Semantic Channel and Shannon Channel Mutually Match and Iterate for Tests and Estimations with Maximum Mutual Information and Maximum Likelihood

Chenguang Lu

College of Intelligence Engineering and Mathematics, Liaoning Engineering and Technology University,

Fuxin, Liaoning, 123000, China

lcguang@foxmail.com

***Abstract*—It is very difficult to solve Maximum Mutual Information (MMI) or Maximum Likelihood (ML) for all possible Shannon Channels so that we have to use iterative methods. According to the Semantic Information Measure (SIM) and *R*(*G*) function proposed by Chenguang Lu (1993) (where *G* is the lower limit of the SMI, and *R*(*G*) is an extension of rate distortion function *R*(*D*)), we can obtain a new iterative algorithm of solving the MMI and ML for tests, estimations, and mixture models. A group of truth functions constitute a semantic channel. Letting the semantic channel and Shannon channel mutually match and iterate, we can obtain the Shannon channel that maximizes mutual information and average log likelihood. This iterative algorithm is called Channels' Matching (CM) algorithm. The convergence can be intuitively explained and proved by the *R*(*G*) function. Several iterative examples show that the CM algorithm for tests and estimations with larger samples is simple, fast, and reliable. Multi-label Logical classification is introduced in passing.**

Keywords—Shannon channel; semantic channel; semantic information; rate distortion; maximum mutual information; maximum likelihood; machine learning; multi-label classification

I. INTRODUCTION

It is a very important to use the Maximum Mutual Information (MMI) or Maximum Likelihood (ML) as criterion to optimize tests, estimation, predictions, classifications, image compression, and clustering. Yet, the Shannon information theory [1] uses distortion criterion instead of the mutual information criterion to optimize tests and estimations because the optimization for the MMI needs to change the Shannon channel. Still, without fixing the Shannon channel, the mutual information cannot be calculated. In this paper, the ML means the ML for uncertain Shannon's channels or uncertain hypothesis-choosing rules. For this ML and the MMI, we can only use iterative methods.

The relationship between the information measure and likelihood [2] has drawn growing attention in recent decades. Akaike notes [3] that the maximum likelihood criterion is equivalent to the minimum Kullback-Leibller (KL) divergence [4] criterion, which is an important discovery. However, the divergence does not mean conveyed information. Although the log relative likelihood [5] and log normalized likelihood [6] used by some researchers are very similar to information measure; yet we cannot directly put them into the KL formula or Shannon mutual information formula because a sampling distribution is generally different from a likelihood function.

There have been some iterative methods for the MMI and ML, including the Newton method [7], the EM algorithm [8], and the minimax method [9]. Still, we want a different iterative method with higher efficiency and clearer convergence reasons.

In a different way, Chenguang Lu [10,11,12] directly defined the semantic information measure by log normalized likelihood. This measure is called the "semantic information measure" because a likelihood function is produced by the truth function of a hypothesis with the source $P(X)$. Lu also proposed the $R(G)$ function (G is the lower limit of the semantic mutual information) [12], which was an extension of Shannon's (information) rate distortion function $R(D)$ [13]. Now it is found that Lu's semantic information measure and the $R(G)$ function can be used to achieve the MMI and ML more conveniently. The new algorithm is called the Channel's matching algorithm, or the CM algorithm. The CM algorithm for tests, estimations, and mixture models can be demonstrated by Excel files.[1] The CM algorithm for mixture models has been discussed by Lu [14].

[1] Excel files can be downloaded from http://survivor99.com/lcg/CM-iteration.zip

In this paper, we first restates Lu's semantic channel, semantic information measure, and $R(G)$ function in a new way that is as compatible with the Likelihood Method (LM) as possible. We then discuss the optimization of truth functions or semantic channels, including the optimization of multi-label logical classifications. Then, we introduce the CM algorithm for tests and estimations with some examples to show its efficiency and reliability.

II. Semantic Channel and Semantic Bayesian Prediction

A semantic channel is supported and affected by a Shannon channel. First, we discuss the Shannon channel.

A. Shannon's Channel and the Transition Probability Function

Let X be a discrete random variable representing an instance with alphabet $A=\{x_1, x_2, \ldots, x_m\}$, let Y be a discrete random variable representing a message with alphabet $B=\{y_1, y_2, \ldots, y_n\}$, and let Z be a discrete random variable representing a observed condition with alphabet $C=\{z_1, z_2, \ldots, z_w\}$. A message sender chooses Y to predict X according to Z. For example, in weather forecasts, X is a rainfall, Y is a forecast such as "There will be light to moderate rain tomorrow", and Z is a group of meteorological data.

We use $P(X)$ to denote the probability distribution of X and call $P(X)$ the source, and we use $P(Y)$ to denote the probability distribution of Y and call $P(Y)$ the destination. We call $P(y_j|X)$ with certain y_j and variable X a transition probability function from X to y_j. Then the Shannon's channel is composed of a group of transition probability functions [1]:

$$P(Y|X) \Leftrightarrow \begin{bmatrix} P(y_1|x_1) & P(y_1|x_2) & \ldots & P(y_1|x_m) \\ P(y_2|x_1) & P(y_2|x_2) & \ldots & P(y_2|x_m) \\ \ldots & \ldots & \ldots & \ldots \\ P(y_n|x_1) & P(y_n|x_2) & \ldots & P(y_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} P(y_j|X) \\ P(y_j|X) \\ \ldots \\ P(y_n|X) \end{bmatrix}$$

where a bidirectional arrow means equivalence. The transition probability function has two properties:

1) $P(y_j|X)$ is different from the conditional probability function $P(Y|x_i)$ or $P(X|y_j)$ in that whereas the latter is normalized, the former is not. In general, $\sum_i P(y_j|x_i) \neq 1$.

2) $P(y_j|X)$ can be used to make Bayesian prediction to get the posterior probability distribution $P(X|y_j)$ of X. To use it by a coefficient, the two predictions are equivalent, i. e.

$$\frac{P(X)kP(y_j|X)}{\sum_i P(x_i)kP(y_j|x_i)} = \frac{P(X)P(y_j|X)}{\sum_i P(x_i)P(y_j|x_i)} = P(X|y_j)$$

B. Semantic Channel and Semantic Bayesian Prediction

In terms of hypothesis-testing, X is a piece of evidence and Y is a hypothesis or a prediction. We need a sample sequence $\{x(t)|t=1.\ 2, \ldots, N\}$ or a sampling distribution $P(X|.)$ to test a hypothesis to see how accurate it is.

Let Θ be a random variable for a predictive model, and let θ_j be a value taken by Θ when $Y=y_j$. The semantic meaning of a predicate $y_j(X)$ is defined by θ_j or its (fuzzy) truth function $T(\theta_j|X) \in [0,1]$. Because $T(\theta_j|X)$ may be constructed with some parameters, we may also treat θ_j as a set of model parameters. We may also consider that $T(\theta_j|X)$ is defined by a normalized likelihood, i. e., $T(\theta_j|X) = k\,P(X|\theta_j)/P(X)$, where k is a coefficient that makes the maximum of $T(\theta_j|X)$ be 1. If $T(\theta_j|X) \in \{0,1\}$, $T(\theta_j|X)$ will be the feature function of a set, whose every element makes y_j true. Therefore, θ_j can also be regarded as a fuzzy set, and $T(\theta_j|X)$ can be regarded as a membership function of a fuzzy set defined by Zadeh [15].

In contrast to the popular likelihood method, the above method uses sub-models θ_1 , θ_2, …, θ_n instead of one model θ or Θ. A sub-model θ_j is separated from a likelihood function $P(X|\theta_j)$ and defined by a truth function $T(\theta_j|X)$. The $P(X|\theta_j)$ here is equivalent to $P(X|y_j, \theta)$ in the popular likelihood method. A sample used to test y_j is also a sub-sample or conditional sample. These changes will make the new method more flexible and more compatible with the Shannon information theory.

When $X=x_i$, $y_j(X)$ becomes $y_j(x_i)$, which is a proposition with truth value $T(\theta_j|x_i)$. Then there is the semantic channel:

$$T(\theta|X) \Leftrightarrow \begin{bmatrix} T(\theta_1|x_1) & T(\theta_1|x_2) & \ldots & T(\theta_1|x_m) \\ T(\theta_2|x_1) & T(\theta_2|x_2) & \ldots & T(\theta_2|x_m) \\ \ldots & \ldots & \ldots & \ldots \\ T(\theta_n|x_1) & T(\theta_n|x_2) & \ldots & T(\theta_n|x_m) \end{bmatrix} \Leftrightarrow \begin{bmatrix} T(\theta_1|X) \\ T(\theta_2|X) \\ \ldots \\ T(\theta_n|X) \end{bmatrix}$$

The truth function is also not normalized，and its maximum is 1. Like $P(y_j|X)$, $T(\theta_j|X)$ can also be used for Bayesian prediction, which is called semantic Bayesian prediction, to produce likelihood function:

$$\begin{aligned} P(X|\theta_j) &= P(X)T(\theta_j|X)/T(\theta_j) \\ T(\theta_j) &= \sum_i P(x_i)T(\theta_j|x_i) \end{aligned} \quad (1)$$

where $T(\theta_j)$ is called the logical probability of y_j. Lu called it the set-Bayes' formula in his earlier papers [10] and used it for semantic information measure. This formula can be found in Thomas' paper published in 1979 [16].

We may also write $T(\theta_j)$ as $T(y_j)$. If $T(\theta_j|X) \propto P(y_j|X)$, then the semantic Bayesian prediction is equivalent to Bayesian prediction. Note that $T(\theta_j)$ is the logical probability of y_j, whereas $P(y_j)$ is the probability of choosing y_j. The two probabilities are very different. For example, when y_j is a tautology, $T(\theta_j)=1$; but $P(y_j)$ is almost 0.

A semantic channel is always supported by a Shannon channel. For weather forecasts, the transition probability

function $P(y_j|X)$ indicates the rule of choosing a forecast y_j. The rules used by different forecasters may be different and have more or fewer mistakes. Whereas, $T(\theta_j|X)$ indicates the semantic meaning of y_j that is understood by audience. The semantic meaning is generally publicly defined and may also come from (or be affected by) the past rule of choosing y_j. To different people, the semantic meaning should be similar.

III. Semantic Information Measure and the Optimization of Semantic Channels

We introduce Lu's semantic information measure and the optimization of semantic channels in relation to likelihood method.

A. Defining Semantic Information with Normalized Likelihood

In Shannon's information formulas, there is only the statistical probability, without the logical probability or likelihood (predicted probability). However, in Lu's semantic information formulas, there are three types of probabilities. The semantic information (measure) conveyed by y_j about x_i is defined as log normalized likelihood [10]:

$$I(x_i;\theta_j)=\log\frac{P(x_i\mid\theta_j)}{P(x_i)}=\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{2}$$

where the semantic Bayesian prediction is used; prior likelihood is assumed to be equal to prior probability. For an unbiased estimation, its truth function and semantic information are illustrated in Fig. 1.

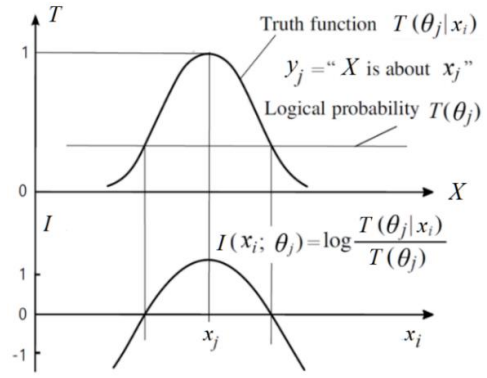


Fig. 1. Semantic information measure is defined by log normalized likelihood. The larger the deviation is, the less information there is; the less the logical probability is, the more information there is; and, a wrong estimation may convey negative information.

This semantic information measure is compatible with Popper's thought [12]. For example, Popper affirms that if a hypothesis can survive tests, then the less its logical probabiity is, the more information it conveys [17].

After averaging $I(x_i;\theta_j)$, we obtain semantic (or generalized) KL information:

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i\mid y_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3}$$

The statistical probability $P(x_i|y_j)$, $i=1, 2, \ldots$, on the left of "log", represents a sampling distribution (note that a sample or sub-sample is also conditional) to test the hypothesis y_j or model θ_j. If $y_j=f(Z|Z\epsilon C_j)$, then $P(X|y_j)=P(X|Z\epsilon C_j)=P(X|C_j)$.

Assume that the size of the sample used to test y_j is N_j; the sample points come from independent and identically distributed random variables. Among N_j sample points, the number of x_i is N_{ij}. When N_j is infinite, $P(X|y_j)= N_{ij}/N_j$. Hence there is log normalized likelihood:

$$\begin{aligned}\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}&=N_j\sum_i P(e_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}\\&=N_j I(X;\theta_j)\end{aligned} \tag{4}$$

The semantic information measure increases with likelihood, whereas the KL divergence does not [5]. After averaging the log normalized likelihood for different y_j, we have

$$\begin{aligned}\sum_j\frac{N_j}{N}\log\prod_i\left[\frac{P(x_i|\theta_j)}{P(x_i)}\right]^{N_{ji}}&=\sum_j P(y_j)\sum_i P(x_i\mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}\\&=I(X;\Theta)=H(X)-H(X\mid\Theta)\end{aligned} \tag{5}$$

where $I(X;\Theta)$ is equal to the semantic mutual information [12], $P(y_j)=N_j/N$, and $N=N_1+N_2+\cdots+N_n$. It shows that the ML criterion is equivalent to the minimum generalized posterior entropy criterion or Maximum Semantic Information (MSI) criterion. It is easy to find that when $P(X|\theta_j)=P(X|y_j)$(for all j), the semantic mutual information $I(X;\Theta)$ will be equal to the Shannon mutual information $I(X;Y)$. Thus, $I(X;Y)$ is the special case of $I(X;\Theta)$.

B. The Optimization of Semantic Channels

Optimizing a predictive model Θ is equivalent to optimizing a semantic Channel or a group of truth functions. For given y_j, optimizing θ_j is equivalent to optimizing $T(\theta_j|X)$ by

$$T^*(\theta_j\mid X)=\arg\max_{T(\theta_j|X)} I(X;\theta_j) \tag{6}$$

$I(X;\theta_j)$ can be written as the difference of two KL divergences:

$$I(X;\theta_j)=\sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i)}-\sum_i P(x_i\mid y_j)\log\frac{P(x_i|y_j)}{P(x_i\mid\theta_j)} \tag{7}$$

Because the KL divergence is greater than or equal to 0, when

$$P(X|\theta_j)=P(X|y_j) \tag{8}$$

$I(X; \theta_j)$ reaches its maximum and is equal to the KL information $I(X; y_j)$. Let the two sides of Eq. (8) be divided by $P(X)$; then there are

$$\frac{T(\theta_j \mid X)}{T(\theta_j)} = \frac{P(y_j \mid X)}{P(y_j)} \text{ and } T(\theta_j|X) \propto P(\theta_j|X) \quad (9)$$

Set the maximum of $T(\theta_j|X)$ to 1. Then we obtain [16]

$$T^*(\theta_j|X)=P(y_j|X)/P(y_j|x_j^*) \quad (10)$$

where x_j^* is the x_i that makes $P(y_j|x_j^*)$ be the maximum of $P(y_j|X)$. Generally, it is not easy to get the $P(y_j|X)$; yet for given $P(X|y_j)$ and $P(X)$, it is easier to get $T(\theta_j|X)$ than to get $P(y_j|X)$ since

$$T^*(\theta_j|X)=[P(X|y_j)/P(X)]/[P(x_j^*|y_j)/P(x_j^*)] \quad (11)$$

where x_j^* is such an x_i that makes $P(x_j^*|y_j)/P(x_j^*)$ be the maximum of $P(X|y_j)/P(X)$. Using this optimized truth function, the semantic Bayesian prediction will be compatible with traditional Bayesian prediction.

In Eq. (5), when $P(Y|X)$ is fixed, we change $T(X|\Theta)$ so that $I(X; \Theta)$ reaches its maximum $I(X; Y)$, which means "the semantic channel matches the Shannon channel". However, conversely, when $T(X|\Theta)$ is fixed, $P(Y|X) \propto T(X|\Theta)$ does not maximize $I(X; \Theta)$. For given $T(X|\Theta)$, there may be other Shannon channels with less noise conveying more semantic information. This matching will be discussed later.

Similar to the Maximum-A-Posterior (MAP) estimation, the MSI estimation also uses the prior. The difference is that the MAP uses the prior of Y or θ, whereas the MSI uses the prior of X. The MSI is more compatible with Bayesian prediction. The Eq. (6) fits parameter estimations, and the Eqs. (10) and (11) fit non-parameter estimations with larger samples.

C. Semantic Channels for Multi-label Logical Classifications and Single-label Selective Classifications

Using truth functions or semantic channels, we can obtain a new machine learning method or classification method.

For example, we classify people with different ages X into fuzzy sets {juveniles}, {youths}, {middle-aged people}, {adults}, …, {old people}, or classify different weathers denoted by X, which means 12 hour rainfall and has negative value that means light intensity, into {sun shine}, {cloud}, {no rain}, {light rain}, {moderate rain}, {heavy rain}, {moderate to heavy rain}, …, {storm}. Among hypotheses involved, one may imply another. For examples, "X is middle-aged" implies "X is adult"; "X is moderate rain" or "X is heavy rain" implies "X is moderate to heavy rain".

For hypotheses y_1="X is juvenile" and y_n="X is old", we may use logistic functions

$$T(\theta_1 \mid X) = \frac{\exp[-k_1(X-a_1)]}{1+\exp[-k_1(X-a_1)]} \quad (12)$$

$$T(\theta_n \mid X) = \frac{1}{1+\exp[-k_n(X-a_n)]} \quad (13)$$

as their truth functions. For hypotheses y_2, …, y_{n-1} about ages, we may use

$$T(\theta_j \mid X) = \exp[-(X-c_j)^2/(2d_j)^2], j=2, 3, \ldots, n-1 \quad (14)$$

or more complicated functions as their truth functions.

For given examples ($x(t)$, $y(t)$), t=1,…, N, if t is big enough, we may obtain $P(x_i, y_j)$ and $P(x_i|y_j)$, i=1,2, …, m; j=1, 2, …, n. Then, we may use Eqs. (6) or (10) to optimize these truth functions, which can also be called logical classification functions. If there is always $T(\theta_j|X) \geqslant T(\theta_k|X)$, then y_k implies y_j.

The above method is suitable to cases where $P(X)$ is variable. If we do not know $P(X)$, we may assume that $P(X)$ is constant. In this case, $T(\theta_j|X)$ is proportional to $P(X|\theta_j)$ and $T(\theta_j)$ is proportional to the area that the curve $T(\theta_j|X)$ covers.

After the truth functions are optimized, for given $X=x_i$, we choose a y_j as the decision or message according to which θ_j maximizes $I(x_i; \theta_j)$. In this case, Y is the function $f(X)$ of X. The $f(X)$ ascertains a classification of set A, which may be called single-label selective classification. If X is uncertain and with probability distribution $P(X|y_?)$, we may choose a y_j as the decision or prediction according to which θ_j maximizes $I(X; \theta_j)$.

Different from popular classification methods [18], above classification method has features:

1) It uses normalized likelihood criterion and hence the class-imbalance has been considered already;
2) It distinguishes classifications into multi-label logical classifications and single-label selective classifications, and allows that the logical classification is independent of the source, and demands that the selective classification makes use of variable source for the sake of ML or MSI;

For multi-label classification, we may use the first-order strategy [19] which decomposes each multi-label example into several single-label examples. Then we can obtain $P(X, Y)$ and logical classification functions. For given $X=x_i$, only those hypotheses with less logical probability may be chosen; however, for uncertain X with probability distribution $P(X|y_?)$, a fuzzy hypothesis, such as "There will be moderate to heavy rain tomorrow" may be chosen. $I(x_i; \theta_j)$ and $I(X; \theta_j)$ can also be used to rank a group of labels for given $X=x_i$ or $P(X|y_?)$.

IV. The Matching Function $R(G)$ of Shannon Information and Semantic Information

The $R(G)$ function is an extension of the (information) rate distortion function $R(D)$. The $R(G)$ function was used for image

compression according to visual discrimination [12]. Now it can be used to explain the CM algorithm.

A. From the R(D) Function to the R(G) Function

In the $R(D)$ function, R is the information rate, D is the upper limit of the distortion. $R(D)$ means that for given D, $R=R(D)$ is the minimum of the Shannon mutual information $I(X; Y)$. The rate distortion function with parameter s [20] is

$$\begin{aligned} D(s) &= \sum_i \sum_j d_{ij} P(x_i) P(y_j) \exp(s d_{ij}) / \lambda_i \\ R(s) &= sD(s) - \sum_i P(x_i) \ln \lambda_i \end{aligned} \tag{15}$$

where $\lambda_i = \sum_j P(y_j)\exp(s d_{ij})$ is the partition function.

Let d_{ij} be replaced with $I_{ij} = I(x_i; y_j) = \log[T(\theta_j|x_i)/T(\theta_j)]$, and let G be the lower limit of the semantic mutual information $I(X; \Theta)$. The $R(G)$ function for a given source $P(X)$ is defined as

$$R(G) = \min_{P(Y|X); I(E;\Theta) \geq G} I(X;Y) \tag{16}$$

Following the derivation of $R(D)$ [20], we can obtain [12]

$$\begin{aligned} G(s) &= \sum_i \sum_j I_{ij} P(x_i) P(y_j) 2^{sI_{ij}}) / \lambda_i = \sum_i \sum_j I_{ij} P(x_i) P(y_j) m_{ij}^s / \lambda_i \\ R(s) &= sG(s) - \sum_i P(x_i) \ln \lambda_i \end{aligned} \tag{17}$$

where $m_{ij} = P(x_i|\theta_j)/P(x_i)$ is the normalized likelihood; $\lambda_i = \sum_j P(y_i) m_{ij}^s$. We may also use $m_{ij} = P(x_i|\ \theta_j)$, which results in the same m_{ij}^s/λ_i. The shape of any $R(G)$ function is a bowl-like curve as shown in Fig. 2.

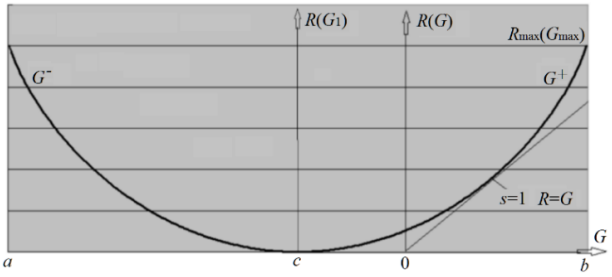


Fig. 2. The $R(G)$ function of a binary source (for details, see [12]). As s=1, R=G, which implies that the semantic channel matches the Shannon channel; $R_{max}(G_{max})$ at the top-right corner means that the Shannon channel matches the semantic channel so that both R and G are at their maxima.

The R(G) function is different from the $R(D)$ function. For a given R, we have the maximum value G^+ and the minimum value G^-, which is negative and means that to bring a certain information loss to enemies, we also need certain objective information R. When R=0, G is negative, which means that if we listen to someone who randomly predicts, the information that we already have will be reduced.

In rate distortion theory, $dR/dD = s$ ($s \leq 0$). It is easy to prove that there is also $dR/dG = s$, where s may be less or greater than 0. The increase of s means the increase of predictive precision or the decrease of noise.

If s changes from positive s_1 to $-s_1$, then $R(-s_1)=R(s_1)$ and G changes from G^+ to G^- (see Fig. 2).

When s=1, λ_i=1 and R=G, which means that the semantic channel matches the Shannon channel and the semantic mutual information is equal to the Shannon mutual information. When s=0, R=0 and $G(s=0)<0$. In Fig. 2, $c = G(s=0)$.

In fact, in the rate distortion theory, if a larger error probability is allowed, the shape of the function $R(D)$ is also a bowl-like curve. We can also use a bowl-like $R(D)$ function to optimize the camouflaged messages to puzzle enemies.

B. Viewing the Maximum Likelihood Ratio Tests from the R(G) Function

For a medical test (see Fig. 3), $A=\{x_0,\ x_1\}$ where x_0 means an uninfected person and x_1 means an infected person, and $B=\{y_0,\ y_1\}$ where y_0 means test-negative and y_1 means test-positive.

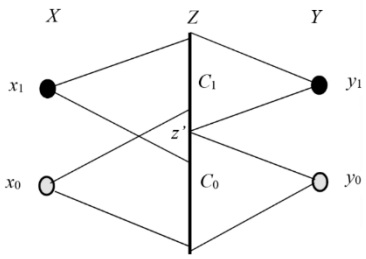


Fig. 3. Illustrating the medical test. The test can be abstracted as a 2×2 Shannon nosy channel. The Shannon mutual information changes with dividing point z'.

In medical tests, for an infected testee x_1, the conditional probability of a test-positive $P(y_1|x_1)$ is called sensitivity. For an uninfected testee x_0, the conditional probability of a test-negative $P(y_0|x_0)$ is called specificity [21]. The sensitivity and specificity form a Shannon channel as shown in Table I.

TABLE I. THE SENSITIVITY AND SPECIFICITY OF MEDICAL TESTS FORM A SHANNON'S CHANNEL $P(Y|X)$

Y	**Infected x_1**	**Uninfected x_0**
Positive y_1	$P(y_1\|x_1)$=sensitivity	$P(y_1\|x_0)$=1-specificity
Negative y_0	$P(y_0\|x_1)$=1-sensitivity	$P(y_0\|x_0)$=specificity

If we absolutely believe that a test-positive means being infected, and a test-negative means not being infected, then there are truth values $T(y_1|x_1)=T(y_0|x_0)=1$, $T(y_1|e_0)=T(y_0|x_1)=0$. If we

use these truth values as the semantic channel, the information will be negatively infinite when one counterexample exists. Thus, we need to consider the confidence levels of y_j. Let the confidence level of y_j be denoted by b, and let the no-confidence level be denoted by b'=1-|b|. Then the truth function of y_j may be defined as

$$T(\theta_j|X)= b' +bT(y_j|X) \quad (18)$$

Here, b' is also the truth value of a counterexample or the degree of falsification of predicate $y_j(X)$.

Assume that the no-confidence level of y_1 and y_0 are b_1' and b_0' respectively; the significance level of a medical test is α. Then α means that there should be $b_0' \leqslant \alpha$. Table II shows the semantic channel for medical tests.

TABLE II. TWO NO-CONFIDENCE LEVELS OF A MEDICAL TEST FORM A SEMANTIC CHANNEL $T(\Theta|X)$

Y	Infected x_1	Uninfected x_0
Positive y_1	$T(\theta_1\|x_1)=1$	$T(\theta_1\|x_0)=b_1'$
Negative y_0	$T(\theta_0\|x_1)=b_0'$	$T(\theta_0\|x_0)=1$

According to Eq. (10), two optimized no-confidence levels are

$$b_1'^*= P(y_1|x_0)/P(y_1|x_1); \quad b_0'^*=P(y_0|x_1)/P(y_0|x_0) \quad (19)$$

In the medical community, Likelihood Ratio is used to indicate how good a test is [21]. Eq. (19) based on the MSI test is compatible with popular Likelihood Ratio (LR) test. There are

$$LR^+ =P(y_1|x_1)/P(y_1|x_0)=1/b_1'^*; \ LR^- =P(y_0|x_0)/P(y_0|x_1)=1/b_0'^* \quad (20)$$

The LR has been used by Thornbury et al for Bayesian prediction [21]. However, it is easier to use the no-confidence level for semantic Bayesian prediction. For example, y_1=HIV-positive, $b_1'^*$=0.0011. If the testees come from ordinary people with $P(x_1)$=0.002, then according to the semantic Bayesian formula,

$$P(x_1|\theta_1)=0.002/(0.002+0.0011*0.998)=0.65.$$

If the testees are gay men with $P(x_1)$=0.1, then

$$P(x_1|\theta_1)=0.1/(0.1+0.0011*0.99)=0.991.$$

Consider the likelihood ratio of tests without a certain partition on C. The likelihood ratio is

$$r_L = \left[\prod_{i=0}^{1}\left(\frac{P(x_i\mid\theta_1)}{P(x_i\mid\theta_0)}\right)^{P(x_i|C_1)}\right]^{NP(C_1)} \left[\prod_{i=0}^{1}\left(\frac{P(x_i\mid\theta_0)}{P(x_i\mid\theta_1)}\right)^{P(x_i|C_0)}\right]^{NP(C_0)} \quad (21)$$

According to Eqs. (5) and (17), $\max(\log r_L)=\max(-NH(X|\Theta))-\min(-NH(X|\Theta))=N(G^+-G^-)$ (see Fig. 2). After R and G^+ are ascertained, s and G^- are also ascertained. Therefore, the maximum likelihood ratio criterion is equivalent to the ML criterion or the MSI criterion.

A binary Shannon channel may be noiseless so that the maximum of R is $R_{max}=H(X)$. Yet, for the test shown in Fig. 3, noise is inevitable, and hence the $P(Y|X)$ for $R_{max}<H(X)$ is not easy to find. As a result, we need an iterative method.

V. THE CM ALGORITHM FOR TESTS AND ESTIMATIONS

This section uses the $R(G)$ function to explain the iterative convergence of the CM algorithm, and provide some examples to show iterative processes，speeds, and reliability.

A. Explaining Channels' Matching and Iterative Convergence by R(G) Function

Matching I (The semantic channel matches the Shannon channel): We keep the Shannon channel $P(Y|X)$ constant, and optimize the semantic channel $T(\Theta|X)$ so that $P(X|\theta_j)=P(X|y_j)$ or $T(\theta_j|X)\propto P(y_j|X)$，and hence $I(X;\Theta)$ reaches its maximum $I(X;Y)$. See Fig. 4 for details.

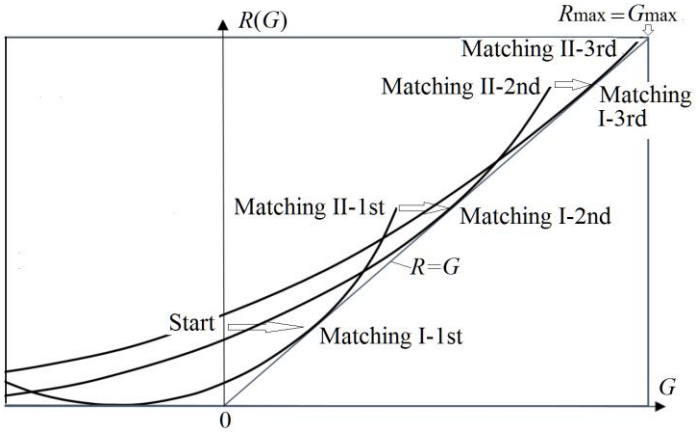


Fig. 4. Illustrating the iterative convergence for tests and estimations. The matching I is for G=R. The matching II is to increase R to the top-right corner of a $R(G)$ function. Repeating the matching I and matching II can maximize R and G to obtain R_{max} and G_{max}.

2) Matching II (The Shannon channel matches the semantic channel): While keeping the semantic channel $T(\Theta|X)$ constant, we change the Shannon channel $P(Y|X)$ to maximize the semantic mutual information $I(X;\Theta)$. The $R(G)$ function reminds us that R and G can be raised by increasing the parameter s. After Matching II，(G, R) locates the top-right corner of a $R(G)$ function curve in Fig. 4.

3) Matching III (The two channels mutually match and iterate): The iterative process is shown in Fig. 4.

B. Iterative Process for Tests

We use some examples to show the iterative process. For the test as shown in Fig. 3, optimizing the Shannon channel is

equivalent to optimizing the dividing point z'. When $Z>z$', we choose y_1; otherwise, we choose y_0.

As an example of the test, $Z \in C=\{1, 2, \ldots, 100\}$ and $P(Z|X)$ is a Gaussian distribution function:

$$P(Z|x_1)=K_1\exp[-(Z-c_1)^2/(2d_1^2)], \quad P(Z|x_0)=K_0\exp[-(Z-c_0)^2/(2d_0^2)]$$

where K_1 and K_0 are normalizing constants. After setting the starting z', say z'=50, as the input of the iteration, we perform the iteration as follows.

The Matching I: Calculate the following items in turn:

- The transition probabilities for the Shannon channel:

$$P(y_0 \mid x_0)=\sum_{z_k=1}^{z'} P(z_k \mid x_0), \qquad P(y_1 \mid x_0)=1-P(y_0 \mid x_0)$$

$$P(y_1 \mid x_1)=\sum_{z_k=z'+1}^{100} P(z_k \mid x_1), \qquad P(y_0 \mid x_1)=1-P(y_1 \mid x_1)$$

- The no-confidence levels b_1'* and b_0'*; the logical probabilities $T(\theta_1)=P(x_1)+b_1'^*P(x_0)$ and $T(\theta_0)=P(x_0)+b_0'^*P(x_1)$.

- Information: $I_{ij}=I(x_i; \theta_j)$ for i=0, 1 and j=0, 1;

- The average semantic information: $I(X;\theta_1|Z)$ and $I(X;\theta_0|Z)$ for given Z (displaying as two curves):

$$I(X;\theta_j|z_k)=\sum_i P(x_i \mid z_k) I_{ij}, \quad k=1, 2, \ldots, 100; j=0, 1 \qquad (22)$$

The matchning II: Compare two information function curves $I(X; \theta_1|Z)$ and $I(X; \theta_0|Z)$ over Z to find their cross point. Use the z_k of this point as new z'. If the new z' is the same as the last z' then let $z^*=z$' (where z^* is the optimized dividing point) and quit the iteration; otherwise go to Matching I.

We may also use the following partitioning function

$$P(y_j \mid Z)=\lim_{s\to\infty} \frac{P(y_j)[\exp(I(X;\theta_j \mid Z))]^s}{\sum_{j'} P(y_{j'})[\exp(I(X;\theta_{j'} \mid Z))]^s}, \quad j=1, 2, \ldots, n \qquad (23)$$

which tells us the optimal dividing point z^*. Even if Z is multi-dimensional, the above equation is also tenable. This formula can save our energy for searching the boundaries of C_j for all j.

C. Three Iterative Examples for Tests and Estimations

The following are three computing examples. In Example 2 and Example 3, there are two dividing points z_1' and z_2'. The iterative principle is the same.

Experiment Report 1 (for a 2×2 Shannon Channel)

Input data: $P(x_0)$=0.8; c_0=30, c_1=70; d_0=15, d_1=10. The start point z'=50.

The iterative process: Matching II-1 gets z'=53; Matching II-2 gets z'=54; Matching II-3 gets z^*=54.

Comparison: To see information loss, we get $H(X)$=0.72 bit; $I(X; Z)$=0.55 bit; and $I(X; \Theta)= I(X; Y)=\sum_k P(z_k)I(X; Y|z_k)$=0.47 bit.

Analysis: If we use minimum error rate as criterion, the optimal dividing point is 57; yet the above optimal dividing point is z^*=54. It is shown that in comparison with minimum error rate criterion, the MSI criterion puts more attention to the correct predictions of small probability events and allow more false positives and less false negatives.

Experiment Report 2 (for a 2×3 Shannon channel)

For this channel, if $z_1'<Z \le z_2'$, $Y=y_2$ means "The test tells nothing".

Input data: $P(x_0)$=0.8; c_0=30, c_1=70; d_0=15, d_1=10. The start point z'_1=50 and z_2'=60.

The iterative progress: Matching II-1 gets z_1'=46 and z_2'=57; Matching II-2 gets z_1'=47 and z_2'=59; Matching II-3 gets z_1*=47 and z_2*=59.

Comparison and analysis: $H(X)$=0.72 bit; $I(X; Z)$=0.55 bit; $I(X; \Theta)$=0.52 bit. Yet in Example 1, $I(X; \Theta)$=0.47 bit. So, This 2×3 channel can convey more semantic information than the above 2×2 channel. This example shows that we may use hypothesis y_2 instead of the significance level α.

Experiment Report 3 (for a 3×3 Shannon channel)

This experiment is to examine a simplified estimation. A pair of good start points and a pair of bad start points are used to check the convergence and speed of the iteration.

Input data: $P(x_0)$=0.5, $P(x_1)$=0.35, and $P(x_2)$=0.15; c_0=20, c_1=50, and c_2=80; d_0=15, d_1=10, and d_2=10.

The iterative results:

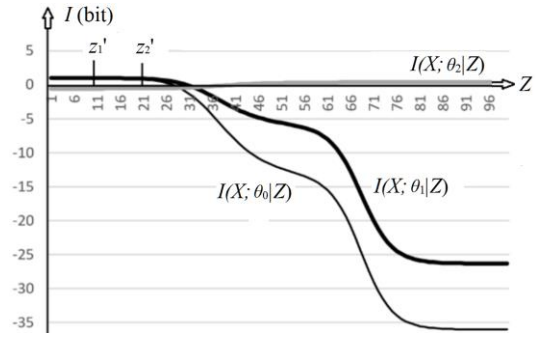
a) *With the good start points*: z_1'=50 and z_2'=60, the number of iterations is 4; z_1*=35 and z_2*=66.

b) *With the bad start points*: z_1'=9 and z_2'=20, the number of iterations is 11; z_1*=35 and z_2*=66 also. Fig. 5 shows the information curves over Z before and after the iteration.

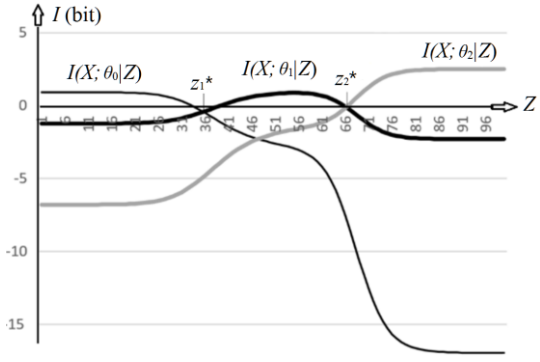
The above estimation does not use parameters and fits larger samples. If m is larger and the sample is smaller, we need parameter estimations. The truth functions may be $T(\theta_j|X)=\exp[-(X-x_j)^2/(2d^2)]$. The CM algorithm also works for parameter estimations for which Eq. (6) is needed.

D. The Comparison of Speeds

In the examples of tests and estimations we used, the most numbers of iterations for convergence are between 3 to 5. We have compared the CM algorithm with the EM algorithm and Newton method for mixture models and found that the CM was clearly faster [14]. The CM algorithm for tests and estimations is much simpler than the CM algorithm for mixture models. It may be expected that The CM algorithm for tests and estimations is also clearly faster than the popular methods.



(a)



(b)

Fig. 2. The iteration with bad start poins. At the beginning of the iteration (a), three information curves cover very small positive areas. At the end of the iteration (b), three information curves cover much larger positive areas so that $I(X;\Theta)$ reaches its maximum.

E. Explaining the Evolution of Natural Language

We may apply the CM algorithm to communications with natural language, such as weather forecasts. Then we can explain the evolution of natural language. A Shannon channel indicates a language usage, whereas a semantic channel indicates the comprehension of the audience. The Matching I is to let the comprehension match the usage, and the Matching II is to let the usage (including the observations and discoveries) match the comprehension. The mutual matching and iterating of two channels means that linguistic usage and comprehension mutually promote. Natural languages should have been evolving in this way.

VI. CONCLUSIONS

This paper restates Lu's semantic information method and reveals that by letting the semantic channel and Shannon channel mutually match and iterate, we can achieve the maximum mutual information and maximum average log-likelihood for tests and estimations. The iterative convergence can be intuitively explained and proved by Lu's $R(G)$ function. Several iterative examples and theoretical analyses show that the CM algorithm for tests and estimations is simple, fast, and reliable. The paper also reveals that the tight combination of the Shannon information theory with the likelihood method and the fuzzy sets theory is necessary and feasible.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal. vol. 27, pp. 379–429 and 623–656, 1948.

[2] R. A. Fisher, "On the mathematical foundations of theoretical statistics," Philo. Trans. Roy. Soc., Lodon, vol. A 222, pp. 309-368, 1922.

[3] H. Akaike, "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, vol. 19, pp. 716–723, 1974.

[4] S. Kullback and R. Leibler, "On information and sufficiency," Annals of Mathematical Statistics. vol. 22, pp.79–86, 1951.

[5] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd Edition, New York: Wiley & Sons, 2006, pp. 375-392.

[6] A. Barron, T, Roos, and K. Watanabe, "Bayesian Properties of Normalized Maximum Likelihood and its Fast Computation, " 2014 IEEE IT Symposium.

[7] M. Kok, J. Dahlin, T. B. Schon, and A. Wills, "Newton-based maximum likelihood estimation in nonlinear state space models", IFAC-PapersOnLine, vol. 48: pp. 398–403, 2015.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B. vol. 39 (1): pp. 1–38. 1977.

[9] Q. Xie, and A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," IEEE Trans. on Information Theory, vol. 46, pp. 431–445, 2000.

[10] C. Lu, A Generalized Information Theory (in Chinese), Hefei: China Science and Technology University Press, 1993

[11] C. Lu, "Coding Meanings of Generalized Entropy and Generalized Mutual Information," J. of China Institute of Communication (in Chinese), vol. 15, pp. 37-44,1994.

[12] C. Lu, "A generalization of Shannon's information theory", Int. J. of General Systems, vol. 28 (6), pp. 453-490, 1999.

[13] C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, IRE Nat. Conv. Rec., Part 4:142–163, 1959.

[14] C. Lu, "Channels' matching algorithm for mixture models," in Intelligence Science I, Z.-Z. Shi, B. Goertzel, and J. Feng, Eds. Switzerland: Springer, 2017, pp. 321-332 [Proceedings of Second IFIP TC 12 international Conference, ICIS 2017, Shanghai, October 25-28, 2017].

[15] L. A. Zadeh, "Fuzzy Sets." Information and Control. vol.8, pp. 338–53, 1965.

[16] C. F. Thomas, " A theory of semantics and possible inference with application to decision analysis", PhD Thesis, University of Toronto, Canada, 1979.

[17] K. Popper, Conjectures and Refutations. Reprinted (2005). London and New York: Routledge. 1963, p. 294.

[18] Z.-H. Zhou, Machine Learning, Beijing: Tsinghua University Press, 2016.

[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithm." IEEE Transactions on Knowledge and Data Engineering, vol. 26(8): pp. 1819-1837, 2014.

[20] J. P. Zhou, Fundamentals of Information Theory (in Chinese), Beijing, People's Posts & Telecom Press. 1983.

[21] J. R. Thornbury, D. G. Fryback, and W. Edwards, "Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information," Radiology. vol. 114: 561–5, 1975.