is enough, the R and G variation range increases with the discrimination increasing (i.e., with σ decreasing). The discrimination determines the semantic channel capacity.

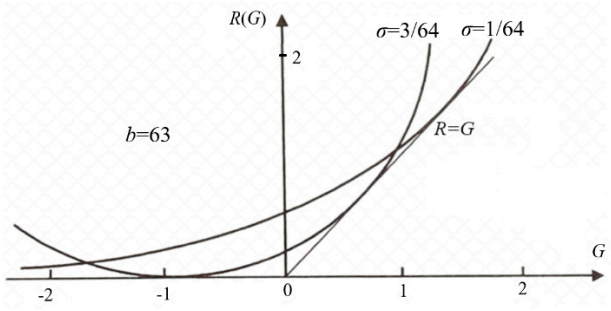


Figure 9. Variation of $R(G)$ function with discrimination (σ=1/64 or σ=3/64) for a given quantization level b=63.

For more discussion on visual information, see Section 6 in [6].

4. Goal-Oriented Information, Information Value, Physical Entropy and Free Energy

4.1. Three Kinds of Information Related to Value

We call the increment of utility the value. Information involves utility and value in three aspects:

1) Information about utility. For example, the information about university admission or the bumper harvest of grain is about utility.

The measurement of this information is the same as the previous semantic information measurement. Before providing information, we have the prior probability distribution $P(x)$ of grain production. The information is provided in the form of range, such as "about 2000 kg per acre", which can be expressed by a truth function. The previous semantic information formula is also applicable.

2) Goal-oriented information. It is also purposeful information or constraint control feedback information.

For example, a passenger and a driver watch GPS maps in a taxi. Assume that the probability distribution of the taxi position without looking at the positioning map (or without some control) is $P(x)$, and the destination is a fuzzy range, which a truth function can represent. The actual position is the probability distribution $P(x|a_j)$ (conditioned on action a_j). The positioning map provides information. For the passenger, this is purposeful information (about how the control result comforts the purpose); for the driver, this is the control feedback information. We call both goal-oriented information. This information involves constraint control and reinforcement learning. The following section discusses the measurement and optimization of this information.

3) **Information that brings value**. For example, Tom made money by buying stocks based on John's prediction of stock prices. The information provided by John brings Tom increased utility, so John's information is valuable to Tom.

The value of information is relative. For example, weather forecast information is different for workers and farmers, and forecast information about stock markets is worth 0 to people who do not buy stocks. The value of information is often difficult to judge. For example, defining value losses due to missed reporting and false reporting is difficult regarding medical cancer tests. In most cases,

 doi:10.20944/preprints202502.0799.v1

missed reporting of low-probability events often causes more loss than false reporting, such as for medical tests and earthquake forecasts. In these cases, the semantic information criterion can be used to reduce missed reporting of low-probability events.

For investment portfolios, quantitative analysis of information value is possible. Section 4.3 focuses on the information value of portfolios.

4.2. Goal-Oriented Information

4.2.1. Similarities and Differences Between Goal-Oriented Information and Prediction Information

Previously, we used the G measure to measure prediction information, requiring the prediction $P(x|\theta_j)$ to conform to the fact $P(x|y_j)$. Goal-oriented information is the opposite, requiring the fact to conform to the purpose.

An imperative sentence can be regarded as a control instruction. We need to know whether the control result conforms to the control purpose. The more consistent the result is, the more information there is.

A truth function or a membership function can represent a control target. For example, there are the following targets:

1. "Workers' wages should preferably exceed 5000 dollars";
2. "The age of death of the population had better exceed 80 years old";
3. "The cruising distances of electric vehicles should preferably exceed 500 kilometers";
4. "The error of train arrival time had better be less than one minute".

The semantic KL information formula can measure purposeful information:

$$I(X;a_j/\theta_j)=\sum_i P(x_i|a_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}. \quad (30)$$

In the formula, θ_j is a fuzzy set, indicating that the control target is a fuzzy range. y_j here becomes a_j, indicating the action corresponding to the j-th control task y_j. If the control result is a specific x_i, the above formula becomes the semantic information $I(x_i; a_j|\theta_j)$.

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

$$I(X;A/\theta)=\sum_j P(a_j)\sum_i P(x_i|a_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}, \quad (31)$$

where A is a random variable taking a value a or a_j. Using SVB, the control ratio $P(a)$ can be optimized to minimize the control complexity (i.e., Shannon MI) when the purposive information is the same.

4.2.2. Optimization of Goal-Oriented Information

Goal-oriented information can be regarded as the cumulative reward in constraint control. However, the goal here is a fuzzy range, which is expressed by a plan, command, or imperative sentence. The optimization task is similar to the active inference task using the MFE principle [46].

The semantic information formulas of imperative and descriptive (or predictive) sentences are the same, but the optimization methods differ (see Figure 10). For descriptive sentences, the fact is unchanged, and we hope that the predicted range conforms to the fact, that is, fix $P(y_j|x)$ so that $T(\theta_j|x)\propto P(y_j|x)$, or fix $P(x|y_j)$ so that $P(x|\theta_j)=P(x|y_j)$. For imperative sentences, we hope that the fact conforms to the purpose, that is, fix $T(\theta_j|x)$ or $P(x|\theta_j)$, and minimize the information difference or maximize the information efficiency G/R by changing $P(y_j|x)$ or $P(x|y_j)$, or balance between the purposiveness and the efficiency.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

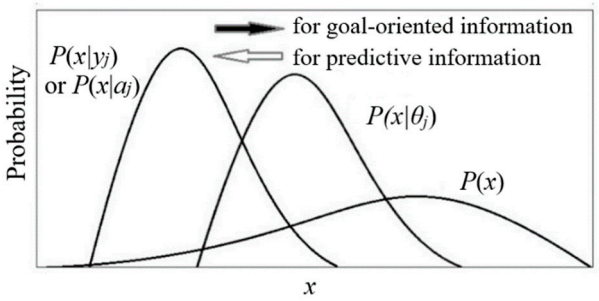


Figure 10. The optimization methods of the two types of semantic information are different. For predictive information, we hope that $P(x|\theta_j)$ is close to $P(x|y_j)$ (see white arrow); while for Goal-oriented information, we hope that $P(x|a_j)$ is close to $P(x|\theta_j)$ (see black arrow).

For multi-target tasks, the objective function to be minimized is:

$$f = I(X; A) - sI(X;A/\theta). \quad (32)$$

When the actual distribution $P(x|a_j)$ is close to the constrained distribution $P(x|\theta_j)$, the information efficiency (not information) reaches its maximum value of 1. To further increase the two types of information, we can use the MID iteration formula to obtain:

$$P(a_j \mid x) = P(a_j) m_{ij}^{s} \Big/ \lambda_i, \quad (33)$$

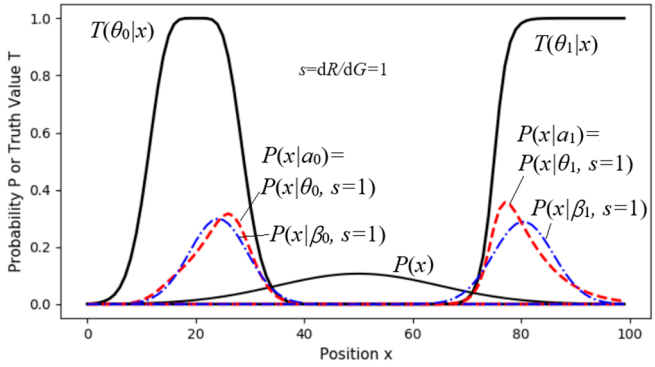
$$P(x_i \mid a_j) = P(a_j \mid x_i) P(x_i) / P(a_j) = P(x_i) m_{ij}^{s} \Big/ \sum_k P(x_k) m_{kj}^{s}. \quad (34)$$

Compared with VB [47,48], the above method is simpler and can change the constraint strength by s.

Because the optimized $P(x|a_j)$ is a function of θ_j and s, we write $P^*(x|a_j)=P(x|\theta_j, s)$. It is worth noting that many distributions $P(x|a_j)$ satisfy the constraint and maximize $I(X; a_j/\theta_j)$, but only $P^*(x|a_j)$ minimizes $I(X; a_j)$.

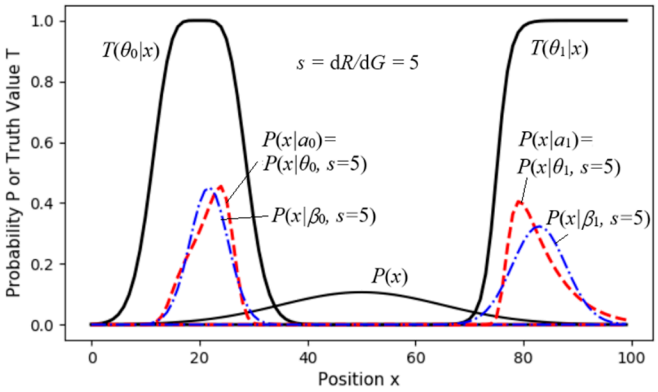
4.2.3. Experimental Results: Trade-Off Between Maximizing Purposiveness and MIE

Figure 11 **s**hows a two-objective control task, with objectives represented by the truth functions $T(\theta_0|x)$ and $T(\theta_1|x)$. We can imagine these as two pastures with fuzzy boundaries where we need to herd sheep. Without control, the density distribution of the sheep is $P(x)$. We need to solve an appropriate distribution $P(a)$.



(a)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1



(b)

Figure 11. A two-objective control task. Dashed lines show $P(x|a_j)=P(x|\theta_j, s)$ (j=0, 1), and dash-dotted lines representing $P(x|\beta_j, s)$ (j=0,1). (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 action a_j.

For different s, we set the initial proportions: $P(a_0)=P(a_1)=0.5$. Then, we used (33) and (34) for the MID iteration to obtain proper $P(a_j|x)$ (j=0,1). Then, we got $P(x|a_j)=P(x|\theta_j,s)$ by using (33). Finally, we obtained $G(s)$, $R(s)$, and $R(G)$ by using (25).

The dashed line for $R_1(G)$ in Figure 12 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.

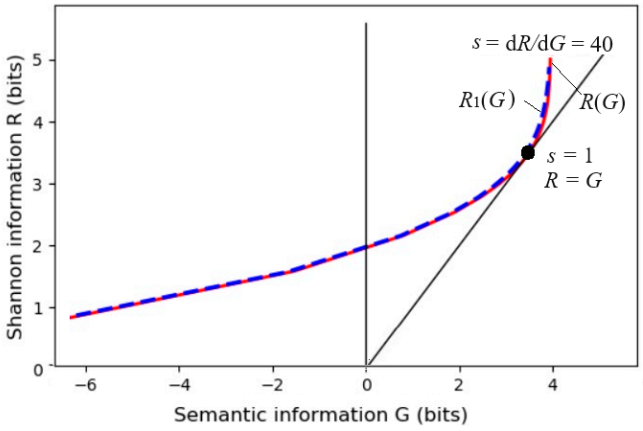


Figure 12. The $R(G)$ for constraint control. G slightly increases when s increases from 5 to 40, meaning s=5 is good enough.

4.3. Investment Portfolios and Information Values

4.3.1. Capital Growth Entropy

Markowitz's portfolio theory [53] uses a linear combination of expected income and standard deviation as the optimization criterion. In contrast, the compound interest theory of portfolios uses compound interest, i.e., geometric mean income, as the optimization criterion.

The compound interest theory began with Kelley [54], followed by Latanne and Tuttle [55], Arrow [56], Cover [57], and the author of this article. The famous American information theory textbook "Elements of Information Theory" [58] co-authored by Cover and Thomas, introduced Cover's research. Arrow, Cover, and the author of this article also discussed the value of information. The author published a monograph, "Entropy Theory of Portfolios and Information Value" [59] 1997, and obtained many different conclusions.

The following is a brief introduction to the Capital Growth entropy proposed by the author.

 doi:10.20944/preprints202502.0799.v1

Assuming that the principal is A, the profit is B, and the sum of principal and interest is C. The investment income is $r=B/A$, and the rate of return on investment (i.e., output ratio: output/input) is $R=C/A=1+r$.

N security prices form an N-dimensional vector, and the price of the k-th security has n_k possible prices, $k=1, 2, ..., N$. There are $W=n_1\times n_2\times...\times n_N$ possible price vectors. The i-th price vector is $x_i = (x_{i1}, x_{i2}, ... x_{iN})$, $i=1, 2, ..., W$; the current price vector is $x_0=(x_{01}, x_{02}, ..., x_{0N})$. Assuming that one year later, the price vector x_i occurs, then the rate of return of the k-th security is $R_{ik}=x_{ik}/x_{0k}$, and the total rate of return is:

$$R_i = \sum_{k=0}^{N} q_k R_{ik} \quad (35)$$

where q_k is the investment proportion in the k-th security, q_0 is the proportion of cash (or risk-free assets) held by the investor; $R_{0k}=R_0=(1+r_0)$, r_0 is the risk-free interest rate.

Suppose we conduct m investment experiments to get the price vectors, and the number of times x_i or r_i occurs is m_i. The average multiple of the capital growth after each investment period or the geometric mean output ratio is

$$R_g = \prod_{i=1}^{W} R_i^{m_i/m}. \quad (36)$$

When $m\to\infty$, $m_i/m=P(x_i)$, we have the capital growth entropy

$$H_g = \log R_g = \sum_{i=1}^{W} P(x_i)\log R_i = \sum_{i=1}^{W} P(x_i)\log \sum_{k=0}^{N} q_k R_{ik}. \quad (37)$$

If the log is base 2, H_g represents the doubling rate.

If the investment turns into betting on horse racing, where only one horse (the k-th horse) wins each time. The winner's return rate is R_k, and the others lose their wagers. Then, the above formula becomes

$$H_g = \log R_g = \sum_{k} P(x_k)\log[q_0 + q_k R_k - (1-q_0-q_k)], \quad (38)$$

where q_0 is the proportion of funds not betted, and $1-q_0-q_k$ is the proportion of funds paid.

4.3.2. Generalization of Kelley's Formula

Kelley, a colleague of Shannon, found that the method used by Shannon's information theory can be used to optimize betting, so he proposed the Kelley formula [54].

Assume that In a gambling game, if you lose, you will lose $r_1=1$ time; if you win, you will earn $r_2>0$ times. The probability of winning is P, then the optimal ratio is:

$$q^*=P-(1-P)/r_2. \quad (39)$$

Using the capital growth entropy can lead to more general conclusions. Let $r_1<0$. The capital growth entropy is:

$$q^* = \arg\max_q H_g = P_1\log(1-qr_1) + P_2\log(1+qr_2). \quad (40)$$

Letting $dH_g/dq=0$, we derive:

$$q^*=E/(r_1r_2), \quad (41)$$

where E is the expected income. For example, for a coin toss bet, if one wins, he earns twice as much; if he loses, he loses 1 time; the probabilities of winning and losing are equal. Then E is 0.5, and the optimal investment ratio is $q^*=0.5/(1*2)=0.25$.

Assuming $r_0=0$ above, if we consider the opportunity cost or the risk-free income, then $r_0>0$. At this time, the optimal ratio is:

$$q^* = \arg\max_q H_g$$
$$= \arg\max_q \{P_1\log[r_0(1-q)+q-qr_1] + P_2\log[r_0(1-q)+q+qr_2]\}. \quad (42)$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

Letting dH_g/dq=0, we can get:

$$q^* = \frac{P_2 d_2 - P_1 d_1}{d_1 d_2} R_0 \quad , \tag{43}$$

where R_0=1+r_0, d_1=r_1+r_0, d_2=r_2–r_0.

The book [59] also discusses optimizing the investment ratio when short selling and leverage are allowed (see Section 3.3) and optimizing the investment ratio for multi-coin betting. The book derives the limit theorem of diversified investment: If the number of coins increases infinitely, the geometric mean income equals the arithmetic mean income.

4.3.3. Risk Measurement, Investment Channels, and Investment Channel Capacity

Markowitz uses expected income E and standard deviation σ to represent the income and risk of a portfolio. Similarly, we use R_g and R_r to represent the return and risk of a portfolio. R_r is defined in the following formula:

$$R_r^2 = R_a^2 - R_g^2 , \tag{44}$$

where R_a=1+E. Assuming that the geometric mean return of any portfolio is equivalent to the geometric mean return of a coin toss bet with an equal probability of gain or loss, then

$$H_g = \log R_g = 0.5\log(R_a - R_r) + 0.5\log(R_a + R_r) . \tag{45}$$

Let sinα=R_r/R_a ∈[0,1], which represents the bankruptcy risk better. When sinα is close to 1, the investment may go bankrupt (see Figure 3.9).

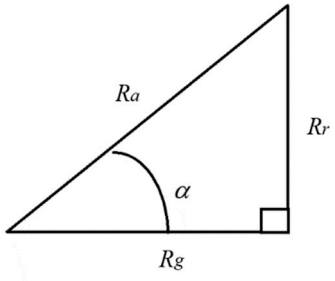


Figure 14. Relationship between relative risk sinα and R_r, R_a, and R_g.

We call the pair (**P, R**) the investment channel, where **P**=(P_1, P_2, ... P_M) is the future price vector, **R**=(R_{ik}) is the return matrix, and the set of all possible investment ratio vectors is $\mathbf{q}_C$. Then, the capacity of the investment channel (abbreviated as investment capacity) is defined as

$$\boldsymbol{H}_C^* = \max_{\boldsymbol{q} \in \boldsymbol{q}_C} \boldsymbol{H}(\boldsymbol{P}, \boldsymbol{R}, \boldsymbol{q}) = \boldsymbol{H}(\boldsymbol{P}, \boldsymbol{R}, \boldsymbol{q}^*) \quad , \tag{46}$$

where **q***= **q***(**R, P**) is the optimal investment ratio.

For example, for a typical coin toss bet (with equal probabilities of winning and losing, and r_0=0), q^*=E/($r_1 r_2$), the investment capacity is:

$$H_C^* = \frac{1}{2}\log\frac{1}{1 - E^2 / R_r^2} . \tag{47}$$

Since 1/(1–x)=1+x+x^2+…≈1+x, when E/R_r<<1, there is an approximate formula:

$$H_C^* \approx \frac{1}{2}\log(1 + \frac{E^2}{R_r^2}) . \tag{48}$$

In comparison with the Gaussian channel capacity formula for communication:

$$C = \frac{1}{2}\log(1 + \frac{P}{N}) \quad , \tag{49}$$

 doi:10.20944/preprints202502.0799.v1

we can see that the investment capacity formula is very similar to the Gaussian channel capacity formula. This similarity means that investment needs to reduce risk, just as communication needs to reduce noise.

4.3.4. Information Value Formula Based on Capital Growth Entropy

Weaver, who co-authored the book "A Mathematical Theory of Communication" [30] with Shannon, proposed three communication levels related to Shannon' 's information, semantic information, and information value.

According to the common usage of "information value", information value mentioned in the academic community does not refer to the value of information on markets but to the utility or utility increment generated by information. We define the information value as the increment of capital growth entropy [59].

Assume that the prior probability distribution of different returns is $P(x)$, and the return matrix is (R_{ik}), then the expected capital growth entropy is $H_g(X)$. The optimal investment ratio vector $\boldsymbol{q}^*$ is $\boldsymbol{q}^*=\boldsymbol{q}^*(P(x),(R_{ik}))$. When the probability distribution of the predicted return becomes $P(x|\theta_j)$, the capital growth entropy becomes

$$H_g^*(X|\theta_j)=\sum_i P(x_i|\theta_j)\log R_i(\boldsymbol{q}^*),\quad R_i(\boldsymbol{q}^*)=\sum_k q_k^* R_{ik} \quad (50)$$

The optimal investment ratio becomes $\boldsymbol{q}^{**}=\boldsymbol{q}^{**}(P(x|\theta_j),(R_{ik}))$. We define the increment of the capital growth entropy obtained after the semantic KL information $I(X;\theta_j)$ is provided as the information value (i.e., the average information value):

$$V(X;\theta_j)=\sum_i P(x_i|y_j)\log\frac{R_i(\boldsymbol{q}^{**})}{R_i(\boldsymbol{q}^*)} \quad (51)$$

It can be seen that $V(X;\theta_j)$ and $I(X;\theta_j)$ have similar structures. For the above formula, when x_i is determined to occur, the information value of y_j becomes

$$v_{ij}=v(x_i;\theta_j)=\log\frac{R_i(\boldsymbol{q}^{**})}{R_i(\boldsymbol{q}^*)} \quad (52)$$

Information value also needs to be verified by facts; wrong predictions may bring negative information value.

4.3.5. Comparison with Arrow's Information Value Formula

The utility function defined by Arrow is [56]:

$$U=\sum_i P_i U(q_i R_i)=\sum_i P_i\log(q_i R_i)=\sum_i P_i\log q_i+\sum_i P_i\log R_i, \quad (53)$$

where $U(q_iR_i)$ is the utility obtained by the investor when the i-th return occurs.

Under the restriction of $\sum_i q_i=1$, $q_i=P_i(i=1,2,\ldots)$ maximizes U so that

$$U^*=\sum P_i\log P_i+\sum_i P_i\log R_i \quad (54)$$

After receiving the information, the investor knows which income will occur and thus invests all his funds in it. Hence, there is

$$U^{**}=\sum_i P_i\log R_i. \quad (55)$$

The information value is defined as the difference in utility between investment with and without information and is equal to Shannon entropy, that is

$$V=U^{**}-U^*=-\sum_i P_i\log P_i=H(X). \quad (56)$$

The optimal investment ratio obtained from the above formula is inconsistent with the Kelley formula and the conclusion of the author of this article. For example, according to the Kelley formula,

 doi:10.20944/preprints202502.0799.v1

the optimal ratio is 25% for the coin toss bet above. The compound interest is 0.061%, and the investment capacity is 0.084<<1 bit.

According to Arrow's theory, how does one bet? Should one bet 50% on each of the profit and loss?

Arrow seems to confuse the k-th security with the i-th return. He uses $U(q_iR_i)=\log(q_iR_i)$, while the author uses

$$U(q_kR_k)=\log[q_0+q_kR_k-(1-q_0-q_k)]. \quad (57)$$

Arrow does not consider the non-bet proportion q_0, nor the paid proportion $1-q_0-q_k$. The utility calculated in this way is puzzling.

Cover and Thomas inherited Arrow's method and concluded that when there is information, the optimal investment doubling rate increment equals Shannon MI [57] (see Section 6.2). Their conclusion has the same problem.

4.4. Information, Entropy, and Free Energy in Thermodynamic Systems

To clarify the relationship between information and free energy in physics, we discuss information, entropy, and free energy in thermodynamic systems.

According to Stirling's formula, $\ln N! = N\ln N - N$ (when $N\to\infty$), there is a simple connection between Boltzmann entropy and Shannon entropy [60]:

$$S'=\mathrm{k}\ln W=\mathrm{k}\ln\frac{N!}{\prod_i N_i!}=-\mathrm{k}N\sum_i P(x_i\mid T)\ln P(x_i\mid T)=\mathrm{k}NH(X\mid T), \quad (58)$$

where S' is entropy, k is the Boltzmann constant, x_i is the i-th microscopic state, N is the number of molecules, and T is the absolute temperature, which equals a molecule's average translational kinetic energy. $P(x_i\mid T)$ represents the probability density of molecules in state x_i at temperature T. The Boltzmann distribution is:

$$P(x_i\mid T)=\exp(-\frac{e_i}{\mathrm{k}T})/Z',\quad Z'=\sum_i\exp(-\frac{e_i}{\mathrm{k}T}), \quad (59)$$

where Z is the partition function.

Considering the information between temperature and molecular energy, we use x_i as energy e_i. Let G_i denote the number of microscopic states with energy e_i and G denote the number of all states. Then $P(x_i)=G_i/G$ is the prior probability of x_i. So, Equation (58) becomes:

$$\begin{aligned}S&=\mathrm{k}\ln\frac{N!}{\prod_i N_i!/G_i^{N_i}}=-\mathrm{k}N\sum_i P(x_i\mid T)\ln\frac{P(x_i\mid T)}{G_i}\\&=-\mathrm{k}N\sum_i P(x_i\mid T)\ln\frac{P(x_i\mid T)}{P(x_i)}+\mathrm{k}N\ln G\\&=\mathrm{k}N[\ln G-KL(P(x\mid T)\,\|\,P(x)].\end{aligned} \quad (60)$$

Under the energy constraint, when the system reaches equilibrium, Equation (59) becomes:

$$P(x_i\mid T)=P(x_i)\exp(-\frac{e_i}{kT})/Z,\quad Z=\sum_i P(x_i)\exp(-\frac{e_i}{kT}), \quad (61)$$

Now, we can interpret $\exp[-e_i/(kT)]$ as the truth function $T(\theta_j\mid x)$, Z as the logical probability $T(\theta_j)$, and Equation (61) as the semantic Bayesian formula.

S and S' differ by a constant c (which does not change with temperature). There is $S'=S+\mathrm{c}$, $\mathrm{c}=\sum_i P(x_i)\ln G_i$. If c is ignored, there is $\ln G=H(X)+\mathrm{c}=H(X)$, and

$$S/(\mathrm{k}N)=H(X)-KL(P(x\mid T)\,\|\,P(x)). \quad (62)$$

Consider a local non-equilibrium system. Different regions $y_j (j = 1, 2, \ldots)$ of the system have different temperatures $T_j (j = 1, 2, \ldots)$, so we have

 doi:10.20944/preprints202502.0799.v1

$$\sum_j P(y_j)KL(P(x_i \mid y_j) \| P(x_i) = \sum_j P(y_j)[H(X) - S_j / (\mathrm{k}N_j)] \quad . \quad (63)$$

Since $P(y_j)=N_j/N$, we can get:

$$I(X;Y) = H(X) - S/(kN). \quad (64)$$

This formula shows the relationship between Shannon MI and physical entropy. It shows that the physical entropy S is similar to the posterior entropy $H(X|Y)$ of x. The above formula shows that the Maximum Entropy (ME) law in physics can be equivalently expressed as the minimum MI law.

According to (47) and (48), when the local equilibrium is reached, there is

$$I(X;Y) = \sum_j \sum_i P(x_i, y_j) \ln \frac{\exp[-e_i / (kT_j)]}{Z_j'} = H(Y_\theta) - H(Y_\theta \mid X) = I(X;Y_\theta). \quad (65)$$

The above formula shows that for a local equilibrium system, the minimum Shannon MI can be expressed by the semantic MI formula.

Helmholtz's free energy formula is:

$$F = E - TS, \quad (66)$$

where F is free energy, and E is the system's internal energy. When the internal energy and temperature remain unchanged, the increase in free energy is

$$\Delta F = -\Delta(TS) = TS - \sum_j T_j S_j = \mathrm{k}NTH(X) - \mathrm{k}N\sum_j T_j H(X \mid Y) \quad . \quad (67)$$

Comparing the above equation with Equations (63) and (64), we can find that Shannon MI is like the increase in free energy; semantic MI is like the increase in local equilibrium systems, which is smaller than Shannon MI, just as work is smaller than free energy. We can also regard kNT and kNT_j as the unit information values [5], so the increase in free energy is similar to the increase in information value.

5. The G Theory for Machine Learning

5.1. Basic Methods of Machine Learning: Learning Functions and Optimization Criteria

The most basic machine learning method is:

1. First, we use samples or sample distributions to train the learning functions with a specific criterion, such as maximum likelihood or RLS criterion;
2. Then, we make probability predictions or classifications utilizing the learning function with minimum distortion, minimum loss, or maximum likelihood criteria.

When learning, we generally use maximum likelihood or RLS criteria; the criteria may differ for different tasks when classifying. For prediction tasks where information is important, we generally use maximum likelihood and RLS criteria. To judge whether a person is guilty or not, where correctness is essential, we may use the minimum distortion (or loss) criterion. The maximum semantic information criterion is equivalent to the maximum likelihood criterion, similar to the RLS criterion. Compared with the minimum distortion criterion, the maximum semantic information criterion can reduce the underreporting of small probability events.

We generally do not use $\mathrm{P}(x|y_j)$ to train $P(x|\theta_j)$, because if $P(x)$ changes, the originally trained $P(x|\theta_j)$ will become invalid. Using parameterized transition probability function $P(\theta_j|x)$ as a learning function is unaffected by $P(x)$ changes. However, using $P(\theta_j|x)$ as a learning function also has essential defects. When category n >2, it is difficult to construct $P(\theta_j|x)(j=1,2,\ldots)$ because of the normalization restriction, that is, $\sum_j P(\theta_j|x)=1$ (for each x). As we will see below, there is no restriction when using truth or membership functions as learning functions.

5.2. For Multi-Label Learning and Classification

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

Consider multi-label learning, a supervised learning task. From the sample {(xk, yk), k = 1,2,…, N}, we can get the sample distribution $P(x, y)$. Then, use formula (16) or (18) for the optimized truth functions.

Assume that a truth function is a Gaussian function, there should be:

$$T(\theta_j \mid x) \propto \frac{P(x \mid y_j)}{P(x)} \propto P(y_j \mid x) \tag{68}$$

So, we can use the expectation and standard deviation of $P(x \mid y_j)/P(x)$ or $P(y_j \mid x)$ as the expectation and standard deviation of $T(\theta_j \mid x)$. If the truth function is like a dam cross-section, we can get it through some transformation.

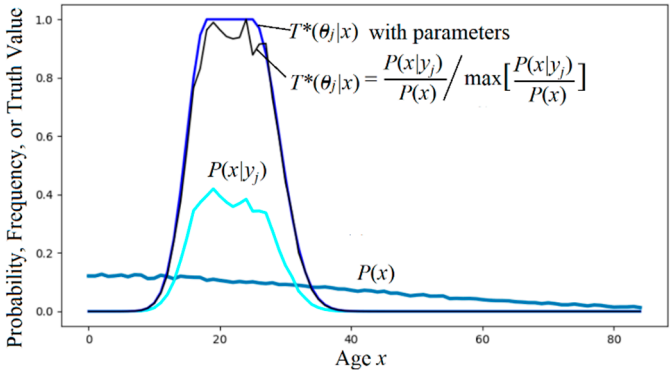


Figure 15. Using prior and posterior distributions $P(x)$ and $P(x \mid y_j)$ to obtain the optimized truth function $T^*(\theta_j \mid x)$. For details, see Appendix B in [8].

If we only know $P(y_j \mid x)$ but not $P(x)$, we can assume that $P(x)$ is equally probable, that is, $P(x)$=1/$\mid U \mid$, and then optimize the membership function using the following formula:

$$I(X;\theta_j) = \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} = \sum_i \frac{P(y_j|x_i)}{\sum_k P(y_j|x_k)} \log \frac{T(\theta_j \mid x_i)}{\sum_k T(\theta_j|x_k)} + \log \mid U \mid \tag{69}$$

For multi-label classification, we can use the classifier:

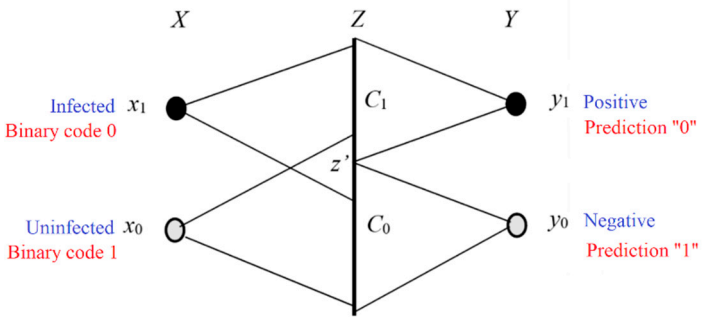
$$y_j^* = \arg\max_{y_j} I(x;\theta_j) = \arg\max_{y_j} \log \frac{T(\theta_j \mid x)}{T(\theta_j)} \tag{70}$$

If the distortion criterion is used, we can use –log $T(\theta_j \mid x)$ as the distortion function or replace $I(X; \theta_j)$ with $T(\theta_j \mid x)$.

The popular binary relevance method (Binary Relevance [61]) converts an n-label learning task into an n-pair label learning task. In comparison, the channel matching method is much simpler.

5.3. Maximum MI Classification for Unseen Instances

This type of classification belongs to semi-supervised learning. We take the medical test and the signal detection as examples (see Figure 16).



 doi:10.20944/preprints202502.0799.v1

Figure 5. Illustrating the medical test and the signal detection. We choose y_j according to $z \in C_j$. The task is to find the dividing point z' that results in MaxMI between X and Y.

The following algorithm is not limited to binary classifications. Let C_j be a subset of C and $y_j = f(z \mid z \in C_j)$; hence $S=\{C_1, C_2, \ldots\}$ is a partition of C. Our task is to find the optimized S, which is

$$S^* = \arg\max_S I(X;Y_\theta|S) = \arg\max_S \sum_j \sum_i P(C_j)P(x_i \mid C_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} . \quad (71)$$

First, we initiate a partition. Then we do the following iterations.

Matching I: Let the semantic channel match the Shannon channel and set the reward function. First, for given S, we obtain the Shannon channel:

$$P(y_j \mid x) = \sum_{z_k \in C_j} P(z_k \mid x),\ j = 1, 2, \ldots, n \quad . \quad (72)$$

Then we obtain the semantic channel $T(y \mid x)$ from the Shannon channel and $T(\theta_j)$ (or $m_\theta(x,y) = m(x, y)$). Then we have $I(x_i; \theta_j)$. For given z, we have conditional information as the reward function:

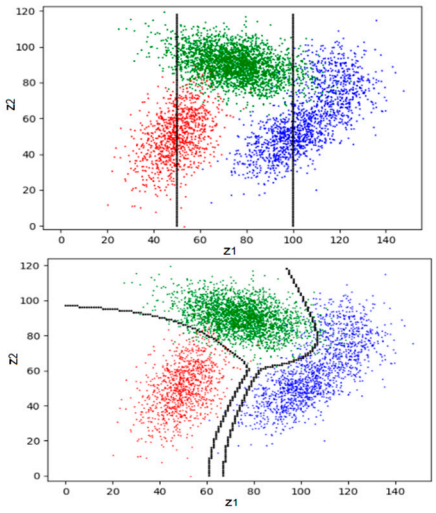
$$I(X;\theta_j|z) = \sum_i P(x_i \mid z) I(x_i;\theta_j) \ , \quad j=0,1,\ldots,n, \quad (73)$$

Matching II: Let the Shannon channel match the semantic channel by the classifier:

$$y_j^* = f(z) = \arg\max_{y_j} I(X;\theta_j|z) \ , j=0,1,\ldots,n. \quad (74)$$

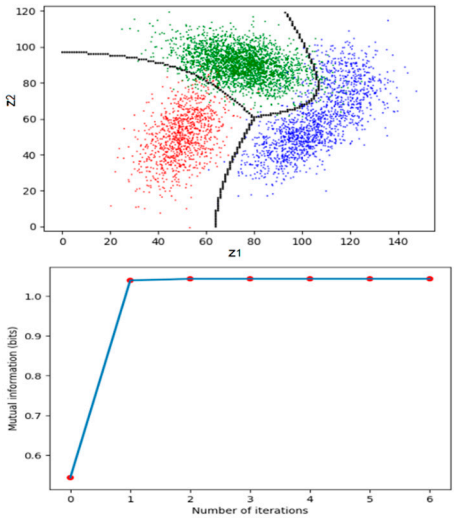
Repeat **Matching I** and **Matching II** until S does not change. Then, the convergent S is S^* we seek. The author explained the convergence with the $R(G)$ function (see Section 3.3 in [13]).

Figure 6 shows an example. The detailed data can be found in Section 4.2 of [13]. The two lines in Figure 6a represent the initial partition. Figure 6d shows that the convergence is very fast.



(a) (b)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1



(c) (d)

Figure 7. The maximum MI classification. (a) A very bad initial partition; (b) after the first iteration; (c) after the second iteration; (d) the MI changes with the iteration number.

However, this method is unsuitable for maximum MI classification in high-dimensional space. We need to combine neural network methods to explore more effective approaches.

5.4. Explanation and Improvement of the EM Algorithm for Mixed Models

The EM algorithm [45,62,63] is usually used for mixed models or clustering, an unsupervised learning method.

We know that $P(x)=\sum_j P(y_j)P(x|y_j)$. Given a sample distribution $P(x)$, we use $P_\theta(x)=\sum_j P(y_j)P(x|\theta_j)$ to approximate $P(x)$ so that the relative entropy or KL divergence $KL(P\|P\theta)$ is close to 0. $P(y)$ is the probability distribution of the latent variable to be sought.

The EM algorithm first presets $P(x|\theta_j)$ and $P(y_j)$, $j = 1, 2, \ldots, n$. E-step obtains:

$$P(y_j|x)=P(y_j)P(x|\theta_j)/P_\theta(x),\ P_\theta(x)=\sum_k P(y_k)P(x|\theta_k) \quad (75)$$

Then, 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: M1-step for

$$P^{+1}(y_j)=\sum_i P(x_i)P(y_j|x_i) \quad (76)$$

and M2-step for

$$P(x|\theta_j^{+1})=P(x)P(y_j|x)/P^{+1}(y_j)=P(x)\frac{P(x|\theta_j)}{P_\theta(x)}\frac{P(y_j)}{P^{+1}(y_j)}, \quad (77)$$

which optimizes the likelihood function. For Gaussian mixture models, we can use the expectation and standard deviation of $P(x)P(y_j|x)/P^{+1}(y_j)$ as the expectation and standard deviation of $P(x|\theta_j^{+1})$.

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. The converged $P(y)$ is the required probability distribution of the latent variable. According to the derivation process of the $R(G)$ function, the E-step and M1-step minimize the information difference $R–G$; the M-step maximizes the semantic MI. Therefore, the optimization criterion used by the EM algorithm is the MIE criterion.

 doi:10.20944/preprints202502.0799.v1

However, there are two problems with the above method to find the latent variable: 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 author improved the EM algorithm to the EnM algorithm [64]. The EnM algorithm includes the E-step for $P(y|x)$, n-step for $P(y)$, and M-step for $P(x|\theta_j)(j=1,2,\ldots)$. The n-step repeats the E-step and M1-step in the EM algorithm n times so that $P^{+1}(y) \approx P(y)$. The EnM algorithm also uses the MIE criterion. The n-step can speed up the solution of $P(y)$. M2-step only optimizes the likelihood functions. Because $P(y_j)/P^{+1}(y_j)$ is approximately equal to 1, we can use the following formula to optimize the model parameters:

$$P(x|\theta_j^{+1}) = P(x)P(x|\theta_j)/P_\theta(x). \quad (78)$$

Without 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 mixed 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.

The following mathematical formula proves that the EnM algorithm converges. After the M-step, the Shannon MI becomes:

$$R = \sum_i \sum_j P(x_i)\frac{P(x_i|\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(y_j|x_i)}{P^{+1}(y_j)}, \quad (79)$$

We define:

$$R'' = \sum_i \sum_j P(x_i)\frac{P(x_i|\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{P_\theta(x_i)}. \quad (80)$$

Then, we can deduce that after E-step, there is

$$KL(P||P_\theta) = R''-G=R-G+KL(P_Y^{+1}||P_Y), \quad (81)$$

where $KL(P||P_\theta)$ is the relative entropy or KL divergence between $P(x)$ and $P_\theta(x)$; the right KL divergence is:

$$KL(P_Y^{+1}||P_Y) = \sum_j P^{+1}(y_j)\log[P^{+1}(y_j)/P(y_j)]. \quad (82)$$

It is close to 0 after the n-step.

Equation (81) can be used to prove that the EnM algorithm converges. Because the M-step maximizes G, and the E-step and the n-step minimize R–G and $KL(P_Y{}^{+1})\|P_Y)$, $H(P\|P_\theta)$ can be close to 0. We can also use the above method to prove that the EM algorithm converges.

In most cases, the EnM algorithm performs better than the EM algorithm [64], especially when $P(y)$ is hard to converge.

Some researchers believe that EM makes the mixture model converge because the complete data log-likelihood $Q = -H(X,Y_\theta)$ continues to increase [39], or negative free energy $F' = H(Y)+Q$ continues to increase [45]. However, we can easily find counterexamples where R–G continues to decrease, but Q and F' do not necessarily continue to increase. Figure 18 shows the example used by Neal and Hinton [45], but the mixture proportion in the true model is changed from 0.3:0.7 to 0.7:0.3.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 February 2025 doi:10.20944/preprints202502.0799.v1

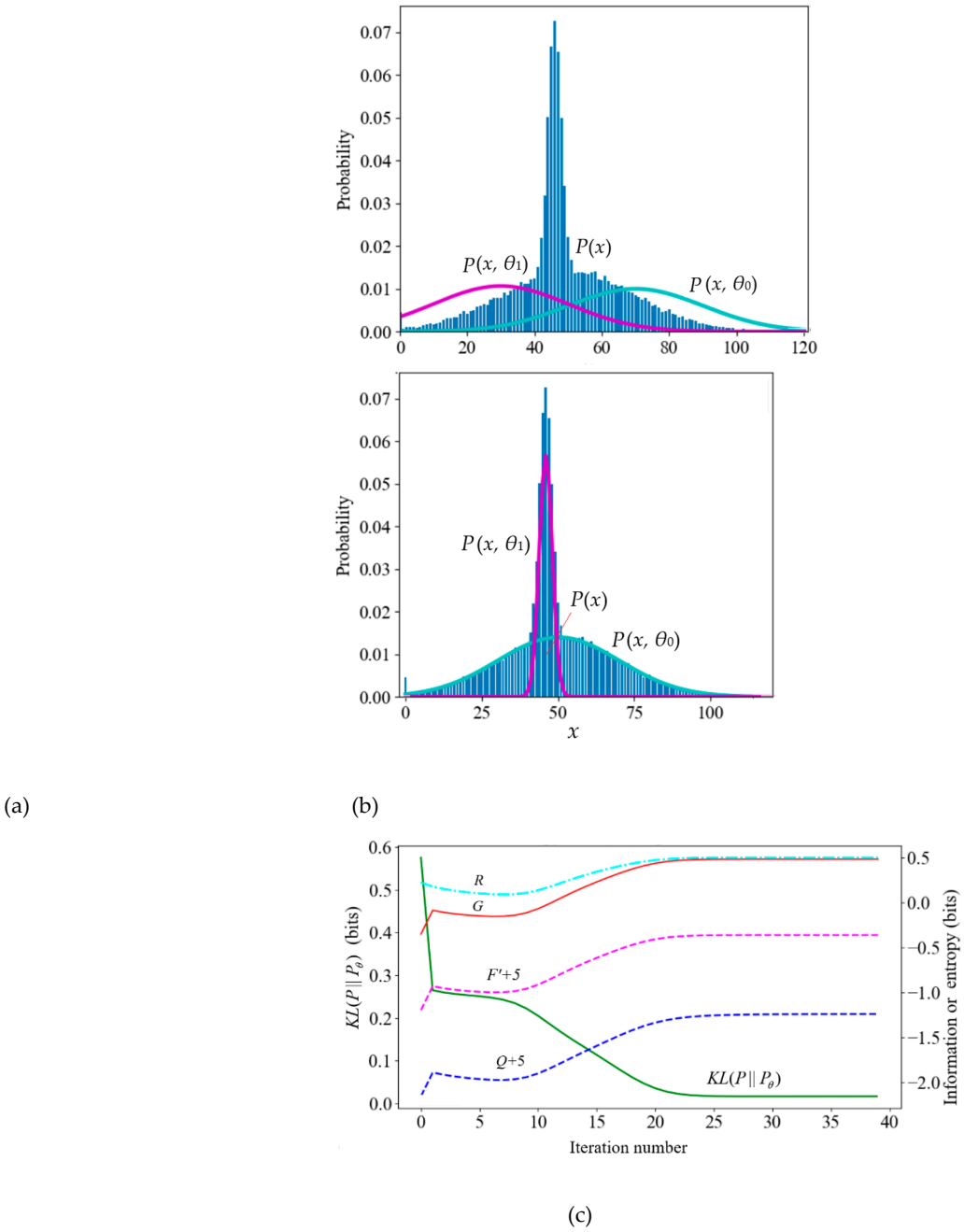


Figure 18. The convergent process of the mixture model from Neal and Hinton [45]. The mixture proportion is changed from 0.7:0.3 to 0.3:0.7. (a) The iteration starts; (b) the iteration converges; (c) the iteration process. $P(x, \theta_j)=P(yj)P(x \mid \theta_j)$ (j=0, 1).

This experiment shows that the decrease in R–G, not the increase in Q or F', is the reason for the convergence of the mixture model.

The free energy of the true mixture model (with true parameters) is the Shannon conditional entropy $H(X \mid Y)$. If the standard deviation of the true mixture components is large, $H(X \mid Y)$ is also large. If the initial standard deviation is small, F is small initially. After the mixture model converges, F must be close to $H(X \mid Y)$. Therefore, F increases (i.e., F' decreases) during the convergence process. Many experiments [25,64] have shown that this is indeed the case.

Equation (77) can also explain pre-training in deep learning, where we need to maximize the model's predictive ability and minimize the information difference R–G (or compress data).

5.5. Semantic Variational Bayes: A Simple Method for Solving Hidden Variables

 doi:10.20944/preprints202502.0799.v1

Given $P(x)$ and constraints $P(x|\theta_j)$, j=1,2,…, we need to solve $P(y)$ that produces $P(x)=\sum_j P(y_j)P(x|\theta_j)$. $P(y)$ is the probability distribution of the latent variable y, sometimes called the latent variable. The popular method is the Variational Bayes method (VB for short) [65]. This method originated from the article by Hinton and Camp [44]. It was further discussed and applied in the articles by Neal and Hinton [45], Beal [66], and Koller [67] (ch.11). Gottwald and Braun's article "*Two Free Energy and the Bayesian Revolution*" [68] discusses the relationship between the MFE principle and ME principle in detail.

VB uses $P(y)$ (usually written as $g(y)$) as a variation to minimize the following function:

$$F=\sum_i P(x_i|y_j)\sum_j P(y_j)\log\frac{P(y_j)}{P_\theta(x_i,y_j)}=-F'(H(Y),Q)=-[Q+H(Y)] \quad (83)$$

It is equal to the semantic posterior entropy $H(X|Y_\theta)$ of X. The smaller F is, the larger the semantic MI $I(X;Y))=H(X)-H(X|Y_\theta)$ is.

It is easy to prove that when the semantic channel matches the Shannon channel, that is, $T(\theta_j|x)\propto P(y_j|x)$ or $P(x|\theta_j)=P(x|y_j)$ (j=1,2,…), F is minimized and the semantic MI is maximized. This can optimize the prediction model $P(x|\theta_j)$(j=1,2,…), but it cannot optimize $P(y)$. For optimizing $P(y)$, the mean field approximation [45,65] is usually used; that is, $P(y|x)$ instead of $P(y)$ is used as the variation. Only one $P(y_j|x)$ is optimized at a time, and the other $P(y_k|x)$ ($k\neq j$) remains unchanged. Minimizing F in this way is actually maximizing the log-likelihood of x or minimizing $KL(P||P_\theta)$. In this way, optimizing $P(y|x)$ also indirectly optimizes $P(y)$.

Unfortunately, when optimizing $P(y)$ and $P(y|x)$, F may not decrease (see Figure 18). So, VB is good as a tool and is imperfect as a theory.

Fortunately, it is easier to solve $P(y|x)$ and $P(y)$ using the MID iteration in solving $R(D)$ and $R(G)$ functions. The MID iteration plus LBI for optimizing the prediction model is the Semantic Variational Bayes' method (abbreviated as SVB) [30]. It uses the MIE criterion.

When the constraint changes from likelihood functions to truth functions or similarity functions, $P(y_j|x_i)$ in the MID iteration formula is changed from

$$P(y|x_i)=P(y)\left[\frac{P(x_i|\theta_j)}{P_\theta(x_i)}\right]^s\Big/\sum_k P(y_k)\left[\frac{P(x_i|\theta_j)}{P_\theta(x_i)}\right]^s \quad (84)$$

to

$$P(y|x_i)=P(y)\left[\frac{T(\theta_j|x_i)}{T(\theta_j)}\right]^s\Big/\sum_k P(y_k)\left[\frac{T(\theta_k|x_i)}{T(\theta_k)}\right]^s. \quad (85)$$

From $P(x)$ and the new $P(y|x)$, we can get the new $P(y)$. Repeating the formulas for $P(y|x)$ and $P(y)$ will lead to convergence of P(y). Using s allows us to tighten the constraints for increasing R and G. Choosing proper s enables us to balance between maximizing semantic information and maximizing information efficiency.

The main tasks of SVB and VB are the same: using variational methods to solve latent variables based on observed data and constraints. The differences are:

1. **Criteria:** In the definition of VB, it adopts the MFE (i.e., minimum semantic posterior entropy) criterion, whereas, for solving $P(y)$, it uses $P(y|x)$ as the variation, actually uses the maximum likelihood criterion that makes the mixture model converge. In contrast, SVB uses the MID criterion, equal to the maximum likelihood criterion (optimizing model parameters) plus the ME criterion.
2. **Variational method**: VB only uses $P(y)$ or $P(y|x)$ as the variation, while SVB alternatively uses $P(y|x)$ and $P(y)$ as the variation.
3. **Computational complexity**: VB uses logarithmic and exponential functions to solve $P(y|x)$ [65]; the calculation of $P(y|x)$ in SVB is relatively simple (for the same task, i.e., when s=1).
4. **Constraints**: VB only uses likelihood functions as constraint functions. In contrast, SVB allows using various learning functions (including likelihood, truth, membership, similarity, and

 doi:10.20944/preprints202502.0799.v1

distortion functions) as constraints. In addition, SVB can use the parameter s to enhance constraints.

Because SVB is more compatible with the maximum likelihood criterion and the ME principle, it should be more suitable for many occasions in machine learning. However, because it does not consider the probability of parameters, it may not be as applicable as VB in some occasions. See [30] for more details of SVB.

5.6. Bayesian Confirmation and Causal Confirmation

Logical empiricism was opposed by Popper's falsificationism [19,20], so it turned to confirmation (i.e., Bayesian confirmation) instead of induction or positivism [69,70]. Bayesian confirmation was previously a field of concern for researchers in the philosophy of science [61,72], and now many researchers in natural sciences have also begun to study it [26,73,74]. The reason is that uncertain reasoning requires major premises, which need to be confirmed.

The main reasons why researchers have different views on Bayesian confirmation are:

1. There are no suitable mathematical tools; for example, statistical and logical probabilities are not well distinguished.
2. Many people do not distinguish between the confirmation of the relationship (i.e. →) in the major premise y→x and the confirmation of the consequent (i.e., x occurs);
3. No confirmation measure can reasonably clarify the Raven Paradox.

To clarify the Raven paradox, the author wrote the article "Channels' confirmation and predictions' confirmation: from medical tests to the Raven paradox" [26].

In the author's opinion, the task of Bayesian confirmation is to evaluate the support of the sample distribution for the major premise. For example, for the medical test (see Figure 16), a major premise is "If a person tests positive (y_1), then he is infected (x_1)", abbreviated as $y_1 \to x_1$. For a channel's confirmation, a truth (or membership) function can be viewed as a combination of a clear truth function $T(y_1|x) \in \{0,1\}$ and a tautology's truth function (always 1):

$$T(\theta_1|x) = b_1 T(y_1|x) + b_1'. \quad (82)$$

A tautology's proportion b_1' is the degree of disbelief. The credibility is b_1, and its relationship with b_1' is $b_1'=1-|b_1|$. See Figure 19.

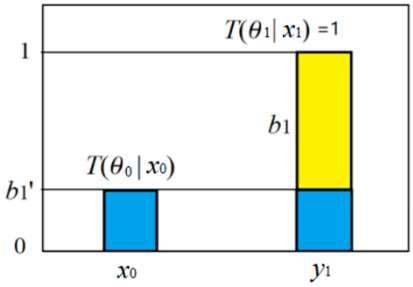


Figure 19. A truth function includes a believable proportion b_1 and unbelievable proportion $b_1' = 1 - |b_1|$.

We change b_1 to maximize the semantic KL information $I(X; \theta_1)$, the optimized b_1, denoted as b_1^*, is the confirmation degree:

$$b_1^* = b^*(y_1 \to x_1) = \frac{P(y_1|x_1) - P(y_1|x_0)}{\max(P(y_1|x_1), P(y_1|x_0))} = \frac{R^+ - 1}{\max(R^+, 1)}, \quad (87)$$

where $R^+=P(y_1|x_i)/P(y_1|x_0)$ is the positive likelihood ratio, indicating the reliability of the test-positive. This conclusion is compatible with medical test theory.

Considering the prediction confirmation degree, we assume that $P(x|\theta_1)$ is a combination of the 0-1 part and the equal probability part. The ratio of the 0-1 part is the prediction credibility, and the optimized credibility is the prediction confirmation degree:

$$c_1^* = c^*(y_1 \to x_1) = \frac{P(x_1|y_1) - P(x_0|y_1)}{\max(P(x_1|y_1), P(x_0, y_1))} = \frac{a - c}{\max(a, c)}, \quad (88)$$

 doi:10.20944/preprints202502.0799.v1

where a is the number of positive examples, and c is the number of negative examples.

Both confirmation degrees can be used for probability predictions, that is, to calculate $P(x|\theta_1)$[26].

Hemple proposed a confirmation paradox, namely the raven paradox [61]. According to the equivalence condition in classical logic, "if x is a raven, then x is black" (Rule 1) is equivalent to "if x is not black, then x is not a raven" (Rule 2). According to this, white chalk supports Rule 2; therefore, it also supports Rule 1. However, according to common sense, a black crow supports Rule 1, and a non-black Raven opposes Rule 1; something that is not a Raven, such as a black cat or a white chalk, is irrelevant to Rule 1. Therefore, there is a paradox between the equivalence condition and common sense. Using the confirmation measure c_1*, we can be sure that common sense is correct and the equivalence condition is wrong (for fuzzy major premises), thus eliminating the Raven paradox. However, other confirmation measures cannot eliminate the Raven paradox [26].

Causal probability is used in causal inference theory [75]:

$$P_d = \max[0, \frac{P(y_1|x_1)-P(y_1|x_0)}{P(y_1|x_1)}] = \max(0, \frac{R^+-1}{R^+}) \quad (89)$$

It indicates the necessity of the cause $x1$ replacing $x0$ to lead to the result y1. Where P(y1|x)=P(y1|do(x)) is the posterior probability of y1 caused by intervention x. The author uses the semantic information method to obtain the channel causal confirmation degree [27]:

$$Cc(x_1/x_0 => y_1) = b_1^* = \frac{P(y_1|x_1)-P(y_1|x_0)}{\max(P(y_1|x_1), P(y_1|x_0))} = \frac{R^+-1}{\max(R^+,1)} \quad (90)$$

It is compatible with the above causal probability but can express negative causal relationships, such as the necessity of vaccines inhibiting infection.

5.7. Emerging and Potential Applications

1) About self-supervised Learning

Applications of estimated MI have emerged in the field of self-supervised learning. The estimated MI is a special case of semantic MI. Both MINE proposed by Belghazi et al. [34] and InfoNCE proposed by Oord et al. [35] use estimated MI.

MINE and InfoNCE are essentially the same as the semantic information methods. Their common features are:

1. The membership function $T(\theta_j|x)$ or similarity function $S(x, y_j)$ proportional to $P(y_j|x)$ is used as the learning function. Its maximum value is generally 1, and its average is the partition function Z_j.
2. The estimated information or semantic information between x and y_j is $\log[T(\theta_j|x)/Z_j]$ or $\log[S(x, y_j)/Z_j]$.
3. The statistical probability distribution $P(x, y)$ is still used when calculating the average information.

However, many researchers are still unclear about the relationship between estimated MI and Shannon MI. The G theory's $R(G)$ function can help readers understand this relationship.

2) About Reinforcement Learning

Goal-oriented information introduced in Section 4.2 can be used as a reward for reinforcement learning. Assuming that the probability distribution of x in state s_k is $P(x|a_{k-1})$, which becomes $P(x|a_k)$ in state s_{k+1}. The reward of a_k is:

$$r_k = I(X;a_k/\theta_j) - I(X;a_{k-1}/\theta_j) = \sum_i [P(x_i|a_k) - P(x_i|a_{k-1})] \log \frac{T(\theta_j|x_i)}{T(\theta_j)}. \quad (91)$$

Reinforcement learning is to find the optimal action sequence $a_1, a_2, \ldots$, so that the sum of rewards $r_1+r_2+\ldots$ is maximized.

Like constraint control, reinforcement learning also needs the trade-off between maximum purposefulness and minimum control cost. The $R(G)$ function should be helpful.

 doi:10.20944/preprints202502.0799.v1

3) About the Truth function and Fuzzy Logic for Neural Networks

When we use the truth, distortion, or similarity function as the weight parameters of the neural network, the neural network contains semantic channels. Then, we can use semantic information methods to optimize the neural network. Using the truth function $T(\theta_j|x)$ as weight is better than using the parameterized inverse probability function $P(\theta_j|x)$ because there is no normalization restriction when using truth functions.

However, unlike the clustering of points on the plane, a point becomes an image for the clustering of graphics, and the similarity function between images needs different methods. A common method is to regard an image as a vector and use cosine similarity between vectors. However, cosine similarity may have negative values, which require activation functions and biases to make necessary conversions. Combining existing neural network methods and channel-matching algorithms needs further exploration.

Fuzzy logic, especially fuzzy logic compatible with Boolean algebra, seems to be useful in neural networks; for example, the activation function $Relu(a–b)$ = max(0, $a–b$) commonly used in neural networks is the logical difference operation $f(a\ \overline{b}\ )$=max(0, $a–b$) used in the author's color vision mechanism model [76–78]. Truth functions, fuzzy logic, and the semantic information method used in neural networks should make neural networks easier to understand.

4) Explaining Data Compression in Deep Learning

To explain the success of deep neural networks such as AutoEncoders [36] and Deep Belief Networks [79], Tishby et al. [39] proposed the information bottleneck explanation, arguing that when optimizing deep neural networks, we maximize the Shannon MI between some layers and minimize the Shannon MI between other layers. However, from the perspective of the $R(G)$ function, each coding layer of the Autoencoder needs to maximize the semantic MI G and minimize the Shannon MI R; pre-training is to let the semantic channel match the Shannon channel so that $G\approx R$ and $KL(P||P_\theta)\approx 0$ (as if for mixture models to converge). Fine-tuning increases R and G at the same time by increasing s (making the partition boundaries steeper).

Not long ago, researchers at OpenAI [80,81] explained General Artificial Intelligence by lossless (actually, loss-limited) data compression, similar to the explanation of using MIE.

6. Discussion and Summary

6.1. Why Is the G Theory a Generalization of Shannon's Information Theory?

First, the semantic information G measure is a generalization of Shannon's information measure. The methods are:

1. In addition to the probability prediction $P(x|y_j)$, the semantic probability prediction $P(x|\theta_j)=P(x)T(\theta_j|x)/T(\theta_j)$ is also used;
2. The G measure also has coding meaning, which means the average code length saved by the semantic probability prediction.

Second, the semantic communication model is essentially the Shannon communication model; the difference is that it changes the distortion constraint to the semantic constraint, including semantic distortion constraint (for $R(\Theta)$), semantic information constraint (for the $R(G)$ function) and the semantic information loss constraint (for electronic communication, see Section 3.2).

Third, the G theory adheres to Shannon's concept of information: information is reduced uncertainty.

6.2. What Is Information?

What is information? This question has many answers [82]. According to Shannon's definition, information is uncertainty reduced. Shannon information is the uncertainty reduced due to the increase of probability, while semantic information is the uncertainty reduced due to the narrowing of concepts' extensions.

 doi:10.20944/preprints202502.0799.v1

From a common-sense perspective, information refers to something previously unknown or uncertain, which encompasses:

Information from natural language: information provided by answers to interrogative sentences (e.g., sentences with "Who?", "What?", "When?", "Where?", "Why?", "How?", or "Is this?").

Perceptual or observational information: information obtained from material properties or observed phenomena.

Symbolic information: information conveyed by various symbols like road signs, traffic lights, and battery polarity symbols.

Quantitative indicators' information: information provided by data such as time, temperature, rainfall, stock market indices, and inflation rates.

Associated information: information derived from event associations, such as the crowing of a rooster signaling dawn and a positive medical test indicating disease.

Items 2, 3, and 4 can also be viewed as answers to questions in item 1, thus providing information. These forms of information involve concept extensions and truth-falsehood considerations and should be semantic information. Associated information in item 5 can be measured using Shannon's or the semantic information formula. When probability predictions are inaccurate (i.e., $P(x\mid\theta_j)\neq P(x\mid y_j)$), the semantic information formula is more appropriate. Thus, the G theory is consistent with the concept of information in everyday life.

In computer science, information is often defined as useful, structured data. What qualifies as "useful"? This utility arises because the data can answer various questions or provide associated information. Therefore, the definition of information in data science also ties back to reduced uncertainty and narrowed concept extensions.

6.3. Relationships and Differences Between the G Theory and Other Demantic Information Theories

6.3.1. Carnap and Bar-Hillel's Semantic Information Theory

The semantic information measure of Carnap and Bar-Hillel is [3]:

$$Ip=\log(1/m_p], \quad (8)$$

where I_p is the semantic information provided by the proposition set p, and m_p is the logical probability of p. This formula reflects Popper's idea that smaller logical probabilities convey more information. However, as Popper noted, this idea requires the hypothesis to withstand factual testing. The above formula does not account for such tests, implying that correct and incorrect hypotheses provide the same information.

Additionally, the G theory differs in calculating logical probability with statistical probability, unlike Carnap and Bar-Hillel's approach.

6.3.2. Dretske's Knowledge and Information Theory:

Dretske [9] emphasized the relationship between information and knowledge, viewing information as content tied to facts and knowledge acquisition. Though he did not propose a specific formula, his ideas about information quantification include:

1. The information must correspond to facts and eliminate all other possibilities.
2. The amount of information relates to the extent of uncertainty eliminated.
3. Information used to gain knowledge must be true and accurate.

The G theory aligns with these principles by providing semantic information formulas and mathematically implementing Dretske's concepts.

6.3.3. Florida's Strong Semantic Information Theory:

Florida's theory [12] emphasizes:

1. The information must contain semantic content and be consistent with reality:
2. False or misleading information cannot qualify as true information.

 doi:10.20944/preprints202502.0799.v1

Floridi elaborated on Dretske's ideas and introduced a strong semantic information formula. However, this formula is more complex and less effective at reflecting factual testing compared to the G theory. For instance, Floridi's approach ensures tautologies and contradictions yield zero information but fails to penalize false predictions with negative information.

6.3.4. Other Semantic Information Theories:

In addition to the semantic information theories mentioned above, other well-known ones include the theory based on fuzzy entropy proposed by Zhong [11] and the theory based on synonymous mapping proposed by Niu and Zhang [17]. Zhong advocated the combination of information science and artificial intelligence, which had a great influence on China's semantic information theory research. He employed fuzzy entropy to define the semantic information measure. However, this approach yielded identical maximum values (1 bit) for both true and false sentences [11], which is counterintuitive. Other people's semantic information measures using DeLuca and Termini's fuzzy entropy also encounter similar problems.

Other authors who discussed semantic information measure and semantic entropy include D'Alfonso [84], Basu et al. [85], and Melamed [86]. These authors improved semantic information measures by improving Carnap and Bar-Hillel's logical probability. The semantic entropy used is mainly in the form of Shannon entropy. The semantic entropy $H(Y_\theta)$ in G theory differs from these semantic entropies. It contains statistical and logical probabilities and reflects the average code length of lossless coding (see Section 2.3).

Niu and Zhang [17] and Guo et al. [87] proposed the semantic information rate-distortion function $R_s(D)$, where R_s represents minimum semantic information. In contrast, $R(G)$ in the G theory still represents the minimum Shannon MI, reflecting the lower limit of data compression. Why do we minimize semantic MI? The reason seems to be that researchers want to establish a semantic information theory parallel to Shannon's information theory. Liu et al. [88] used the dual-constrained rate-distortion function $R(D_s, D_x)$; Guo et al. [87] also used $R(Ds, D_x)$, which is meaningful. In contrast, G in $R(G)$ already contains dual constraints because G means fidelity and semantic information.

In addition, fuzzy information theory [89,90] and generalized information theory [91] also involve semantics. However, these theories are further from Shannon's information theory.

6.4. Relationship Between the G Theory and Kolmogorov Complexity Theory

Kolmogorov [92] defined the complexity of a string of data as the shortest code length under the requirement of lossless recovery. The information provided by knowledge is defined as the complexity reduced by knowledge. Shannon's information measured can be understood as the average information, while Kolmgorov's information is the information provided by knowledge about a data string. Shannon's information theory does not consider the complexity of individual data, while Kolmogorov's theory does not consider statistical averages. It can be said that Kolmogorov defined the amount of information in microdata, while Shannon provided the mutual information formula for measuring the information of macrodata. The two theories are complementary.

Because knowledge includes the extensions of concepts and the logical relationships between concepts, as well as the correlations (including causality) between things, the information defined by Kolmogorov contains semantic information. However, Kolmogorov did not provide a specific formula for measuring information. The G theory provides semantic Bayesian prediction, semantic information measure, and the $R(G)$ function, which should supplement the above two theories.

The relationship between the G theory and Kolmogorov complexity needs further study.

6.5. Comparison of the MIE Principle and the MFE Principle

Friston proposed the MFE principle, which he believed was a universal principle organisms use to perceive the world and adapt to the environment (including transforming the environment). The

 doi:10.20944/preprints202502.0799.v1

core mathematical method he uses is VB. A similar principle used by the G theory is the MIE principle.

The main differences between the two are:

1. The G theory regards Shannon's MI $I(X;Y)$ as free energy, while Friston's theory regards the semantic posterior entropy $H(X|Y_\theta)$ as free energy.
2. The methods for finding the latent variable $P(y)$ and the Shannon channel $P(y|x)$ are different. Friston uses VB, and the G theory uses SVB.

When optimizing the prediction model $P(x|\theta_j)(j=1,2,\ldots)$, the two are consistent; when optimizing $P(y)$ and $P(y|x)$, the results of the two are similar, but the methods are different. SVB is simpler. The reason why the results are similar is that VB uses the mean-field approximation when optimizing $P(y|x)$, which is equivalent to using $P(y|x)$ instead of $P(y)$ as a variation and actually uses the MID criterion. So, the two results are almost the same. Figure 18 shows that in a mixture model's convergence process, information difference $R-G$ instead of free energy F continuously decreases.

In physics, free energy is energy that can be used to do work; the more, the better. Why should it be minimized? In physics, there are two situations in which free energy is reduced. One is passive reduction because of the increase in entropy. The other reason is to save the consumed free energy while doing work. This is for considering thermal efficiency, which is a conditional reduction. Reducing the consumed free energy conforms to Jaynes' maximum entropy principle. Therefore, from a physics perspective, it is not easy to understand that one would actively minimize free energy.

MIE is like the maximum doing-work efficiency W/F when using free energy F to do work W. The MIE principle is easier to understand.

The author will discuss these two principles further in other articles.

6.6. Limitations and Areas That Need Exploration

The G theory is still a basic theory. It has limitations in many aspects, and many elements need improvement.

1) Semantics and Distortion of Complex Data

Truth functions can represent the semantic distortion of labels. However, it is difficult to express the semantics, semantic similarity, and semantic distortion of complex data (such as a sentence or an image). Many researchers have made valuable explorations [15,93]. The author's research is insufficient.

The semantic relationship between a word and many other words is also very complex. Innovations like Word2Vec [94,95] in deep learning have successfully modeled these relationships, paving the way for advancements like Transformer [96] and ChatGPT. Future work in the G theory should aim to integrate such developments to align with the progress in deep learning.

2) Feature Extraction

The features of images encapsulate most semantic information. There are many efficient feature extraction methods in deep learning, such as Convolutional Neural Networks and AutoEncoders. These methods are ahead of the G theory. Whether the G theory can be combined with these methods to obtain better results needs further exploration.

3) The Channel-matching Algorithm for Neural Networks

Establishing neural networks to enable mutual alignment of Shannon and semantic channels appears feasible. Current deep learning practices, relying on gradient descent and backpropagation, demand significant computational resources. If the channel matching algorithm can reduce reliance on these methods, it would save computational power and physical free energy.

4) **Neural Networks Utilizing Fuzzy Logic**

Using truth functions or their logarithms as network weights facilitates the application of fuzzy logic. The author previously used a fuzzy logic method compatible with Boolean algebra to set up a symmetrical color vision mechanism model: the 3-8 decoding model [76–79]. Combining the G theory with fuzzy logic and neural networks holds promise for further exploration.

5) **Optimizing Economic Indicators and Forecasts:**

 doi:10.20944/preprints202502.0799.v1

Forecasting in weather, economics, and healthcare domains provides valuable semantic information. Traditional evaluation metrics, such as accuracy or average error, can now be enhanced by converting distortion into truth. Using truth functions to represent semantic information offers a novel method for evaluating and optimizing forecasts, which merits further exploration.

6.7. Conclusion

The G theory's validity is supported by its broad applications across multiple domains, particularly in solving problems related to semantic communication, semantic compression theory, multi-label learning, maximum mutual information classification, mixture models, latent variable resolution, Bayesian confirmation, constraint control, investment portfolios and information value. Particularly, the G theory allows us to utilize the existing coding methods for semantic communication. The crucial approach is to replace distortion constraints with semantic constraints, where the information rate-distortion function becomes the information rate-fidelity function.

However, the G theory's primary limitation lies in the semantic representation of complex data. In this regard, it has lagged behind the advancements in deep learning. Bridging this gap will require learning from and integrating insights from other studies and technologies.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: ……. The Author particularly thanks Viacheslav Kovtun. Without his encouragement, the Author would not have finished this review.

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

Appendix A. Abbreviations

Abbreviation	Original text
EM	Expectation-Maximization
EnM	Expectation-n-Maximization
GPS	Global Positioning System
G theory	Semantic information G theory (G means generalization)
InfoNCE	Information Noise Contrast Estimation
KL	Kullback–Leibler
LBI	Logical Bayes' Inference
ME	Maximum Entropy
MI	Mutual Information
MIE	Maximum Information Efficiency
MID	Minimum Information Difference
MINE	Mutual Information Neural Estimation
SVB	Variational Byes
VB	Semantic Variational Byes

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* 1948, *27*, 379–423.
2. Weaver, W. Recent contributions to the mathematical theory of communication. In *The Mathematical Theory of Communication*, 1st ed.; Shannon, C.E., Weaver, W., Eds; The University of Illinois Press: Urbana, IL, USA, 1963; pp. 93–117.
3. Carnap, R.; Bar-Hillel, Y. *An Outline of a Theory of Semantic Information*; Tech. Rep. No. 247; Research Laboratory of Electronics, MIT: Cambridge, MA, USA, 1952.

4. 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).
5. Lu, C. *A Generalized Information Theory*; China Science and Technology University Press: Hefei, China, 1993; ISBN 7-312-00501-2. (in Chinese)
6. Lu, C. Meanings of generalized entropy and generalized mutual information for coding. *J. China Inst. Commun.* 1994, *15*, 37–44. (in Chinese)
7. Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* 1999, *28*, 453–490.
8. Lu, C. Semantic Information G Theory and Logical Bayesian Inference for Machine Learning. Information, 2019, *10*, 261.
9. Dretske, F., 1981, *Knowledge and the Flow of Information*, Cambridge, Massachusetts The MIT Press. ISBN 0-262-04063-8
10. Wu, W. General source and general entropy. Journal of Beijing University of Posts and Telecommunications, 1982, *5*, 29-41. (in Chinese). 吴伟陵 [1]吴伟陵.广义信息源与广义熵[J].北京邮电大学学报, 1982, 5(1):29-41.
11. Zhong, Y. A Theory of Semantic Information. Proceedings 2017, 1, 129. https://doi.org/10.3390/IS4SI-2017-04000
12. Floridi L. Outline of a theory of strongly semantic information. Minds and Machines, 2004, *14*, 197-221.
13. Floridi, L. Semantic conceptions of information. In *Stanford Encyclopedia of Philosophy*; Stanford University: Stanford, CA, USA, 2005. Available online: http://seop.illc.uva.nl/entries/information-semantic/ (accessed on 17 June 2020).
14. D'Alfonso, S. On Quantifying Semantic Information. *Information* 2011, *2*, 61–101.
15. Xin, G.; Fan, P.; Letaief, K.B. Semantic Communication: A Survey of Its Theoretical Development. *Entropy* 2024, *26*, 102. https:// doi.org/10.3390/e26020102
16. Strinati, E.C.; Barbarossa, S. 6G networks: Beyond Shannon towards semantic and goal-oriented communications. Comput. Netw. 2021, 190, 107930.
17. Kai Niu, Ping Zhang, A mathematical theory of semantic communication, *Journal on Communications* 2024, *45*, 8-59.
18. Davidson, D. Truth and meaning. *Synthese* 1967, *17*, 3, 304-323.
19. Popper, K. *Logik Der Forschung: Zur Erkenntnistheorie Der Modernen Naturwissenschaft*; Springer: Vienna, Austria, 1935; English translation: *The Logic of Scientific Discovery*; Routledge Classic: London, UK; New york, NY, USA, 2002.
20. K. Popper, Conjectures and Refutations, 1st ed.; London and New York: Routledge, 2002.
21. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc.* 1922, *222*, 309–368.
22. Zadeh, L.A. Fuzzy sets. *Inf. Control* 1965, *8*, 338–353.
23. Zadeh, L.A. Probability measures of fuzzy events. *J. Math. Anal. Appl.* 1986, *23*, 421–427.
24. Lu, C. The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning. Philosophies 2020, *5*, 25.
25. Lu C. Reviewing evolution of learning functions and semantic information measures for understanding deep learning. Entropy 2023, 25, 802.
26. Lu, C. Channels' confirmation and predictions' confirmation: From the medical test to the raven paradox. Entropy 2020, *22*, 384. https://www.mdpi.com/1099-4300/22/4/384.
27. Lu, C. Causal Confirmation Measures: From Simpson's Paradox to COVID-19. Entropy 2023, *25*, 143. https://doi.org/10.3390/e25010143.
28. 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.
29. Lu C. Semantic Information G Theory for Range Control with Tradeoff between Purposiveness and Efficiency, Available online: https://arxiv.org/abs/2411.05789. (accessed on 1 January 2025).
30. 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)

 doi:10.20944/preprints202502.0799.v1

31. Shannon, C.E. Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 1959, 4, 142–163.
32. Berger, T. *Rate Distortion Theory*; Prentice-Hall: Enklewood Cliffs, NJ, USA, 1971.
33. Zhou, J.P. Fundamentals of information theory, Beijing, China: People's Posts and Telecommunications Press, 1983. (in Chinese).
34. 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.
35. 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 2023).
36. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* 2006, *313*, 504-507.
37. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Trischler, A.; Bengio, Y. Learning Deep Representations by Mutual Information Estimation and Maximization. Available online: https://arxiv.org/abs/1808.06670. (accessed on 22 December 2022).
38. Tschannen, M.; Djolonga, J.; Rubenstein, P.K.; Gelly, S.; Luci, M. On Mutual Information Maximization for Representation Learning. Available online: https://arxiv.org/pdf/1907.13625.pdf (accessed on 23 February 2023).
39. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5.
40. Tarski, A. The semantic conception of truth: and the foundations of semantics. *Philos. Phenomenol. Res.* 1994, *4*, 341–376.
41. Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitrechnung*; Ergebnisse Der Mathematik (1933); translated as *Foundations of Probability*; Dover Publications: New York, NY, USA, 1950.
42. von Mises, R. *Probability, Statistics and Truth*, 2nd ed.; George Allen and Unwin Ltd.: London, UK, 1957.
43. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* 1974, *19*, 716–723.
44. 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.
45. 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.*
46. Friston, K. The free-energy principle: a unified brain theory?. *Nat Rev Neurosci* 2010, 11, 127–138. https://doi.org/10.1038/nrn2787.
47. Wang, P.Z. From the fuzzy statistics to the falling random subsets. In *Advances in Fuzzy Sets, Possibility Theory and Applications*; Wang, P.P., Ed.; Plenum Press: New York, NY, USA, 1983; pp. 81–96.
48. Wang, P.Z. *Fuzzy* Sets *and* Falling Shadows of Random Set; Beijing Normal University Press: Beijing, China, 1985. (In Chinese).
49. Farsad, N.; Rao, M.; Goldsmith, A. Deep learning for joint source-channel coding of text. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: New York, NY, USA, 2018; pp. 2326–2330.
50. Güler, B.; Yener, A.; Swami, A. The semantic communication game. IEEE Trans. Cogn. Commun. Netw. 2018, 4, 787–802.
51. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
52. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.H. Deep learning enabled semantic communication systems. IEEE Trans. Signal Process. 2021, 69, 2663–2675.
53. Markowitz, H.M. Portfolio selection. *The Journal of Finance* 1952, 7, 77–91. doi:10.2307/2975974
54. Kelly, J. L. A new interpretation of information rate. *Bell System Technical Journal* 1956, 35, 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x.
55. Latané H. A.; Tuttle, D. A. Criteria for Portfolio Building. *The Journal of Finance* 1967, 22, 359-373.

 doi:10.20944/preprints202502.0799.v1

56. Arrow, K.J. The economics of information: An exposition. *Empirica* 1996, 23, 119–128. https://doi.org/10.1007/BF00925335
57. Cover, T. M. Universal portfolios. *Mathematical Finance*. 1991, *1*, 1–29. doi:10.1111/j.1467-9965.1991.tb00002.x. S2CID 219967240.
58. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2006.
59. Lu, C. The Entropy Theory of Portfolios and Information Values. Hefei, China: China Science and Technology University Press, 1997. ISBN7-312-00952-2F.36. (in Chinese).
60. Jaynes, E.T. Probability Theory: *The* Logic of Science. Bretthorst, G.L., Ed.; Cambridge University Press: Cambridge, UK, 2003.
61. Zhang, M.L.; Li, Y.K.; Liu, X.Y.; Geng, X. Binary relevance for multi-label learning: An overview. *Front. Comput. Sci.* 2018, *12*, 191–202.
62. 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.
63. Ueda N.; Nakano, R. Deterministic annealing EM algorithm, *Neural Networks*, 1998, *11*, 271-282, 1998.
64. C. Lu, Understanding and accelerating EM algorithm's convergence by fair competition principle and rate-verisimilitude function. Available: https://arxiv.org/abs/2104.12592. (accessed on 20 January 2025).
65. Wikipedia, Variational Bayesian methods. Available online: https://en.wikipedia.org/wiki/Variational_Bayesian_methods. (accessed on 22 Dec. 2024).
66. Beal, M. J. Variational algorithms for approximate Bayesian inference. Doctoral thesis (Ph.D), University College London, 2003.
67. Koller, D. Probabilistic Graphical Models: Principles and Techniques. The MIT Press, Cambridge, Massachusetts, USA. 2009.
68. Sebastian Gottwald1,2 and Daniel A. Braun1, The Two Kinds of Free Energy and the Bayesian Revolution, Available online: https://arxiv.org/abs/2004.11763. (accessed on 20 January 2025).
69. Hempel, C.G. Studies in the logic of confirmation. *Mind* 1945, *54*, 1–26.
70. Carnap, R. *Logical Foundations of Probability*, 1st ed.; University of Chicago Press: Chicago, IL, USA, 1950.
71. *Scheffler, I.; Goodman, N.J. Selective confirmation and the ravens: A reply to Foster. J. Philos. 1972, 69, 78–83.*
72. Fitelson, B.; Hawthorne, J. How Bayesian confirmation theory handles the paradox of the ravens. In *the Place of Probability in Science*; Eells, E., Fetzer, J., Eds.; Springer: Dordrecht, German, 2010; pp. 247–276.
73. Crupi, V.; Tentori, K.; Gonzalez, M. On Bayesian measures of evidential support: Theoretical and empirical issues. Philos. Sci. 2007, *74*, 229–252.
74. Greco, S.; Slowi ´nski, R.; Szcz ¸ech, I. Properties of rule interestingness measures and alternative approaches to normalization of measures. Inf. Sci. 2012, *216*, 1–16.
75. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.*, 2009, 3: 96–146.
76. Lu, C. Decoding model of color vision and verifications. *Acta Opt. Sin.* 1989, *9*, 158–163. (In Chinese)
77. Lu, C. B-fuzzy quasi-Boolean algebra and a generalized mutual entropy formula. *Fuzzy Syst. Math.* 1991, *5*, 76–80. (in Chinese).
78. Lu, C. Explaining color evolution, color blindness, and color recognition by the decoding model of color vision. In Proceedings of the 11th IFIP TC 12 International Conference, IIP 2020, Hangzhou, China, 3–6 July 2020; Shi, Z.; Vadera, S.; Chang, E., Eds.; Springer: Cham, Switzerland, 2020; pp. 287–298. Available online: https://www.springer.com/gp/book/9783030469306.
79. Hinton, G. E. Deep belief networks. *Scholarpedia* 2009, 4, 5947. doi:10.4249/scholarpedia.5947
80. RAE J. Compression for AGI, Stanford MLSys Seminar, Available online: https://www.youtube.com/watch?v=dO4TPJkeaaU. (accessed on 18 January 2025).
81. Sutskever L. An observation on Generalization, Bekeley: Simons Institute, Available online: https://simons.berkeley.edu/talks/ilya-sutskever-openai-2023-08-14. (accessed on 18 January 2025).
82. Mark Burgin, *Theory of Information: Fundamentality, Diversity And Unification*, World Science (in Chinese), Knowledge Copyright Publishing, Beijing, China, 2015, ISBN 978-7-5130-3095-3.
83. De Luca, A.; Termini, S. A definition of a non-probabilistic entropy in setting of fuzzy sets. *Inf. Control* 1972, *20*, 301–312.
84. D'Alfonso, S. On quantifying semantic information. *Information* 2011, *2*, 61–101.

85. Basu, P.; Bao, J.; Dean, M.; Hendler, J. Preserving quality of information by using semantic relationships. *Pervasive Mobile Comput.* 2014, *11*, 188–202.
86. Melamed, D. Measuring semantic entropy, 1997, Available online: https://api.semanticscholar.org/CorpusID:7165973. (accessed on 20 January 2025).
87. Guo, T.; Wang, Y. Han, J. et al. Semantic compression with side information: a rate-distortion perspective. Available online: https://arxiv.org/pdf/2208.06094. (accessed on 20 January 2025).
88. Liu, J.; Zhang, W.; Poor, H.V. A rate-distortion framework for characterizing semantic information, in *Proceedings of 2021 IEEE International Symposium on Information Theory* (ISIT). Piscataway: IEEE Press, 2021: 2894-2899.
89. Kumar, T.; Bajaj, R.K.; Gupta, B. On some parametric generalized measures of fuzzy information, directed divergence and information Improvement. *Int. J. Comput. Appl.* 2011, *30*, 5–10.
90. Ohlan, A.; Ohlan, R. Fundamentals of fuzzy information measures. In *Generalizations of Fuzzy Information Measures*; Springer: Cham, Switzerland, 2016. https://doi.org/10.1007/978-3-319-45928-8_1.
91. Klir, G. Generalized information theory. *Fuzzy Sets Syst.* 1991, 40, 127–142.
92. Kolmogorov, A. N., Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1965, 1, 1--7.
93. Wikipedia, Semantic similarity, Available online: https://en.wikipedia.org/wiki/Semantic_similarity (accessed on 20 January 2025).
94. Tao, X. et al. Federated Edge Learning for 6G: Foundations, Methodologies, and Applications, in Proceedings of the IEEE, doi: 10.1109/JPROC.2024.3509739.
95. Letaief, K.B. et al. AI empowered wireless networks. IEEE Commun. Mag. 2019, *57*, 84–90.
96. Gündüz, D. et al. Beyond transmitting bits: Context, semantics, and task-oriented communications. IEEE J. Sel. Areas Commun. 2022, *41*, 5–41.
97. Niu, K; et al. A paradigm shift towards semantic communications. IEEE Communications Magazine, 2022, *60*, 113-119.
98. *Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space.* Available online: arXiv:1301.3781 *(accessed on 20 January 2025)..*
99. *Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems.* Available online: arXiv:1310.4546 *(accessed on 20 January 2025).*
100. Vaswani et al. Attention is all you need. Available online: https://arxiv.org/abs/1706.03762. (accessed on 18 January 2025).