语义变分贝叶斯
——基于语义信息论的求隐含变量方法

Semantic Variational Bayes Based on a Semantic Information Theory for Solving Latent Variables

鲁晨光

主页：http://survivor99.com

摘要：变分贝叶斯求解隐含变量(的概率分布)使用最小自由能准则，这个准则不容易被理解，变分计算也很复杂。为此，本文提出语义变分贝叶斯(SVB)。笔者以前提出的语义信息论把信息率失真函数 $R(D)$推广为信息率逼真函数 $R(G)$——它是给定语义互信息 G 的最小香农互信息。SVB 来自 $R(G)$的参数解，其中变分和迭代方法起源于香农等人对信息率失真函数的研究. SVB 使用的约束函数不限于似然函数，也可以是真值、隶属、相似、或失真函数。不同的是 SVB 使用最大信息效率(G/R)准则——在优化模型参数时使用最大语义信息准则(它兼容最大似然准则)，而在优化香农信道时使用最小互信息(或最大熵)准则。对于同样任务，SVB 的计算比 VB 更简单。文中计算实验包括：举例说明混合模型随 G/R 增大(而不是随负的自由能)增大而收敛；用给定误差范围的数据压缩说明 SVB 的应用；通过给定范围约束的控制任务说明语义信息测度和 SVB 如何用于最大熵控制和强化学习，并为权衡控制的合目的性和效率提供数字依据。要把 SVB 用于神经网络和深度学习，我们需要进一步研究。

关键词：语义信息论，变分贝叶斯，隐含变量，信息率失真，信息率逼真，混合模型，最大熵控制

1. 引言

机器学习经常需要从观察数据(或实例先验分布 $P(x)$)和预测模型(比如似然函数 $P(x|y,\theta)$)求解隐含变量 y 的概率分布 $P(y)$。流行的方法是变分贝叶斯方法【1，2】，简称变分贝叶斯(VB)。VB 已经成功用于很多场合，比如混合模型【3，4】，自动编码器(AutoEncoder)【5】，主动(active)推断(使用最小自由能原理或准则) 【6】, 等等【7】。贝叶斯推断(BI)是贝叶斯主义者用于推断模型参数的方法, 参数包括似然函数 $P(x|y,\theta)$ 中的参数和产生 $P(y)$的参数。和 BI 不同，频率主义者用的似然推断只考虑似然函数中的参数，也不考虑参数的概率分布。笔者使用频率主义方法，认为频率主义也应有求上述隐含变量(实际指概率分布 $P(y)$)的一般方法。

虽然频率主义使用的 EM(期望-最大)算法也能求解混合模型的隐含变量，但是在似然函数即预测模型不变时，如何求解 $P(y)$仍然是个问题。对于存在模糊范围约束而不是似然函数约束时，更是如此。比如，对于主动推断，给定无控制状态 x 的概率分布 $P(x)$和若干目标 $y_1, y_2, \ldots$(代表模糊范围)作为约束，我们需要选择动作 a_j 实现目标 y_j，并且优化控制在不同目标上的分配比例 $P(a_1), P(a_2)$，…。

最近，信息理论的正则化已经被用于深度学习【8】和强化学习

【9】，并取得很好效果。这种方法用两种信息的差作为目标函数，然后用变分方法最小化该函数, 并求解隐含变量。本文作者努力方向与此一致，并且希望有个在理论上更好理解的一般方法。本文使用香农互信息和语义互信息的差作为目标函数，其理论基础可以追溯到香农开创的信息率失真理论【10，11】。

一般说来，x 可能取值数目远大于 y 的可能取值数目。给定观察数据或样本分布 $P(x)$和约束 $P(x|y, \theta)$时 , 不存在 $P(y)$的正确解(因为约束太多)，它使得 $P_\theta(x)=\sum_j P(x|y_j,\theta)P(y_j)$等于 $P(x)$。但是我们可以使用某种准则得到近似解，比如 VB 算法中使用最小 KL 离散度 $KL(P(y)||P(x, y|\theta))$或 $KL(P(y|x)||P(x, y|\theta))$作为优化准则（即所谓的最小自由能原理），得 $P(y)$的近似解。本文提出语义变分贝叶斯方法(Semantic Variational Bayes' method, 缩写：SVB)，它使用最大信息效率准则——兼容最大似然准则和最大熵原理。

作者使用"语义"("Semantic")一词，一是因为 SVB 基于语义信息论——即 Shannon-Lu 理论或 G 理论【12，13】(G 意思是香农信息论的推广)；二是因为 SVB 使用似然，真值，隶属，相似、失真等函数作为约束，这些函数和语义相关。虽然 SVB 是频率主义的，但是它使用了各种推广的贝叶斯公式，完成的任务和 VB 类似，所以，笔者仍然使用"Variational Bayes"。但是，笔者用"Bayes Inference"而不是 Bayesian Inference(后者确切翻译是贝叶斯主义推断)是为了避免让人认为 SVB 是贝叶斯主义方法。

G 理论使用似然，真值，隶属，相似、失真等函数构造语义信息测度，它把信息率失真函数 $R(D)$推广为信息率逼真函数 $R(G)$——R 是给定语义互信息 G 时的最小香农互信息。G/R 等于信息效率。SVB 就来自 $R(G)$函数的参数解，其中变分和迭代方法起源于香农等人对信息率失真函数的研究【11，14，15，16】。笔者于三十年前提出语义信息 G 理论【17，18】，最近十年把这一理论用于机器学习。笔者先前的文章已经讨论了从样本分布求解或优化各种学习函数的方法，并将这些方法用于多标签分类，混合模型【13】，贝叶斯确证【19】，语义压缩【20】等。一些文章已经涉及求解隐含变量的概率分布【13, 20】。

本文首次提出 SVB，希望 SVB 成为 VB 的替代选择，这是因为：

1） VB 使用的最小自由能原理不容易被理解，并且存在混合模型反例(见 4.1 节)使得我们不能通过最大化负的自由能函数 $F=Q+H(Y)$【1，4】, 使混合模型收敛。

2） 虽然 VB 使用平均域方法(用 $P(y|x)$代替 $P(y)$作为变分)【3】时不存在反例，但是计算 $P(y|x)$比较复杂，比如 $P(y|x)$的表达式同时用到指数和对数函数。我们希望有更简单的方法。

3） 给定观察数据和其他许多学习或约束函数中的一种时，都可能存在求解隐含变量的问题。我们需要更加一般的求解方法。

本文主要目的：

- 提供理论上容易理解，计算相对简单的从 $P(x)$和各种学习或约束函数求解隐含变量概率分布 $P(y)$的一般方法；

- 使香农等人研究信息率失真函数时使用的变分和迭代方法发扬光大；
- 增加大家对语义信息 G 理论的理解。

本文主要贡献：提出、系统地介绍、并举例验证了 SVB；和 VB 比，

- 它在理论上更加容易理解，因为它兼容最大似然准则和最大熵原理，继承并发展了香农等人研究信息率失真函数发展出来的变分和迭代方法；
- 它允许使用多种约束函数，还可以通过指数 s 加强约束(参看(28))；
- 其计算更加简单，因为 $P(y|x)$和 $P(y)$的解的表达式比较简单；对于同样任务(不用 s 时)，计算不包含指数和对数函数。

因为本文使用信息理论方法，使用数据的概率分布或相对频率 $P(x_i)$ $(i = 1, 2, \ldots, m)$而不用数据序列 $x(1), x(2), \ldots, x(N)$。请读者注意：本文求平均的表达方式和统计学中表达方式有所不同。

2. 语义信息 G 理论(Shannon-Lu 理论)和 SVB

2.1 P-T 概率框架和语义贝叶斯公式

Kolmogorov 定义的概率【21】是一个随机变量属于一个集合的概率，简称集合的概率。但是要得到这种概率，我们还是需要知道集合中元素的概率。比如要知道年轻人、成年人、老年人…的概率，我们还要知道不同年龄的概率分布。一种年龄的概率就是频率主义者定义的概率【22】。所以两种概率是互补的。P-T 概率框架试图统一这两种概率，并且把集合推广到模糊集合。我们称集合的概率是逻辑概率（记为 T），集合中元素的概率是统计概率(记为 P)。我们定义：

- 随机变量 X取值于 U=$\{x_1,x_2,\ldots\}$,代表一个实例；随机变量 Y取值于 $V=\{y_1, y_2, \ldots\}$, 代表一个标签或假设。$y_j(x_i)$ 是一个命题. θ_j 是论域 U 上的一个模糊子集【23】, 其元素使得 y_j 为真. 我们有 $y_j(x) = \text{“}x \in \theta_j\text{”}$. θ_j 也可以理解为一个模型或一组模型参数。
- 用“=”定义的概率，比如 $P(y_j) \equiv P(Y = y_j)$，是统计概率；用“$\in$”定义的概率，比如 $P(X \in \theta_j)$, 是逻辑概率. 为区分 $P(Y = y_j)$ 和 $P(X \in \theta_j)$, 我们记 y_j 的逻辑概率为 $T(y_j) \equiv T(\theta_j) \equiv P(X \in \theta_j)$。
- 条件逻辑概率 $T(y_j|x) \equiv T(\theta_j|x) \equiv P(X \in \theta_j|X = x) \in [0,1]$ 是 y_j 的真值函数，也是模糊集合 θ_j的隶属函数 $m_{\theta j}(x)$，即

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

根据 Davidson 的真值条件语义学【24】，y 的真值函数反映了 y 的语义。

根据上面定义，香农使用的概率是统计概率，Kolmgorov 定义的概率是逻辑概率。一个假设 y_j的逻辑概率一般不等于它的统计

概率。比如，一个永真句的逻辑概率是 1，而它的统计概率接近 0. 我们有 $P(y_1) + P(y_2) + \ldots + P(y_n)=1$，但是可能有 $T(y_1) + T(y_2) + \ldots + T(y_n) > 1$. 比如：$T$(“非成年人”)+$T$(“成年人”)+$T$(“年轻人”) = 1+$T$(“年轻人”) > 1.

逻辑概率也需要通过统计概率得到。我们有：

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

这就 Zadeh 定义的模糊事件的概率【25】。

我们可以把 $T(\theta_j|x)$ 和 $P(x)$放进贝叶斯公式得到语义贝叶斯预测【13】：

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

$P(x|\theta_j)$就是流行方法中的似然函数 $P(x|y_j, \theta)$. 我们这里使用 $P(x|\theta_j)$是因为第 j 种参数和 y_j 是绑定的。我们称上面公式是语义贝叶斯公式。因为 $T(\theta_j|x)$的最大值是 1，从 $P(x)$ 和 $P(x|\theta_j)$, 我们推导出:

$$T(\theta_j \mid x) = \frac{T(\theta_j)P(x \mid \theta_j)}{P(x)},\ T(\theta_j) = 1 \Big/ \max_x \left(\frac{P(x \mid \theta)}{P(x)}\right). \tag{4}$$

一组转移概率函数 $P(y_j|x), j=1, 2, \ldots, n$ 构成一个香农信道，一组真值函数 $T(\theta_j|x)$ (j=1,2,…)构成一个语义信道。当语义信道匹配香农信道时，即 $T(\theta_j|x) \propto P(y_j|x)$或 $P(x|\theta_j) = P(x|y_j)$ (j=1,2,…)，语义信息量达到其最大值并等于香农互信息。

2.2 真值、隶属、相似、分辨率、和失真函数之间的关系

假设对于每个y_j存在一个典型或柏拉图的理念x_j, 它使 $T(\theta_j|x_j)=1$, 则真值或隶属度 $T(\theta_j|x_i)$就是 x_i 和 x_j 之间的相似度。如果 $U=V$，y_j 就变成估计，即 $y_j=\hat{x}_j$=“估计 x 是 x_j”。 x 和 x_j 之间的相似函数——记为 $S(x, x_j)$——等价于两者之间的混淆概率函数。比如：GPS 指示的位置和实际位置的相似度就是两者的混淆概率。

真值函数和失真函数可以相互转换。设 $d(y_j|x_i)$是 x_i 使得 y_j 失真的失真量，我们定义：

$$T(\theta_j|x_i) = \exp[-d(y_j|x_i)], \tag{5}$$

令 exp 和 log 是一对反函数，于是有

$$d(y_j|x_i) = -\log T(\theta_j|x_i). \tag{6}$$

有时候，直接定义失真函数是困难的，我们可以先定义真值函数，然后用(6)得到失真函数。比如直接定义标签“老年人”相对于不同年龄 x 的人的失真函数比较困难，我们可用一个 Logistic 函数表示“老年人”的真值函数，然后用(6)得到“老年人”的失真函数。

因为失真一般是不对称的。所以我们用 $d(y_j|x_i)$表示用 y_j 代表 x_i

的失真量。对于估计，失真函数是对称的。比如对于 GPS，失真正比于距离的平方，真值或相似函数是：

$$T(\theta_j \mid x) = S(x, x_j) = \exp[-d(x, x_j)] = \exp[-\frac{(x - x_j)^2}{2\sigma^2}]. \tag{7}$$

相似函数也就是观察者的分辨率函数。GPS 的精度 RMS(平均误差平方的开方, 上式中 σ)就反映其分辨率【12】。

隶属函数、相似函数、失真函数都反映标签或估计的语义, 因为一个隶属函数等价于一个真值函数；相似函数是真值函数的特例；一个失真函数可以转换为一个真值函数。

2.3 语义信息测度

香农互信息可以表示为：

$$I(X;Y) = \sum_j \sum_i P(y_j) P(x \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}. \tag{8}$$

香农互信息意味着因预测而节省的平均码长。我们用似然函数 $P(x_i|\theta_j)$代替 log 右边的 $P(x_i|y_j)$(左边不变), 就得到语义互信息公式：

$$I(X;Y_\theta) = \sum_j \sum_i P(y_j) P(x_i \mid y_j) \log \frac{P(x_i \mid \theta_j)}{P(x_i)} = \sum_j \sum_i P(y_j) P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \tag{9}$$

语义互信息意味着因根据语义预测而节省的平均码长【17,18】。注意：上面公式中 log 左边的 $P(x|y_j)$是用以求平均的，表示样本分布。它可以是相对频率，可能不平滑或不连续。正是因为 $P(x|\theta_j)$和 $P(x|y_j)$可能不同，上式反映信息需要事实检验。

当 $Y=y_j$时, 语义互信息就变成语义 Kullback-Leibler (KL) 信息:

$$I(X;\theta_j) = \sum_i P(x_i \mid y_j) \log \frac{P(x_i|\theta_j)}{P(x_i)} = \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{10}$$

.

进一步，$X=x_i$ 时，语义 KL 信息就变成单个事件之间(y_j 提供关于 x_i)的语义信息：

$$I(x_i;\theta_j) = \log \frac{P(x_i \mid \theta_j)}{P(x_i)} = \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \tag{11}$$

图1图解了上面公式。它表明：逻辑概率越小，信息量绝对值越大；偏差越大，信息量越小; 错误的假设传递负的信息。这些结论符合 Popper 思想[26](10.1-II 节).

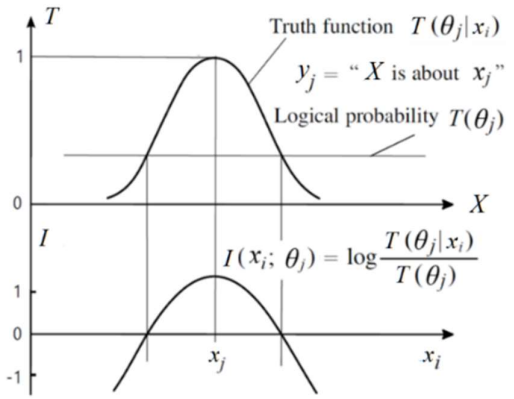


Fig. 1. Illustrating the amount of semantic information. The semantic information conveyed by y_j about x_i decreases with the deviation or distortion increasing.

图 1. 语义信息量图解。y_j 传递关于 x_i 的语义信息随偏差增大而减少。

把(5)带进(9)，我们得到：

$$I(X;Y_\theta) = -\sum_j P(y_j)\log T(\theta_j) - \mathrm{E}_{P(x,y)}d(x,\ y) = H(y_\theta) - \overline{d}. \tag{12}$$

其中 $H(Y_\theta)$ 是语义熵，$\overline{d}$ 是平均失真。可见 $I(X;\ Y_\theta)$ 就像负的正则化误差平方。因此，我们可以把最大语义互信息准则看作特殊的最小正则化误差平方(RLS)准则。

如果(9)中真值函数变成相似函数，语义互信息就变成估计互信息【27】，它已经被深度学习研究者用于信息噪声对比估计(InfoNCE)【28】和互信息神经估计(MINE)【29】。

2.4 用逻辑贝叶斯推断优化各种学习函数

作者曾提出逻辑贝叶斯推断, 英文是"Logical Bayesian Inference". 因为它并不是贝叶斯主义的方法，以后都写成"Logical Bayes' Inference"。缩写仍然是 LBI。

最常用的学习函数是似然函数 $P(x|\theta_j)$. 我们可以用语义 KL 公式优化它。因为语义 KL 信息等于两个 KL 离散度的差：

$$I(X;\ \theta_j) = KL(P(x|y_j)||P(x)) - KL(P(x|y_j)||P(x|\theta_j), \tag{13}$$

容易证明：$KL(P(x|y_j)||P(x|\theta_j) = 0$ 时，即

$$P(x|\theta_j) = P(x|y_j) \tag{14}$$

时，$I(X;\ \theta_j)$达到其最大值，$KL(P(x|y_j)||P(x))$.

在很多情况下，我们希望使用 $P(\theta_j|x)$——它是参数化的 $P(y_j|x)$——作为学习函数。Fisher 称之为反概率【31】(函数)。当 n=2 时，我们能用一对 Logistic 函数作为反概率函数；但是，当 n>2 时，构造 $P(\theta_j|x), j = 1, 2, \ldots, n$ 是困难的，因为有归一化限制：对于每个 $x, \sum_j P(\theta_j|x) = 1$. 然而，使用真值函数或相似函数就没有这一限

制。

从(4)和(14)，我们能推导出优化的 $T(\theta_j|x)$：

$$T^*(\theta_j \mid x)=\frac{P^*(x \mid \theta_j)}{P(x)}\Big/\max_x\left(\frac{P^*(x \mid \theta)}{P(x)}\right)=\frac{P(x \mid y_j)}{P(x)}\Big/\max_x\left(\frac{P(x \mid y)}{P(x)}\right)=\frac{P(y_j \mid x)}{\max_x(P(y_j \mid x))}. \tag{15}$$

上式中我们假定 $P(x|y_j)$是平滑的分布；否则，我们需要使用下式得到参数化的平滑的 $T^*(\theta_j|x)$:

$$T^*(\theta_j \mid x)=\arg\max_{\theta_j}\sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)}. \tag{16}$$

如果我们只知道 $P(y_j|x)$ 而不知道 $P(x)$, 我们可以假设 $P(x)=1/m$ ，从而得到 $T^*(\theta_j|x)$【13】。

我们称上述求解真值和相似函数的方法，(14)-(16), 为逻辑贝叶斯推断，它是为了使语义信道匹配香农信道。显然，它兼容最大似然准则。英文中，我们用“Logical Bayes' Inference” (LBI)代替“Logical Bayesian Inference” [13]，因为 LBI 不是贝叶斯主义的方法.

下面我们假设真值函数和相似函数是优化好的，样本分布等于相应的概率分布。我们称

$$c(x, y_j) \equiv P(x|\theta_j)/P(x)=P(x, y_j)/[P(x)P(y_j)] \tag{17}$$

是相关函数。Sklar 提出 copula 函数【32，33】。$c(x, y_j)$就是二维 copula 密度函数。因为 $c(x, y_j)$和 $P(y_j|x)$成正比，我们可以在贝叶斯公式中用 $c(x, y_j)$代替 $P(y_j|x)$，如下：

$$P(x \mid y_j)=\frac{P(x)c(x, y_j)}{c(y_j)},\ c(y_j)=\sum_i P(x_i)c(x_i, y_j), \tag{18}$$

其中 $P(x)$可能和先前的 $P(x)$不同。我们也称 $c(x, y)$是贝叶斯芯，因为 $P(x, y)=P(x)c(x, y)P(y)$。如果 $P(x)$没有改变，$P(x|y_j)=P(x_i)c(x, y_j)$；如果 $P(x)$改变了，则我们必须用(18)求 $P(x|y_j)$。

Copula 密度函数也可以作为约束求隐含变量【34】。给定 $P(x)$和 $c(x, y)$, 我们可以用 SVB 求解 $P(y)$。 给定 $P(x)$和某种约束求 $P(y)$时，如果形如 $c(x, y)P(y)$的表达式构成香农信道 $P(y|x)$。 应当有

$$\sum_i P(x_i \mid y_j)=\sum_i P(x_i)c(x_i, y_j)=1;\ \sum_i P(x_i)c(x_i, y_j)P(y_j)=P(y_j). \tag{19}$$

否则，$P(y)$和 $P(y|x)$是不适当的求解结果。比如 EM 算法中，收敛前的 E-step 求解出的 $P(y)$和 $P(y|x)$结果是不适当的。

如果 x 和 y 的论域相同，y_j 就变成估计 $\hat{x}_j$，并且当 $i=j$ 时，

$T(\theta_j|x_i)=T(\hat{x}_j|x_i)=1$，则真值成为对称的相似度：

$$S(\hat{x}_j \mid x_i)=\frac{c(x_i,y_j)}{\max\limits_x(c(x,y_j))}=\frac{c(x_i,y_j)}{c(x_j,y_j)}=\frac{c(x_i,y_j)}{c(x_i,y_i)}=S(\hat{x}_i \mid x_j).$$

(20)

然而，真值函数或隶属函数可能是不对称的。设和 y_i 相似的 y 构成模糊子集 θ_{xi}，我们有

$$T(\theta_{xi} \mid y_j)=\frac{P(x_i \mid y_j)}{\max\limits_y(P(x_i \mid y))}\neq\frac{P(y_j \mid x_i)}{\max\limits_x(P(y_j \mid x))}=T(\theta_j \mid x_i).$$

(21)

最近，真值函数或相似函数也被用作深度学习的学习函数【27】。例如，Oord 等人[28]在提出信息噪声对比估计，明确指出学习函数和 $c(x, y_j) = P(x|y_j)/P(x)$成正比。他们论文中的表述是：

$$f_k(x_{t+k},c_t)\propto P(x_{t+k} \mid c_t)/P(x_{t+k}), \tag{22}$$

其中 c_t 是从前面的数据得到的特征矢量, x_{t+k} 是要预测的特征矢量。文中 $f_k(x_{t+k}, c_t))$ 就是根据 c_t 预测的 x_{t+k} 和实际的 x_{t+k} 之间的相似函数, 用它构造的估计互信息就是语义互信息的特例。

2.5 SVB 的起源——从信息率失真函数到信息率逼真函数

作者使用”信息率逼真”有两个原因，一是“逼真度”(verisimilitude 【26】或 truthlikeness 【35】)一直是哲学家们讨论的话题。它表明了从不太真实到真实的改进或失真的减少。由于科技文章中 “verisimilitude” 和“ truthlikeness”不是常用词，所以英语中，我们用信息论常用的近义词“Fidelity”(保真度或逼真度)表示逼真度。二是因为香农最初就提出信息率逼真(information rate fidelity)准则【10,11】，后来才使用最小失真(distortion)作为最大逼真(fidelity)的一种表示方法。现在我们用语义信息测度代表逼真度应当符合 Popper 和香农原意。

由于率失真等于信息率失真，Cover 和 Thomas 在【36】(10.2 节)中表示他们也用“率失真” 表示“信息率失真”，即对两者不加区分。基于同样的原因，本文对“信息率逼真“和”率逼真”也不加区分。

如果将 $R(D)$的失真限制 $\bar{d}\le D$ 改为 $I(X; Y_\theta)\ge G$，则 $R(D)$成为信息率逼真函数 $R(G)$【12，13】。这时，我们将 $d_{ij}=d(x_i, y_j)$替换为 $I_{ij}=I(x_i; \theta_j)$。

除了给定 $P(x)$和 G，还有下面限制：

$$\sum_i P(x_i \mid y_j)=1,\ j=1,2,...,n; \tag{23}$$

$$\sum_j P(y_j)=1. \tag{24}$$

我们用拉格朗日乘子法求最小互信息。拉格朗日函数是：

$$L(P(y \mid x),P(y))=I(X;Y)-sI(X;Y_\theta)-\mu_j\sum_i P(x_i \mid y_j)-\alpha\sum_j P(y_j) \tag{25}$$

其中 $I(X; Y_\theta)$和 $I(X; Y)$用 $P(y|x)$和 $P(y)$表示是：

$$I(X;Y_\theta)=\sum_i P(x_i)\sum_j P(y_j\mid x_i)\log m_{ij},\ m_{ij}=T(\theta_j\mid x_i)/T(\theta_j)=P(x_i\mid\theta_j)/P(x_i) \tag{26}$$

$$I(X;Y)=\sum_i P(x_i)\sum_j P(y_j\mid x_i)[\log[P(y_j\mid x_i)-\log P(y_j)]. \tag{27}$$

因为 $P(y|x)$ 和 $P(y)$相互依赖，我们交替地使用其中一个做变分。首先我们固定 $P(y)$，令：

$$\partial L/\partial P(y_j\mid x_i)=0,\ j=1,2,...,n;\ i=1,2,...,m\ .$$

于是得到：

$$P(y_j\mid x_i)=P(y_j)m_{ij}^s/\lambda_i,\ \lambda_i=\sum_j P(y_j)m_{ij}^s,\ i=1,2,...;\ j=1,2,... \tag{28}$$

我们可以把 m^s_{ij}/λ_i 当做贝叶斯芯，并且把(28)看做推广的贝叶斯公式。然后我们固定 $P(y|x)$, 令

$$\partial L/\partial P(y_j)=0,\ j=1,2,...,n.$$

我们得到：

$$P(y_j)=\sum_i P(x_i)P(y_j\mid x_i). \tag{29}$$

重复(28)和(29)，就像我们为 $R(D)$函数所做的那样【16】(326 页)，我们就能得到适当的 $p(y)$和 $p(y|x)$。如果没有这个迭代，m^s_{ij}/λ_i 就是不合适的贝叶斯芯或 copula 密度函数，因为 $P(x|y_j)=P(x)m_{ij}s/\lambda_i$ 不是归一化的。(28)和(29)就是最小互信息迭代, 让香农信道匹配语义信道的方法。它和 LBI 一起构成语义变分贝叶斯。

最后，我们得到 $R(G)$函数的参数解(参看图 2)：

$$\begin{aligned} G(s)&=\sum_i\sum_j P(x_i)P(y_j\mid x_i)I_{ij},\\ R(s)&=sG(s)-\sum_i P(x_i)\log\lambda_i, \end{aligned} \tag{30}$$

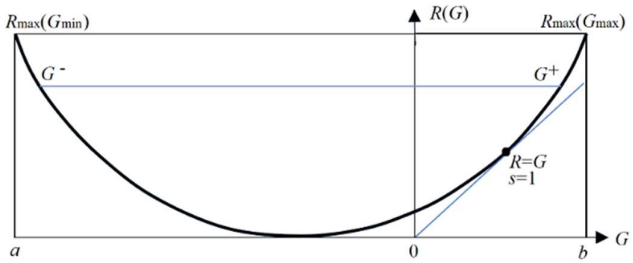


Fig. 2. The information rate fidelity function $R(G)$ for binary communication. Any $R(G)$ function is bowl-like and has a point where s=1 and $R = G$. For given R, there are two anti-functions $G^-(R)$ and $G^+(R)$.

图 2. 二元通信的信息率逼真度函数 $R(G)$。任何 $R(G)$函数都是碗状的，其中有一个 $R(G)=G$ 的点 (s=1)。给定 R 有两个反函数 $G^+(R)$和 $G^-(R)$。

任何 $R(G)$函数都是碗状的(有可能不太对称)【12】，其中二阶导数大于 0。s =dR/dG 在右边部分是正的。当 s=1 时, $G = R$, 这意味着

语义信道匹配香农信道。G/R 表示信息效率。给定 R，G 有最大值 G^+和最小值 G^-, G^-意味着有意使用谎言时，G 能小到什么程度.

$R(G)$函数可以应于根据视觉分辨率的图像压缩【12】,不可见实例最大互信息分类和混合模型收敛证明【13】，和语义压缩【20】。

值得注意的是，给定语义信道 $T(y|x)$时, 让 $P(y_j|x) \propto T(y_j|x)$ 或 $P(x|y_j) = P(x|\theta_j)$并不最大化语义互信息，而是最小化香农互信息或最大化信息效率 G/R。然后，我们能在此基础上通过增大 s 同时增大 G 和 R。当(28)中 s->∞, $P(y_j|x)$ (j = 1, 2, …, n) 就仅取值 0 或 1，成为分类函数。

SVB 使用最小香农互信息准则加上最大语义互信息准则，两个准则加一起就等于最大信息效率准则. 不过有时候它并不追求 G/R=1，而是在给定预测和控制的不同质量要求时最大化 G/R。

我们也能用模糊熵 $H(Y_\theta|X)$ 代替平均失真 $\bar{d}$，得到信息率真函数 $R(\Theta)$ [21]. 在信息而不是真更加重要的场合，$R(G)$比 $R(D)$和 $R(\Theta)$更加合适. 求 $R(\Theta)$得到的 $P(y)$和 $P(y|x)$与求 $R(G)$得到的不同，因为优化准则不同。在最大真值准则下，$P(y|x)$变为：

$$P(y_j \mid x_i) = P(y_j)[T(\theta_{xi} \mid y_j)]^s \Big/ \sum_j P(y_j)[T(\theta_{xi} \mid y_j)]^s, i = 1,2,...; j = 1,2,... \tag{31}$$

如果 $T(\theta_{xi}|y)$=exp[-$d(x_i|y_j)$], $R(\Theta)$就变成 $R(D)$【21】. 如果 $T(\theta_j)$很小，$R(G)$需要的 $P(y_j)$会大于 $R(D)$或 $R(\Theta)$需要的 $P(y_j)$.

3. 用 SVB 求解隐含变量

3.1 隐含变量的近似解和优化准则

从语义信息论的角度看，我们可以把观察数据的概率分布 $P(x)$看成信源，把要求解的 $P(y|x)$看成信道，即香农信道。通常的约束条件是 $P(x|y_j)$或 $P(x|\theta_j)$ (j = 1, 2, …, n).

已知 $P(x)$和 $P(x|y)$, 求解 $P(y)$, 我们可以列出 m 个等式：

$$P(x_i|y_1)P(y_1) + P(x_i|y_2)P(y2) + \ldots + P(x_i|y_n)P(y_n) = P(x_i),\ i = 1, 2, \ldots, m. \tag{32}$$

上面公式加上 $P(y_1) + P(y_2) + \ldots + P(y_n) = 1$, 一共 m+1 个等式。当 n=m+1 时，我们可能得到 $P(y)$的精确解。当 $n > m$+1 时，$P(y)$有多种解。当 $n < m$+1 时，n 无解；但是，使用某种准则，我们可以得到损失最小或信息最大解。后面我们只考虑 $n > m$+1 的情况，使用最大信息效率准则，得到 $P(y)$近似解。

如果约束不是 $P(x|y_j)$或 $P(x|\theta_j)$而是 $T(\theta_j|x)$，我们可以通过(3)得到 $P(x|\theta_j) = P(x)T(\theta_j|x)/T(\theta_j)$。

需要注意的是， 约束函数是 x 上的分布，比如 $T(\theta_j|x)$和 $s(x, x_j)$, exp[-$d(x, y_j)$] (j = 1, 2, …, n)，不是归一化的；而要求解的函数 $P(y|x)$

是 y 上的分布, 即 $P(y|x_i)$, $i = 1, 2, \ldots, m$. 它应当是归一化的。log 的左边只能放 $P(y|x_i)$以便和 $P(x_i)$一起求平均。

作者曾考虑过：既然最小互信息迭代是为了让 $P(y_j|x) \propto T(\theta_j|x)$, $j = 1, 2, \ldots$, 我们令 $P(y) \propto T(\theta_j)$即 $P(y_j)=T(\theta_j)/\sum_j T(\theta_j)$, 再令 $P((y_j|x)=T(\theta_j|x)P(y_j)/T(\theta_j)$ 就行了。然而，这样做不能保证对于每个 x_i, $P(y|x_i)$是归一化的，即 $\sum_j P(y|x_i)=1$。

3.2 解释和改进用于混合模型的 EM 算法

我们知道 $P(x)=\sum_j P(y_j)P(x|y_j)$。给定样本分布 $P(x)$, 我们使用 $P_\theta(x)=\sum_j P(y_j)P(x|\theta_j)$逼近 $P(x)$ ($P(y)$就是隐含变量的概率分布), 使得下面的相对熵接近 0：

$$H(P \| P_\theta) = \sum_i P(x_i) \log[P(x_i) / P_\theta(x_i)]. \tag{33}$$

EM 算法首先预设 $P(x|\theta_j)$和 $P(y_j)$，$j = 1, 2, \ldots, n$。它的 E-step 得到:

$$P(y_j \mid x) = P(y_j)P(x \mid \theta_j) / P_\theta(x),\ P_\theta(x) = \sum_k P(y_k)P(x \mid \theta_k), \tag{34}$$

然后在 M-step 最大化完全数据对数似然度(通常用 Q 表示)——分成两步， 包括 M1-step：

$$P^{+1}(y_j) = \sum_i P(x_i)P(y_j \mid x_i). \tag{35}$$

和 M2-step：优化似然函数

$$P(x \mid \theta_j^{+1}) = P(x)P(y_j \mid x) / P^{+1}(y_j) = P(x)\frac{P(x \mid \theta_j)}{P_\theta(x)}\frac{P(y_j)}{P^{+1}(y_j)}. \tag{36}$$

如果是高斯混合模型，就是用 $P(x)P(y_j \mid x) / P^{+1}(y_j)$ 的期望和标准偏差作为 $P(x \mid \theta_j^{+1})$ 的期望和标准偏差。

重复上面两个步骤可以使混合模型收敛。根据信息率逼真函数的推导过程，可见 E-step 和 M1-step 最小化香农互信息；根据(14)，M-step 最大化语义互信息。所以 EM 算法使用的优化准则是最大信息效率准则。

但是，用上面方法求隐含变量有两个问题：1)$P(y)$可能收敛很慢；2)如果似然函数也是固定的，如何求解 $P(y)$?。

基于信息率逼真函数的分析，我把 EM 算法改进为 EnM 算法。EnM 算法包含 EM 算法中的 E-step, n-step 和 EM 算法中的 M2-step。n-step 步重复 EM 算法中的 E-step 和 M1-step n 次，使得 $P^{+1}(y) \approx P(y)$。显然，EnM 算法使用的也是最大信息效率准则。n-step 可以加速求

解隐含变量概率分布 $P(y)$。因为 $P(y_j)/P^{+1}(y_j)$约等于 1，我们可用下式优化模型参数：

$$P(x|\theta_j^{+1})=P(x)P(x|\theta_j)/P_\theta(x). \tag{37}$$

如果没有 n-step, 会有 $P(y_j)\neq P^{+1}(y_j)$，$\sum_i P(x_i)P(x|\theta_j)/P_\theta(x_i)\neq 1$, 这说明 $P(y)$和贝叶斯芯 $P(x|\theta_j)/P_\theta(x)$对于 $P(x)$来说是不合适的。

在求解混合模型时，我们可以选取 n 小一点， 比如选 n=3。在专门求解 $P(y)$时，我们可以选取较大 n，直至 $P(y)$收敛。当 n=1 时，EnM 算法就变为 EM 算法。

下面我们证明 EnM 算法收敛。E-step 之后，香农互信息变为：

$$R=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(y_j|x_i)}{P^{+1}(y_j)}, \tag{38}$$

我们定义：

$$R''=\sum_i\sum_j P(x_i)\frac{P(x_i|\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(x_i|\theta_j)}{P_\theta(x_i)}. \tag{39}$$

然后我们能推导出： 在 E-step 之后，有：

$$H(P\|P_\theta)=R''-G=R-G+H(P_Y^{+1}\|P_Y), \tag{40}$$

其中 $H(P||P_\theta)$是 x 的相对熵或 KL 离散度; 另一个相对熵是：

$$H(P_Y^{+1}||P_Y)=\sum_j P^{+1}(y_j)\log[P^{+1}(y_j)/P(y_j)]\cdot \tag{41}$$

它在 n-Step 之后接近 0.

式(40)可以用来证明混合模型收敛。因为 M2-step(语义信道匹配香农信道)最大化 G, 且 E-step 和 n-step(香农信道匹配语义信道)最小化 R 和 $H(P^{+1}(y)\|P(y))$, 使 $H(P\|P_\theta)$接近 0。上面方法也能用来证明 EM 算法收敛。

EnM 算法在大多数情况下比 EM 算法表现更好【37】, 特别是在初始的 $P(y)$和收敛的 $P(y)$相差较大时。

3.3 信道混合模型和数据压缩

混合模型的似然函数 $P(x|\theta_j)(j=1, 2, \ldots, n)$也可以通过语义信道 $T(\theta_j|x)(j = 1, 2, \ldots, n)$和信源 $P(x)$产生。给定 $P(x)$和语义信道求香农互信息的最小值，这是数据压缩问题。如果同时通过改变语义信道中参数最大化语义互信息，那就是求信道混合模型。

在上面 EnM 算法中，$I_{ij}=\log[P(x|\theta_j)/P(x_i)]$。如果似然函数变成真值函数，就有 $I_{ij}=\log[T(\theta_j|x_i)/T(\theta_j)]$。E-step 变为：

$$P(y_j \mid x_i) = \frac{P(y_j)P(x_i \mid \theta_j)}{\sum_k P(y_k)P(x_i \mid \theta_k)} = \frac{P(y_j)T(\theta_j \mid x_i)/T(\theta_j)}{\sum_k P(y_k)T(\theta_k \mid x_i)/T(\theta_k)}.$$

(42)

假设 $T(\theta_j|x)$是高斯分布，优化 $T(\theta_j|x)$时，我们可用 $P(y_j|x)$的期望和标准偏差作为 $T(\theta_j|x)$的期望和标准偏差。用(42)和(29)可以求出隐含变量概率分布 $P(y)$。加上 s 如(28)就可以得到 $R(G)$函数参数解——用于数据压缩。

真值函数 $T(\theta_j|x)$表示 y_j 所代表的 x 的模糊范围。失真限制是范围限制的特例。如果给定失真函数而不是真值函数，我们用 $T(\theta_j|x) = \exp[-d(x, y_j)]$作为真值函数，其他同上。

如果我们用最小失真准则代替最大语义信息准则，则(42)变为：

$$P(y_j \mid x_i) = \frac{P(y_j)\exp[-d(x_i, y_j)]}{\sum_k P(y_k)\exp[-d(x_i, y_j)]}, \text{对于所有 } i \text{ 和 } j.$$

(43)

如果在(43)中加上 s，我们就得到 $R(\Theta)$或 $R(D)$函数的参数解。

3.4 优化约束控制和强化学习中的隐含变量

信息最大最小方法【8】,【9】已经用于约束控制和强化学习。SVB 应能提供更简洁方法。

G 测度也能用来度量随机事件控制的合目的信息。一个不确定控制目标可用一个真值函数或隶属函数表示。比如下面目标：

- “粮食产量接近或超过 500kg/亩";
- "工人工资最好超过 5000 元";
- "人口死亡年龄最好超过 80 岁";
- "列车到达时间误差最好不超过 1 分钟".

语义 KL 信息能用来度量合目的信息：

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

(44)

公式中 θ_j 是一个模糊集合，表示控制目标是一个模糊范围。y_j 在这里变成 a_j，表示相应第 j 个控制任务 y_j 采用的行动。如果有数个控制目标 $y_1, y_2, \ldots$ 我们可以用语义互信息公式表达合目的信息:

$$I(X; A / \theta) = \sum_j P(a_j) \sum_i P(x_i \mid a_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)}.$$

(45)

其中 A 是取值 a 的随机变量，$P(a)$是隐含变量的概率分布，我们需要用最大信息效率准则优化 $P(a)$。和混合模型不同，这里我们不需要改变 $P(x|\theta_j)$使得 $P_\theta(x)$接近 $P(x)$。

控制的合目的信息(即祈使句的语义信息)公式和陈述句的语义信息公式在形式上是相同的，但是优化方式不同。对于陈述句，事实 $P(x|y_j)$是不变的，我们希望语义概率预测 $P(x|\theta_j)$符合事实。而对于祈使句，我们希望事实 $P(x|a_j)$ 符合目的 $P(x|\theta_j)$。

对于多任务，我们需要最小化香农互信息 $I(X; A)$——在给定语义互信息 G 和归一化限制时。当实际分布 $P(x|a_j)$ 接近约束后的分布 $P(x|\theta_j)$时, 信息效率达到其最大值 1。

使用(28)和(29)，我们得到:

$$P(a_j \mid x) = P(a_j) m_{ij}^s \Big/ \lambda_i \,,\ \lambda_i = \sum_j P(y_j) m_{ij}^s \,, \tag{46}$$

$$P(a_j) = \sum_i P(x_i) P(a_j \mid x_i). \tag{47}$$

重复上面二式直至 $P(a)$不变，使得 $m_{ij}^s \Big/ \lambda_i$ 成为适当的贝叶斯芯，

然后得到：

$$P(x_i \mid a_j) = P(x_i) m_{ij}^s \Big/ \lambda_i \,. \tag{48}$$

要进一步增大合目的信息, 我们可以通过增大 s，得到随 s 变化的隐含变量 $P(a_j)$, j=1, 2, …

因为优化的 $P(x|a_j)$是 θ_j 和 s 的函数, 我们记 $P^*(x|a_j)=P(x|\theta_j, s)$。值得注意的是，许多分布 $P(x|a_j)$都满足约束并使 $I(X; a_j/\theta_j)$达到最大，但是只有 $P^*(x|a_j)$ 使 $I(X; a_j)$ 达最小.

假设实际控制结果是呈高斯分布的似然函数 $P(x|\beta_j)$，我们可以用 $P(x|\beta_j)$代替 $P(x|\theta_j, s)$， 即用 $P(x|\theta_j, s)$的期望和标准偏差作为 $P(x|\beta_j)$的期望和标准偏差，得到近似最优控制结果。

假设强化学习中行动的效果 $P(x|\beta_j)$是已知的,不需要学习，则上面方法也可以用来优化强化学习。

4. 实验结果

4.1 混合模型——使用最大信息效率准则

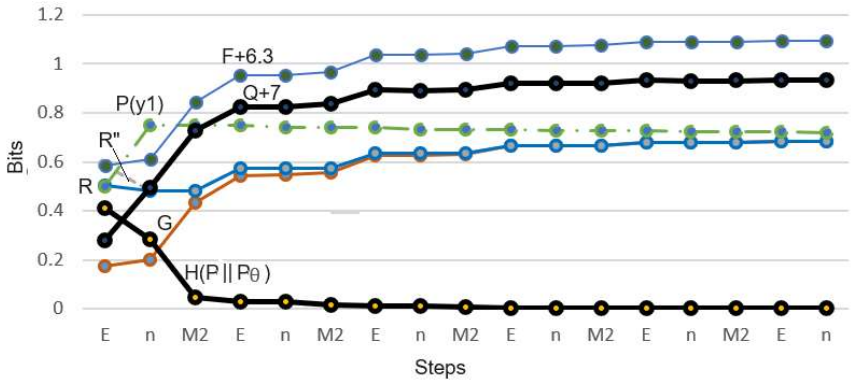
作者用EnM算法求解了一个简单的混合模型，并记录了迭代过程中 G, R, R'', Q, F 随模型参数的变化(见表 1 和图 3). 该模型包含两个组件。实际的、起始的和收敛的 $P(y)$和 θ 如表 1 所示。在例 A 中，起始 $P(y)$和 θ 使 $R<R^*$(R^* 是真模型的 R)。在例 B 中，起始 $P(y)$和 θ 使

$R > R^*$。$P(x)$是用真模型参数产生的。

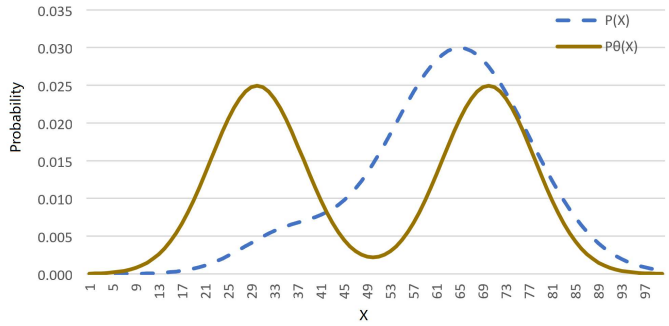
表 1. 实际的、开始的和收敛的模型参数。

Table 1. Real, start, and guessed model parameters.

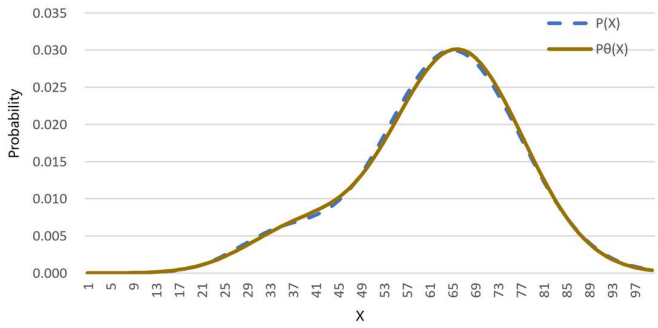
Example		Real parameters			Start parameters $H(P\|\|P_\theta)$=0.680 bit			Parameters after 5 Iterations $H(P\|\|P_\theta)$=0.00092 bit		
		μ^*	σ^*	$P^*(y)$	μ	Σ	$P(y)$	μ	σ	$P(y)$
A	y_1	35	8	0.7	30	15	0.5	35.4	8.3	0.720
	y_2	65	12	0.3	70	15	0.5	65.2	11.4	0.280
B	y_1	35	8	0.1	30	8	0.5	38	9.3	0.134
	y_2	65	12	0.9	70	8	0.5	65.8	11.5	0.866



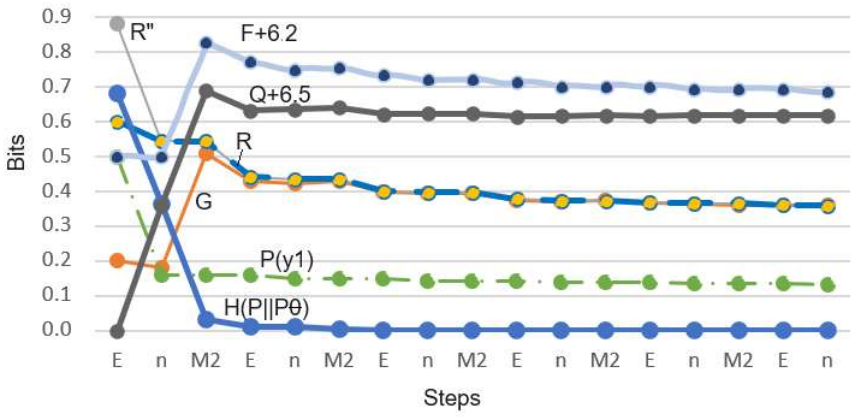
(a)



(b)



(c)



(d)

图 3. 例 A 和 B 中混合模型的迭代过程。(a) 例 A 的迭代过程中，R，G，Q 和 F 都是增加的; (b)例 B 的迭代开始; (c)例 B 的迭代结束; (d)例 B 的迭代过程，其中 R，G，Q 和 F 在大多数步骤中是减少的。

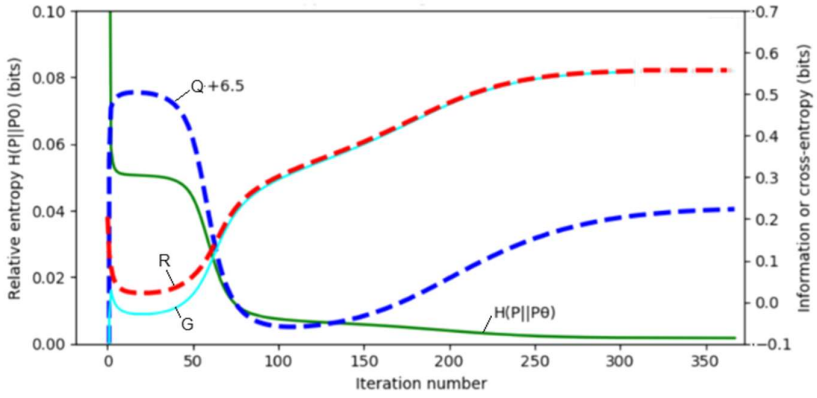
Fig. 3 The iterative processes of the mixture models. (a) The iterative process of Example A, where R, G, Q, and F are increasing. (b)The iterative start of Example B. (c) The iterative end of Example B. (d) The iterative process of Example B, where R, G, Q, and F are decreasing in most steps.

图 3 显示，$H(P||P_\theta) = R'' - G$ 在每一步中是减少的。完整数据对数似然度 Q、负自由能 F 可能增加，如图 3a 所示，也可能减少，如图 3d 所示。

例 B 是一个反例——对于通过增大 Q【34】或 F【4】来证明 EM 算法收敛的收敛证明来说。

4.2 E3M 和 EM 算法比较

图 4 显示一个不易收敛的高斯混合模型例子，用以比较 EM 和 E3M 算法。真实模型参数是：$(\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*, P^*(y_1)) = (100, 125, 10, 10, 0.7)$. 初始模型参数是: $(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1)) = (80, 95, 5, 5, 0.5)$.



(c)

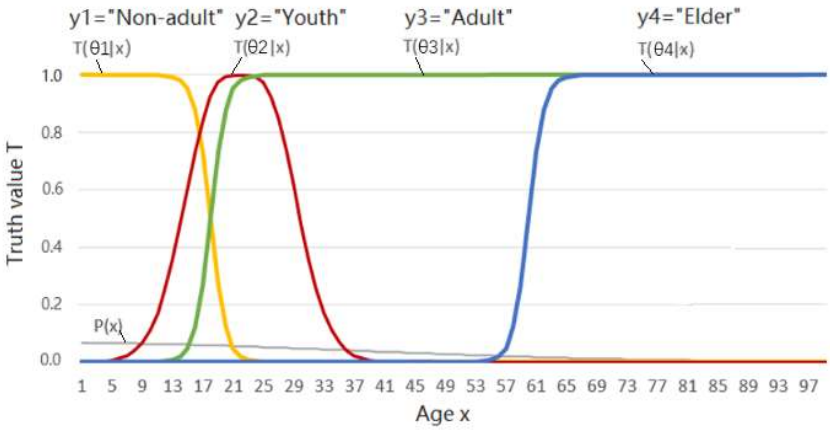
图 4. 用一个不易收敛的例子比较 EM 和 E3M 算法。EM 算法需要大约 340 次迭代，而 E3M 算法需要大约 240 次迭代。收敛过程中，完整数据对数似然度 Q 并不是单调增加的。$H(P||P_\theta)$随 $R - G$ 减小。

Fig. 4. Comparing EM and E3M algorithms with an example that is hard to converge. The EM algorithm needs about 340 iterations, whereas the E3M algorithm needs about 240 iterations. In the convergent process, complete data log-likelihood Q is not monotonously increasing. $H(P||P_\theta)$ decreases with $R - G$.

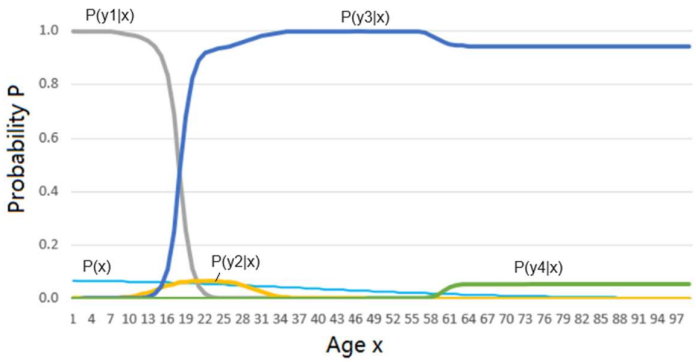
这个例子揭示：EM 算法收敛是因为语义互信息 G 和香农互信息 R 相互靠近，并不是因为完整数据对数似然度 Q 不断增大(流行观点)。详细讨论见【37】中的 3.3 节。

4.3. 数据压缩——根据最大信息效率准则

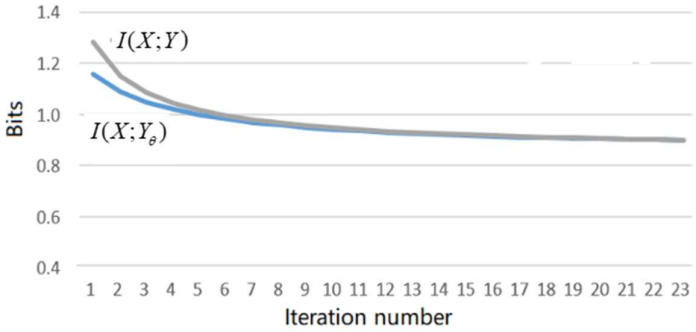
设 4 个标签 y_1="非成年人"，y_2="年轻人",y_3="成年人"和 y_4="老年人"的真值函数随年龄 x 的分布如图 5a 所示 (【20】中附录 B 有产生这些曲线的公式)。任务是：使用最大信息效率准则，用 4 个真值函数作为约束条件(s=1)获得产生最小互信息的香农信道 $P(y|x)$.



(a)



(b)



(c)

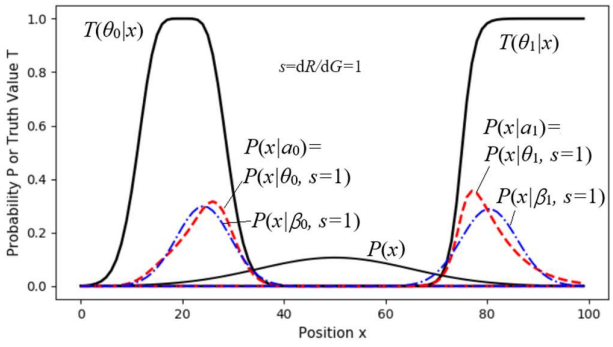
图 5. 给定约束范围求传递最小互信息的香农信道 P(y|x)。(a)关于年龄的 4 个标签的真值函数；(b)收敛的香农信道 $P(y|x)$; (c) $I(X; Y_\theta)$和 $I(X; Y)$在迭代过程中的变化.

Figure 5. Finding the iterative result of the $R(G)$ function for the given constraint ranges. (a) The truth function of four labels about ages; (b) Convergent Shannon channel $P(y|x)$; (c) The variation of $I(X; Y_\theta)$ and $I(X; Y)$ during the iterative process.

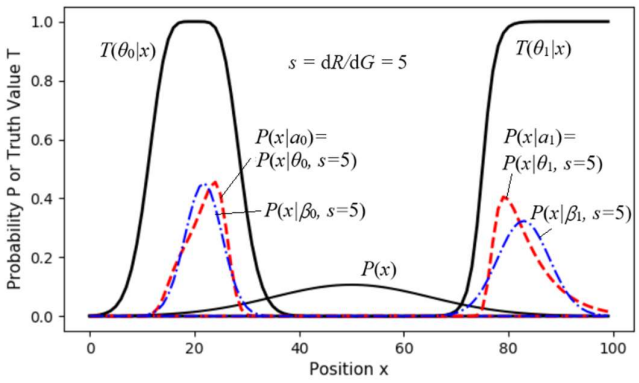
图 5b 显示：4 个转移概率函数分别和四个真值函数覆盖区域相近，但是最大值不同。$P(y_j|x)$和 $T(\theta_j|x)$只是近似成正比，这是因为前者是归一化的。从图 5c 可见语义互信息和香农互信息随着迭代而逐渐相等。收敛的 4 个隐含变量是$\{P(y_1), P(y_2), P(y_3), P(y_4)\}$ = {0.3619, 0.0200, 0.6120, 0.0057}。$P(y_1)$和 $P(y_3)$较大是因为这样能节省香农互信息 R。最小 R 是 0.883 比特。我们也计算了 $R(\Theta)(s=1)$。$R(G)$需要的 $P(y_2)$和 $P(y_4)$比 $R(\Theta)$需要的大。原因是：一个带有较小逻辑概率的标签能传递更多信息，因此应该被更多选择——如果我们使用最大语义信息准则而不是最小失真准则。对于同样的约束, $R(G) = 0.883$ 比特，它大于 $R(\Theta) = 0.845$ 比特.

4.4 多目标控制——根据难度和效果优化控制频率比例

图 6 显示了一个两目标控制任务，目标用真值函数 $T(\theta_0|x)$和 $T(\theta_1|x)$表示. 我们可以想象，它们是两个边界模糊的牧场，我们需要赶羊去那里。如果没有控制，羊群的密度分布如 $P(x)$所示。我们需要合适的分布 $P(a)$=$P(y)$.



(a)



(b)

图 6. 一个两目标控制任务。$P(x|a_j)=P(x|\theta_j, s)$如虚线所示；$P(x|\beta_j, s)$是 a_j 产生的正态分布，如点划线所示。(a)显示 了 s=1 时的情况；(b)显示 s=5 时的情况。

Figure 6. A two objective control task. The dashed line shows $P(x|a_j)=P(x|\theta_j, s)$; the dotted line displays $P(x|\beta_j, s)$, which is a normal distribution produced by action a_j. (a) The case with s = 1; (b) The case with s = 5.

图 6 中, $P(x)$ 是正态分布，期望是 μ = 50，标准差是 σ =15; 两个真值函数表示两个目标，它们是：

$$T(\theta_0|x) = 1-[1-\exp(-(x-20)^2/(2*25))]^3, \tag{49}$$

$$T(\theta_1|x) = 1/[1+\exp[-0.8(x-c)]. \tag{50}$$

对于不同的 s，我们设置初始比例：$P(a_0)=P(a_1)=0.5$。然后，我们用 (46) 和(47)做最小互信息迭代，得到收敛的 $P(a_j|x)$和 $P(a_j)$ (j=0,1)。然后我们通过(48)得到 $P(x|a_j)=P(x|\theta_j, s)$。最后，我们用(30)得到 $G(s)$, $R(s)$, 和 $R(G)$。

表 2 显示了 $P(a)$随(46)中 s 和(50)中 c 的变化。图 7 显示了 $R(G)$ 函数。表 2 显示 G 和 R 随 s 增大而增大, G/R 随 s 增大而减小；$P(a_1)$ 随 c 增大而减小。$P(a)$的变化揭示了：求 $P(a)$的迭代算法能减小控制次数在困难任务 y_1 上的分配比例 $P(a_1)$。

表 2. $R(G)$、$P(a)$和 G/R 随 s 和 c 的变化

s	c	$P(a_0)$	$P(a_1)$	G (bits)	R(bits)	G/R
1	75	0.535	0.465	3.43	3.43	1
1	80	0.579	0.421	3.80	3.80	1
5	75	0.540	0.460	3.89	4.29	0.907
5	80	0.592	0.408	4.28	4.71	0.909
40	75	0.540	0.460	3.95	5.01	0.803
40	80	0.592	0.408	4.33	5.34	0.811

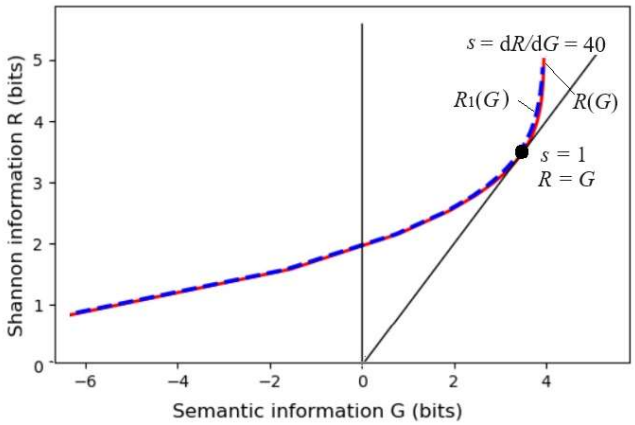


图 7. 一个两目标控制任务中的 $R(G)$函数。虚线表示 $R_1(G)$函数——是用正态分布 $P(x|\beta_j, s)$ (j=0,1)代替 $P(x|\theta_j, s)$ 产生的.

Figure 7. The $R(G)$ function for a two-objective control task. The dashed line represents the $R_1(G)$ function, which is generated by replacing $P(x|\beta_j, s)$ with a normal distribution $P(x|\beta_j, s)$ (j=0,1).

表 2 和图 7 告诉我们：当 s 从 5 增大到 40 时，G 稍许增大，这意味着 s=5 就足够好了。继续增大 s 和香农互信息，控制效率会明显降低。图 7 中 $R_1(G)$ (虚线)告诉我们：如果我们用一个正态分布代替 $P(x|\theta_j, s)$ (j=0,1) 作为控制结果, 得到的 $R_1(G)$和 $R(G)$ 类似。

5. 讨论

5.1 语义变分贝叶斯和信息率失真理论及最大熵理论之间的关系

语义变分贝叶斯使用的最小互信息迭代(见(28)和(29))直接来自信息率逼真函数 $R(G)$的参数解，然而，它起源于香农等人的信息率失真函数研究。虽然用于信息率失真函数的约束只有失真函数，用于信息率逼真函数的约束可以是各种学习函数, 包括似然函数、真值函数，隶属函数，相似函数和失真函数。我们用这些函数构造语义信息函数 $I(x; \theta_j)$ (j = 1, 2, …)，用 $I(x; \theta_j)_j$ 和其均值 G 作为约束求最小香农互信息。因为 $I(x; \theta_j) = \log[P(x|\theta_j)/P(x)]$,所以 SVB 兼容最大似然准则。

给定 $P(x)$时，香农互信息最小化等价于后验熵 $H(X|Y)$最大化。所以SVB使用的最大信息效率准则包含最大熵准则。另外，SVB 允许使用 s(见(28))加强约束，它有助于我们在增大 G 和增大信息效率 G/R 之间权衡。

5.2 关于混合模型收敛的原因

有人认为，EM 使混合模型收敛是因为完整数据对数似然度 $Q = -H(X, Y_\theta)$不断增大【38】,【39】。还有人使用变分贝叶斯方法

【1】,【4】，认为不断增大 $F = H(Y)+Q=H(Y)-H(X, Y_\theta)$ (负的自由能)可使混合模型收敛。容易看出，F 在 M-step 之后变成 X 的负的语义后验熵: $-H(X|Y_\theta)$。 它和语义互信息 $I(X; Y_\theta) = H(X) - H(X|Y_\theta)$同时增大或减小（参看图 3）。然而，4.1 节中反例表明：Q、F 和 G 在混合模型收敛的过程中可能是减小的。这说明，流行的混合模型收敛理论存在问题。使用 VB 不易发现上面问题是因为，在 VB 的实际应用时，常用的是平均域方法。这种方法用 $P(y|x)$代替 $P(y)$作为变分【3】，使得 F 不再近似于 $-H(X|Y_\theta)$，而是近似于 x 的对数似然度或负的交叉熵： $-H_\theta(x)$. 这时 F 就随混合模型收敛而增大，没有反例。

3.2 节提供了一个新的混合模型的收敛证明。它证明不断增大 G 并减小 R（即增大信息效率）能使相对熵 H(P‖Pθ)接近 0.

求解混合模型不但求隐含变量的概率分布，还求某种类型的模型(似然函数或真值函数)的参数。SVB 有助于混合模型求解；反过来，求解混合模型的 EnM 算法检验了 SVB。

5.3 语义变分贝叶斯和变分贝叶斯比较

SVB 和 VB 的主要任务一样——使用变分方法，根据观察数据和约束求解隐含变量。不同的是：

- **准则:**
  - VB 在使用 P(y)作为变分时，它采用最小自由能准则（接近于最大语义信息准则)；而在使用 P(y|x)作为变分时，它使用的是使混合模型收敛的最大似然准则。.
  - SVB 采用最大信息效率准则，即在优化香农信道 $P(y|x)$ 时使用最小(香农)互信息或最大熵准则，而在优化预测模型或语义信道 T(y|x)时使用最大语义信息准则。
- **变分方法:**
  - VB 使用平均域近似方法——把所有 $P(y)$或 $P(y|x)$划分成若干子集， 然后轮流用每个子集中的元素作为变分【1, 3】.
  - SVB 交替用 $P(y|x)$和 $P(y)$做变分。
- **计算复杂性:**
  - VB 中的 P(y|x)表达式包含对数和指数函数【1，3】，计算比较复杂.
  - SVB 中的 $P(y|x)$表达式不包含对数和指数函数，计算比较简单(对于同样任务，即 s=1 时)。
- **约束:**
  - VB 只用似然函数作为约束。
  - SVB 允许使用各种学习函数（包括似然、真值、隶属、相似、失真函数）作为约束，并且允许用参数 s 增强约束。。
- **理论基础:**
  - VB 基于 BI(Bayesian Inference)，其中 Bayesian 意味着贝叶斯主义方法。

- SVB基于语义信息论——Shannon-Lu理论，其中Bayes意味着推广的贝叶斯公式。

按照流行的VB定义，$P(y)$用作变分，最大化F。但是很多人实际上用$P(y|x)$作为变分【1】,【3】。具体做法是把y1, y2, …分成若干组，轮流选择每一组的$P(yj|x)$用作变分. 如果用$P(y|x)$代替$P(y)$，F就近似于x的负的交叉熵。这时，这时VB和SVB的差别就比较小。但是，SVB中还有求最小互信息的n-step，这是VB中没有的。

因为SVB更加兼容最大似然和最大熵准则，计算较为简单，它应该更加适合机器学习的很多场合。但是，因为它不考虑参数的概率，在某些场合，可能不如VB适用。

5.4 需要进一步研究如何把SVB和神经网络相结合

虽然SVB还没有用于神经网络，但是它提供了应用的可能性。首先因为SVB为使用相似函数和真值(或隶属)函数提供了方便。

把神经网络的权重解释成概率往往是困难的， 因为概率有归一化要求。比如，如果希望$m*n$个权重表示n个呈某种分布的转移概率函数$P(y_j|x)(j = 1, 2, …, n)$, 因为有归一化要求，构造这些函数是困难的。如果我们用权重表示真值函数或相似函数，我们对权重就没有归一化要求。我们需要的只是每一组权重的相对大小， 一组权重可以表示一个负指数或指数函数。如果一组权重有的大于 1，我们可以把它们看作一个等比例放大了的真值函数。它同样可以用于语义贝叶斯预测(见(3))。

SVB 使用迭代方法，无需考虑梯度。理论上，在学习阶段，我们可以考虑用高斯信道混合模型【27】实现有限波尔茨曼机(RBM)的功能，从而简化运算；在分类阶段，我们令 s->∞得到的函数$P(y_j|x)(j = 1, 2, …, n)$就表现为矩形，成为分类函数。实际效果如何，有待检验。

深度学习领域已经有了很多有效方法，SVB 和这些方法结合，有希望取得更好结果。

6. 结论

SVB 来自信息率逼真函数的参数解，其中变分和迭代方法来自香农等人对信息率失真函数的研究。SVB 和 VB 一样，用来求给定观察数据和某种约束时的隐含变量(的概率分布)。但是 SVB 允许我们使用多种学习或约束函数——包括似然、真值、隶属、相似、copula 密度…函数——作为约束条件。SVB 和 VB 在理论上主要差别是它们的优化准则不同。VB 使用最小自由能准则，而 SVB 使用最大信息效率准则，即最大似然准则加最大熵准则。SVB 求解隐含变量的计算更简单(见(28)和(29))，因为对于同样的任务，SVB 不需要指数和对数运算。

本文通过实验数据支持了上述结论。4.1 节显示：混合模型收敛过程中，信息效率 G/R 不断增大，而不是完整数据对数似然度 Q 或负的自由能 F 不断增大。4.2 节显示：使用 SVB 的 EnM 算法比

EM 算法收敛更快。4.3 节显示了：我们能使用 SVB 在给定误差范围限制时进行数据压缩。4.4 节显示了语义信息测度和 SVB 可用于给定目标范围的约束控制和强化学习，便于我们在控制的合目的性和控制效率之间做权衡。

如何把 SVB 用于深度学习，还有待进一步研究和实验。

参考文献

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

[2] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes." [Online]. Available: https://arxiv.org/abs/1301.6676

[3] Wikipedia, "Variational Bayesian methods," [Online]. Available: https://en.wikipedia.org/wiki/Variational_Bayesian_methods

[4] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants." In: *Learning in Graphical Models*, edited by Michael I. Jordan, PP. 355–368, MIT Press, Cambridge, 1999.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," [Online], Available: https://arxiv.org/abs/1312.6114

[6] K. Friston, "The free-energy principle: a unified brain theory?" *Nat Rev Neurosci*, vol. 11, no. 2, pp. 127–138, Feb. 2010, doi: 10.1038/NRN2787.

[7] M. S. Yellapragada and C. P. Konkimalla, "Variational Bayes: A report on approaches and applications," [Online]. Available: https://arxiv.org/abs/1905.10744

[8] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Information-theoretic regularization for learning global features by sequential VAE," *Mach Learn*, vol. 110, no. 8, pp. 2239-2266, 2021, doi: 10.1007/s10994-021-06032-4.

[9] S. Ding, W. Du, L. Ding, J. Zhang, L. Guo, and B. An, "Robust Multi-agent Communication with Graph Information Bottleneck Optimization," *IEEE Trans Pattern Anal Mach Intell*, vol. 46, no. 6, pp. 3096-3107, 2023, doi: 10.1109/TPAMI.2023.3337534.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 379–429, 623–656, 1948.

[11] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.* vol. 4, 142–163, 1959.

[12] C. G. Lu, "A generalization of Shannon's information theory," *International Journal of General System*, vol. 28, no. 6, pp. 453–490, 1999.

[13] C. Lu, "Semantic information G theory and logical Bayesian inference for machine learning," *Information*, vol. 10, no. 8, p. 261, Aug. 2019, doi: 10.3390/INFO10080261.

[14] T. Berger, Rate Distortion Theory, Enklewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[15] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, 1998.

[16] J. P. Zhou et al., Fundamentals of information theory, Beijing, China: People's Posts and Telecommunications Press, 1983.

[17] C. Lu, "Meanings of generalized entropy and generalized mutual information for coding,"(in Chinese) *J. of China Institute of Communication*, vol. 5, no. 6, pp. 37-44, June 1994.

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

[19] C. Lu, "The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning," *Philosophies*, vol. 5, no. 4, p. 25, Oct. 2020, doi: 10.3390/philosophies5040025.

[20] C. Lu, "Using the semantic information G measure to explain and extend rate-distortion functions and maximum entropy distributions," *Entropy*, vol. 23, no. 8, Aug. 2021, doi: 10.3390/E23081050.

[21] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitrechnung*; Ergebnisse Der Mathematik (1933); translated as *Foundations of Probability*; Chelsea Publishing Company: New York, NY, USA, 1950.

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

[23] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–53,1965.
[24] L. A. Zadeh, "Probability measures of fuzzy events," *J. of Mathematical, Analysis and Applications,* vol. 23, pp. 421-427, 1962.
[25] D. Davidson, "Truth and meaning," *Synthese*, vol. 17, no. 3, pp. 304-323, 1967.
[26] K. Popper, Conjectures and Refutations, 1st ed.; London and New York: Routledge, 2002.
[27] C. Lu, "Reviewing evolution of learning functions and semantic information measures for understanding deep learning," *Entropy,* vol. 25, no. 5. 2023. doi: 10.3390/e25050802.
[28] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," [Online]. Available: https://arxiv.org/abs/1807.03748.
[29] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, S., Y. Bengio, A. Courville, and R. D. Hjelm, "MINE: Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 1-44, https://doi.org/10.48550/arXiv.1801.04062.
[30] S. Kullback and R. Leibler, "On information and Sufficiency," *Annals of Mathematical Statistics,* vol 22, pp. 79–86, 1951.
[31] S. E. Fienberg, "When Did Bayesian Inference Become "Bayesian?" Bayesian Analysis, vol. 1, no. 1, pp. 1-37, 2003.
[32] Wikipedia, Copula, [online], Available: https://en.wikipedia.org/wiki/Copula_(probability_theory)
[33] J. Ma and Z. Sun, "Mutual information is copula entropy," *Tsinghua Sci. Technol.* Vol. *16*, no. 1, pp. 51–54, 2011.
[34] P. Krupskii, H. Joe, "Approximate likelihood with proxy variables for parameter estimation in high-dimensional factor copula models, " *Statistical Papers,* vol. 63, pp. 543–569, 2022.
[35] G. Oddie, "Truthlikeness," in The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta, Ed. [online], Available: https://plato.stanford.edu/archives/win2016/entries/truthlikeness/
[36] T. M. Cover and J. A. Thomas, "Elements of Information Theory," John Wiley & Sons: New York, USA, 2006.
[37] C. Lu, "Understanding and accelerating EM algorithm's convergence by fair competition principle and rate-verisimilitude function," [online]. Available: https://arxiv.org/abs/2104.12592
[38] 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, no. 1, pp. 1–38, 1997.
[39] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Networks*, vol. 11, no. 2, pp. 271-282, 1998.