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

改进最小自由能原理为最大信息效率原理

鲁晨光

摘要：Friston 提出最小自由能 (MFE) 原理，它基于变分贝叶斯(VB)方法。该原理继承了进化系统论的基本思想，但是它强调强调大脑和行为和环境相互协调，增加有序，对抗熵增大。但是 VB 和 MFE 原理存在两个问题：1)作者发现，用 VB 优化隐含变量的时候，变分自由能(VFE)*F* 未必持续减小；2)VFE 和物理学自由能的概念有矛盾。作者早在 30 年前提出语义信息 G 理论和 *R*(*G*)函数(*R* 是给定语义互信息 *G* 时的最小香农互信息)——它是信息率失真函数的推广。本文基于 *R*(*G*)函数的研究提出语义变分贝叶斯(SVB)和最大信息效率 (MIE)原理，用以解决 VB 和 MFE 原理的两个问题。文中简要介绍了语义信息 G 理论和 *R*(*G*)函数，说明优化隐含变量时 *R–G=F-H*(*X*|*Y*)而不是 *F* 持续减小——这一结论得到实验结果的支持。实验还显示了 SVB 用于求隐含变量和主动推断同样可靠，但是计算更简单。文中还通过局域非平衡和平衡系统中信息、熵和自由能之间的关系，说明香农互信息信息相当于自由能的增量，语义互信息相当于 Exergy 的增量，VFE 相当于物理学熵。这样我们就好解释：人类通过获取更多信息和自由能抵抗熵增大。文中还以色觉和美感机制为例说明，MIE 原理和生物学功利主义相互支持。

关键词：变分贝叶斯，最小自由能原理，自由能，物理学熵，香农信息，语义信息，信息率失真函数，EM 算法，主动推断

1 引言

早在 1993 年，Hinton 等人【1,2】提出使用最小自由能作为优化准则，优化神经网络学习，取得重大突破。这种方法后来被发展为变分贝叶斯 VB 方法【3,4】，在机器学习(包括强化学习)领域都有很多成功的应用【5】。Friston【6,7】推广变分贝叶斯(VB)的应用到神经科学，把最小自由能准则发展为最小自由能 (the Minimum Free-Energy Principle ，MFE)原理，希望它成为一个统一的脑和生物行为的科学理论 。该理论整合了预测编码、感知预测、主动推断、信息、随机动力系统等概念【7-9】。主动推断包括对环境的适应(比如选择生态位)和干预(比如约束控制)。这使得：Friston 的理论不同于《熵——一个新的世界观》【10】持有的消极观点，它继承了现存的进化系统论【11,12】中积极的基于熵的世界观，强调生物预测、适应和改造环境，实现自组织和有序，对抗熵增大。

Friston 的理论受到广泛关注并引起深入思考【13，14】。应用也相继出现【15,16】。但是也有人提出不同意见。有人认为自由能就是负熵【14】，是生命需要的，最小化自由能意味着死亡【17】。有人认为，MFE 原理作为工具是有价值的，但是作为一个普遍原理或可证伪科学规律，有待商榷。

Silverstein 和 Pimbley 早在 1990 年发表的文章【18】中就在文章标题中使用了“最小自由能方法”，这种方法可以作为另一种自由能方法的代表。其目标函数被定义为平均误差能量表示和信号熵表示。Gottwald 和 Braun 的文章《两种自由能和贝叶斯革命》【19】争论说，Friston 等人的自由能只是两种种的一种。另外一种是使用其他目标函数的最大熵方法。这些目标目标函数要么等于奖励函数加上熵函数，需要最大化；要么等于平均损失减去熵函数，需要最小化。两种自由能方法都是很有意义的。但是，在本文作者看来，MFE 原理和

最大熵原理【20,21】有明显不同，理由是：在 MFE 原理中，

1） 主观预测和客观事实(或客观事实和主观目标)是双向靠近，而且是动态的；

2） 用信息和熵测度表示损失函数，这样，主动推断问题就成为样本学习问题的逆问题。

3） 存在多任务协调和权衡，需要求隐含变量。

看来 MFE 原理比最大熵原理更能解释生物的主观能动性。但是，在本文作者看来，MFE 原理也是不完善的，最小自由能准则本身就有问题。作者 30 年前就推广香农信息论，得到一种语义信息论【22-24】，其中语义互信息公式是 $I(X;Y_\theta)=H(X)-H(X|Y_\theta)$, 其中 $H(X|Y_\theta)$是语义后验熵或交叉后验熵。粗略说来，$H(X|Y_\theta)$就是自由能 F；最小化自由能就是最大化语义互信息。后来作者推广香农信息论得到的理论为语义信息 G 理论【25,26】(或简称为 G 理论，G 意思是推广)。作者早就知道在香农信道匹配语义信道的时候 $H(X|Y_\theta)$未必单调减小。作者也研究过混合模型和 EM 算法【25】。在试验中，看到 $H(X|Y_\theta)$和 F 也未必随混合模型收敛持续减小。虽然使用 VB 能使混合模型收敛【2,3】，但是理论解释并不正确。这是 VB 和 MFE 原理的第一个问题。

第二个问题是，VFE 和物理学中的熵和自由能究竟有怎样的关系？显然矛盾的是：在热力学中，自由能是能做功的能量，通常需要最大化【27】；系统内能一定时，自由能随熵增大而减小。如果我们主动最小化自由能，那不是顺应熵增大趋势？——如 Martyushev【17】指出的。为此，我们需要知道，VFE 或 $H(X|Y_\theta)$在热力学系统中究竟代表什么？热力学系统中，信息(比如温度提供关于分子能量的信息)、熵和自由能究竟有怎样的关系？

Haken 和 Portugali [15,16]探讨了这一问题，并得到结论：信息论和协同学原理（包括 MFE 原理）能帮助我们回答薛定谔的“生命是什么”。而 Ben-Naim【28-30】在分析热力学系统中两个分子之间的互信息之后，否定用熵和信息理论解释生命系统。本文作者赞成前者。

笔者早在 1993 年【23】就通过局域平衡自由能公式得出结论：香农互信息就相当于局域非平衡系统中的自由能增量，语义互信息相当于局域平衡系统中的自由能增量。另外，作者通过推广香农的信息率失真函数 $R(D)$ 得到信息率逼真函数 $R(G)$ 【23,25】(R 是给定语义互信息 G 时的最小香农互信息， G 反映逼真度)。求解 $R(G)$函数的方法可谓语义变分贝叶斯方法(SVB)【28】，SVB 能解决 VB 要解决的问题，但是并不总是最小化变分自由能(VFE) F，而是最小化 $R-G=F-H(X|Y)$($H(X|Y)$是香农后验熵)。因为最小化 $R-G$ 也就是最大化信息效率 G/R。所以笔者提出最大信息效率 (MIE) 原理。MIE 原理能克服 VB 和 MFE 原理中存在的两个问题，并且算法更简单。G 理论在机器学习中的应用包括：多标签学习，不可见实例最大互信息分类，混合模型【25】，贝叶斯确证【26】，语义压缩【29】等。这些应用的合理性增加了 G 理论的合理性。

本文的动机是：澄清 VB 和 MFE 原理中存在的上面两个问题，并改进它们。

本文目的是：提供最小自由能原理的改进版本或替代选择——最大信息效率 MIE 原理，从而更好解释生物体(bio-organism)特别是人类，如何高效利用信息和自由能，对抗物理世界的熵增大趋势。

本文贡献：

- 从数学视角澄清 VB 和 MFE 原理中存在的理论和实践不一致问题。
- 从 G 理论和统计物理学视角阐明香农互信息，语义互信息，变分自由能，与物理学熵和自由能之间的关系。
- 通过几个混合模型的例子(包括 Neal 和 Hinton 使用的例子)，证明混合模型收敛过程中 VFE 可能增大，并解释增大的原因。

本文遵循 Popper 的思想——科学理论在本质上是猜想，其价值在于信息【33】(p.294)。如果一个假设逻辑越小并且越能经得起经验事实的检验，那么他提供的信息量就越大。据此，最小自由能和最大信息效率原理都是有待进一步确证【26】的猜想，只要一个猜想解释观察

到的经验现象比其他猜想好，它就值得保留。

本文使用信息理论方法而不是统计理论方法描述和解决问题。比如使用样本分布 $P(x)$ 和交叉熵优化似然函数，而不是用样本和似然度优化似然函数。SVB 中的最小信息差迭代方法来自香农等人求解信息率失真函数用到的迭代方法【30-32】。

本文缩写和解释见附录 A。

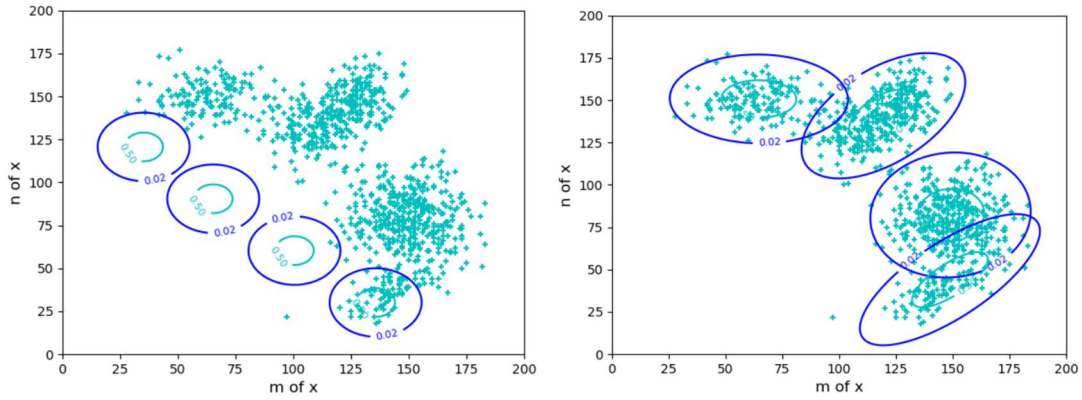
2. 从机器学习的两个典型任务谈起

2.1 羊群聚类——混合模型问题

为了直观解释 VB 要完成的任务。我们以羊群聚类和放羊为例。本节说明羊群聚类问题(参看图 1)。

假设有几群羊分布在边界模糊的草地上，我们能观察到密度分布(即单位面积上羊的数量比例)是 $P(x)$，也知道有 n 群羊，每群羊的分布有某种规律，比如呈高斯分布。我们可以建立一个混合模型：$P_\theta(x)=P(y_1)P(x|\theta_j)+P(y_2)P(x|\theta_j)+\ldots$ 然后用最大似然准则或最小交叉熵准则优化混合比例 $P(y_1)$, $P(y_2)\ldots$ 和模型参数。

通常使用期望最大(Expectation and Maximization， EM)算法【33,34】求解混合模型. 这一算法非常巧妙，它能自动调整不同部件即似然函数(图 1 中圆圈所示)，让每个部件覆盖一组实例(羊群)，而且能提供恰当比例 $P(y_j)$, j=1,2,3,4。$P(y)$是混合模型部件的比例，又叫隐含变量的概率分布。我们有时也称 $P(y)$是隐含变量。



(a) (b)

图 1 以羊群聚类为例解释高斯混合模型。(a)迭代开始，预测和实际相差很远；(b)迭代收敛，预测和实际符合。其中 x=(m,n)是二维的. 聚类使用 EM 算法。

本文附录 B 包含产生本文各图片的 Python 源代码下载信息。

EM 算法不理想在两方面：1)EM 算法的收敛证明一直存在问题【25,34】，以至于出现盲目改进；2)$P(y)$有时候收敛很慢，且在似然函数不变时不好求解 $P(y)$。研究者专为求解隐含变量发展出变分贝叶斯(VB)。这不仅是因为要改进 EM 算法【2,3】，也因为其它任务也需要求解隐含变量【5】。

混合模型属于机器学习中无监督学习，很有代表性。有限波尔茨曼机和深度学习预训练中都有它的影子。利用优化的似然函数，我们可以做概率预测和分类。

2.2 赶羊到牧场——约束控制和主动推断问题

赶羊到牧场是随机事件约束控制问题，也是主动推断(active inference)问题，涉及强化学习。这时候香农互信息反映控制复杂性，我们需要最大化合目的性(或效用)并且最小化香

农互信息即控制成本。香农互信息最小化等价于 X 的后验熵 $H(X|Y)$最大化.

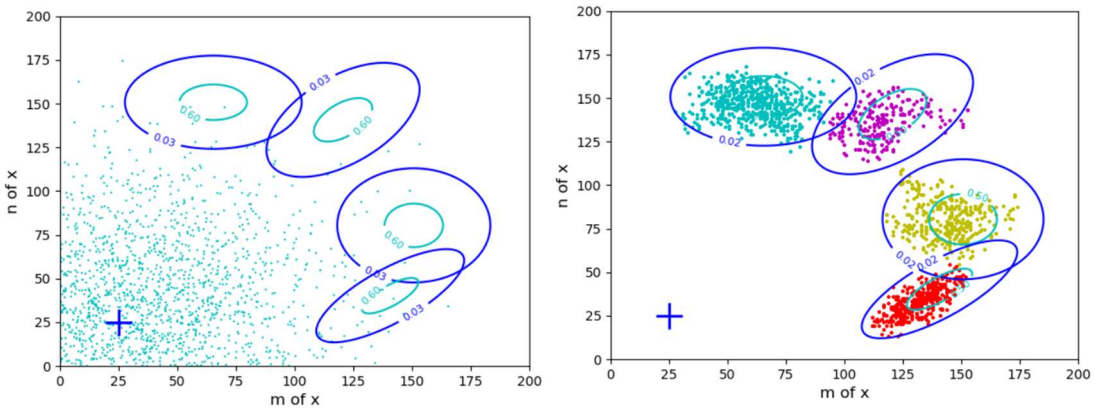


图 2 以放羊为例说明不确定事件的约束控制(约束条件是模糊范围)。(a)控制开始；(b)控制结束。

图中圆圈表示控制目标；图 2a 中点表示初始的羊群密度分布 $P(x)$。通常控制目标或约束条件有两种：

1） 圆圈反映概率分布 $P(x|\theta_j)(j=1,2,3,4)$, 给定 $P(x)$和 $P(x|\theta_j)$, 求解香农信道 $P(y|x)$和羊群比例 $P(y)$. 要求 $P(x|y_j)$接近 $P(x|\theta_j)$，并且 $P(y)$能最小化控制成本。

2） 圆圈反映模糊范围(x 在隶属度等于 1 的地方最好，离开越远误差越大)。从 $P(x)$和模糊范围可以得到 $P(x|\theta_j)$，其他同样。

图 2 中圆圈表示约束范围。羊群聚类(见图 1)和放羊(见图 2)的区别：羊群聚类时，$P(x)$是固定的，羊群数量比例是客观的；放羊时，$P(x)$转移到目标区域，羊群数量比例是按控制成本最低原则调节的。比如在聚类时，中间两群羊比例较大；而在放羊时，两边的两群比例较大，因为把它们赶到目的地容易些。另外，图 2b 中羊群中心偏离目标中心，这是因为羊群中心离初始中心“+”近时，控制成本低些。

我们也可以提高约束强度，让羊群向理想位置集中。但是也要考虑控制成本，羊群聚类和放羊都需要求解隐含变量 $P(y)$.

虽然 SVB 和 MIE 原理是 VB 和 MFE 原理的改进版本，但是为了便于理解 VB 和 MFE 原理，本文下面先介绍语义信息 G 理论, SVB 和最大信息效率原理。

3 语义信息 G 理论和最大信息效率原理

3.1. P-T 概率框架和语义信道

为什么我们需要新的概率框架？原因是：

- P-T 概率框架【26，29】允许我们除了用似然函数，还可以用真值、隶属、相似、失真等函数作为约束条件，求解隐含变量。
- 一个假设或标签，比如“成年人”，同时有两种概率：标签的逻辑概率，即标签代表的集合的概率，这是 Kolmgorov 定义的集合的概率【35】，它不是归一化的；另一个是标签被选择的概率，这是 Mises 定义的概率【36】，它是归一化的。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【37】。

香农的概率框架可以表示为一个三元组($\boldsymbol{U}$, $\boldsymbol{V}$, P). 其中 $\boldsymbol{U}=\{x_1, x_2, \cdots\}$ 和 $\boldsymbol{V}=\{y_1, y_2, \cdots\}$ 是两个随机变量 X 和 Y 的论域。P 是 Mises 定义的统计概率，可理解为相对频率或其极限。注意：统计概率是用“=”定义的， 比如 $P(x_i)=P(X=x_i)$，$P(y_j|x_i)=P(Y=y_j|X=x_i)$.

我们在香农的概率框架基础上定义 P-T 概率框架，它可以表示为一个五元组 ($\boldsymbol{U}$, $\boldsymbol{V}$, P,

$\boldsymbol{B}$, T)。其中 $\boldsymbol{B}$ 包含 $\boldsymbol{U}$ 中所有模糊子集，T 是一个模糊子集的概率。对于每个模子集 θ_j，$\boldsymbol{V}$ 中有一个 y_j，它是 θ_j 的标签，两者之间一一对应。这样 y_j 可看做一个谓词，即 $y_j(\cdot)$=”$\cdot$ 在 θ_j 中”。y_j 的逻辑概率就是 X 属于 θ_j 的概率，即 $T(y_j)=P(X\in\theta_j)$。注意逻辑概率是用“$\in$”定义的，它就是 Kolmogorov 定义的集合的概率【35】。同理，$y_j(x_i)$=“x_i 在 θ_j 中”是一个命题，命题的真值就是 x_i 在 θ_j 中的隶属度，也是 y_j 的条件逻辑概率，即：

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

根据 Davidson 的真值条件语义学【39】，$T(y_j\,|\,x)$反映了 y_j 的语义。一个标签的逻辑概率不等于它的统计概率，最极端的例子是：一个永真句的逻辑概率是 1，而它的统计概率(被选择的概率)接近 0. 我们有 $P(y_1)+P(y_2)+\ldots+P(y_n)=1$，但是可能有 $T(y_1)+T(y_2)+\ldots+T(y_n)>1$.

根据上面定义，我们有 y_j 的逻辑概率

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

后面将看到，它就是统计物理学中的配分函数和机器学习中的正则化项。θ_j 也被看作是模型参数。我们可以把 $T(\theta_j\,|\,x)$ 和 $P(x)$放进贝叶斯公式得到语义概率预测公式【25】：

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

$P(x\,|\,\theta_j)$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 θ_j 和 y_j 是绑定的。我们称上面公式是语义贝叶斯公式。

就像一组转移概率函数 $P(y_j\,|\,x)$(j=1, 2, …)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。后面我们有时省略(j=1,2,…)和(i=1,2,…)。

真值函数和失真函数之间的关系是【29】：

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

3.2 推广香农信息测度到语义信息 G 测度

香农互信息是：

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

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：

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

其中 $H(Y_\theta\,|\,X)$是模糊熵，等于平均失真 $\bar{d}$,因为根据等式(4)，有：

$$H(Y_\theta\,|\,X)=-\sum_j\sum_i P(x_i,y_j)\log T(\theta_j\,|\,x_i)=\bar{d}\,, \tag{7}$$

$H(X|Y_\theta)$是 x 的语义后验熵：

$$H(X|Y_\theta)=-\sum_j\sum_i P(x_i,y_j)\log P(x_i|\theta_j). \tag{8}$$

粗略说来，它就是变分贝叶斯方法和最小自由能原理中变分自由能 F。它越小，语义信息量越大。$H(Y_\theta)$是语义熵:

$$H(Y_\theta)=-\sum_i P(y_j)\log T(\theta_j), \tag{9}$$

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。$P(x|\theta_j)$和 $P(x|y_j)$可能不同，反映信息需要事实检验。容易证明，最大语义互信息准则等价于最大似然准则，并且类似于正则化最小误差平方(RLS)准则。语义熵就是正则化项。

当 $Y=y_j$时，语义互信息就变成语义 KL 信息或广义 KL 信息:

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

当 $P(x|\theta_j)=P(x|y_j)$, 语义 KL 信息最大。令 $T^*(\theta_j|x)$的最大值是 1， 将用公式(3)中的 $P(x|\theta_j)$ 代入 $P(x|\theta_j)=P(x|y_j)$，我们可以得到优化的真值函数【24】。

$$T^*(\theta_j|x)=\frac{P(x|y_j)}{P(x)}\Big/\max_x\left(\frac{P(x|y)}{P(x)}\right)=\frac{P(y_j|x)}{\max_x(P(y_j|x))}. \tag{11}$$

用上式求解 $T^*(\theta_j|x)$要求样本分布是连续且平滑的。否则。我们要用下式得到 $T^*(\theta_j|x)$：

$$T^*(\theta_j|x)=\arg\max_{\theta_j}\sum_i P(x_i|y_j)\log\frac{T(\theta_j|x_i)}{T(\theta_j)}. \tag{12}$$

上面求解 $T^*(\theta_j|x)$的方法被称之为逻辑贝叶斯推断【24】。

如果真值函数变成相似函数，语义互信息就变成估计互信息。比如要计算 GPS 指针和颜色感觉的信息，真值函数就是相似函数，可用高斯函数表示【25】。估计互信息已经被深度学习研究者用于互信息神经估计(MINE)【39】信息噪声对比估计(InfoNCE)【40】。

3.3 信息率逼真函数和语义变分贝叶斯

香农【31】定义：给定信源 $P(x)$, 失真函数 $d(x,y)$和平均失真 $\bar{d}$ 的上限 D, 我们改变信道 $P(y|x)$求香农最小互信息 $R(D)$。$R(D)$就是信息率失真函数，它能指导我们经济地传递信息，特别是用于数据压缩。

现在我们用 $I(x_i;\theta_j)=\log[T(\theta_j|xi)/T(\theta_j)]$取代 $d(x_i,y_j)$, 用 $I(X;Y_\theta)$取代 $\bar{d}$， 用语义互信息的下限 G 取代 D，求最小香农互信息 $R(G)$，$R(G)$就是信息率逼真函数。因为 G 反映因语义预测而节省的平均码长，用 G 代替 D 和缩短码长的目的更加一致，并且 G/R 能反映信息效率。

$R(G)$函数被定义为：

$$R(G)=\min_{P(Y|X):I(X;\theta)\geq G} I(X;Y). \tag{13}$$

我们用拉格朗日乘子法求最小互信息。约束条件除了 $I(X;Y_\theta)\geq G$, 还有

$$\sum_j P(y_j \mid x_i) = 1,\ i=1, 2, \ldots; \qquad \sum_j P(y_j) = 1. \tag{14}$$

所以拉格朗日函数是：

$$L(P(y \mid x), P(y)) = I(X;Y) - sI(X;Y_\theta) - \mu_i \sum_j P(y_j \mid x_i) - \alpha \sum_j P(y_j). \tag{15}$$

用 $P(y|x)$作为变分，令 $\partial L / \partial P(y_j \mid x_i) = 0$，就得到：

$$P^*(y_j \mid x_i) = P(y_j) m_{ij}^s / Z_i,\ \ Z_i = \sum_j P(y_j) m_{ij}^s, \tag{16}$$

其中 $m_{ij}=P(x_i|\theta_j)/P(x_i)=T(\theta_j|x_i)/T(\theta_j)$, Z_i 代替了 μ_i，两者关系是 $\log(1/Z_i)= \mu_i/P(x_i)$.

用 $P(y)$作为变分，令 $\partial L / \partial P(y_j) = 0$ 就得到：

$$P^*(y_j) = \sum_i P(x_i) P(y_j \mid x_i). \tag{17}$$

因为 $P(y|x)$和 $P(y)$相互依赖，我们可以先假设一个 $P(y)$，然后，轮流使用上面两个公式就得到收敛的 $P^*(y)$和 $P^*(y|x)$ (参见【32】(P. 326)). 上面两个公式构成最小信息差迭代(MID 迭代)。有人会问，为什么要通过变分得到式(17)，而不是直接使用式(17)。回答是：如果直接使用式(17)，我们还需要证明: 更新 $P(y)$会减小 $R - G$。

一个 $R(G)$函数的参数解如下(参看图 3)：

$$\begin{aligned} G(s) &= \sum_i \sum_j P(x_i) P^*(y_j \mid x_i) I_{ij} = \sum_i \sum_j I_{ij} P(x_i) P^*(y_j) m_{ij}^{\ s} / Z_i, \\ R(s) &= sG(s) - \sum_i P(x_i) \log Z_i, \quad Z_i = \sum_k P(y_k) m_{ij}^{\ s}. \end{aligned} \tag{18}$$

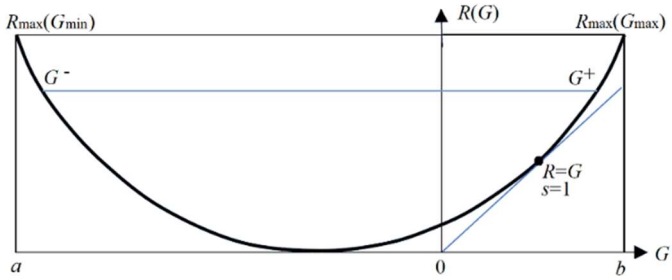


图 3. 二元通信的信息率逼真度函数 $R(G)$。任何 $R(G)$函数都是碗状的，其中有一个 $R(G)=G$ 的点 (s=1)。给定 R 有两个反函数 $G^+(R)$和 $G^-(R)$。

任何 $R(G)$函数都是碗状的(有可能不太对称)【24】，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^-意味着发信人有意说谎时，收信者接收的语义信息能小到什么程度.

值得注意的是，给定语义信道 $T(y|x)$时, 让香农信道匹配语义信道，即让 $P(y_j|x) \propto T(y_j|x)$ 或 $P(x|y_j) = P(x|\theta_j)$，并不最大化语义互信息，也不一定最小化香农互信息，而是最小化信息差 $R–G$ 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(23)中 s–>∞, $P(y_j|x)$就仅取值 0 或 1，成为分类函数。

3.4. 语义变分贝叶斯和最大信息效率原理

如果我们用 $P(x|\theta_j)(j=1,2,\ldots)$作为变分， 令 $\partial L/\partial P(x_i|\theta_j)=0$，就得到 $P(x|\theta_j)=P(x|y_j)$ 或 $T(\theta_j|x)\propto P(y_j|x)$. 这正是 LBI 的结果。所以，最小信息差迭代加上 LBI 加等于语义变分贝叶斯(Semantic Variational Bayes, SVB)。用 SVB 求解隐含变量时，约束函数可以是似然函数， 也可以是真值函数(或隶属函数)，相似函数， 失真函数(或损失函数)。

当约束函数是似然函数时，最小信息差迭代公式是：

$$P^*(y_j|x_i)=P(y_j)[P(x_i|\theta_j)]^s\Big/Z_i,\ Z_i=\sum_j P(y_j)[P(x_i|\theta_j)]^s,$$
$$P^*(y_j)=\sum_i P(x_i)P(y_j|x_i). \tag{19}$$

当约束由似然函数变为真值函数的时候，最小信息差迭代变为：

$$P^*(y|x_i)=P(y)\left[\frac{T(\theta_j|x_i)}{T(\theta_j)}\right]^s\Big/Z_i,\ Z_i=\sum_k P(y_k)\left[\frac{T(\theta_k|x_i)}{T(\theta_k)}\right]^s,$$
$$P^*(y_j)=\sum_i P(x_i)P(y_j|x_i). \tag{20}$$

超参数 s 允许我们加强约束，减少边界的模糊性。所以香农信道也可以看作模糊分类函数。模糊性会降低语义信息量，但是会节省香农互信息。

在 G 足够大前提下，使用最小信息差准则，我们就得到最大信息效率准则。把这一准则应用于各种领域，就是使用最大信息效率原理。注意，最大信息效率原理并不是只追求效率，使 $G/R=1$，而是在 G 满足需要的条件下最大化 G/R。这需要我们在最大化 G 和最大化 G/R 之间权衡。信息率失真理论中已经包含最大信息效率思想，但是由于失真和信息之间不好比较大小，无法得到信息效率表达式。

G 作为约束条件，不止在于限制语义信息量，还在于限制何种语义信息，它和人的目的和需求相关。所以，最大信息效率原理和信息价值有关。

4. 最大信息效率原理用于混合模型和约束控制

4.1 混合模型和 EM 算法的新解释(关于羊群聚类)

EM 算法【33,34】通常用于混合模型(聚类)——一种无监督学习方法。

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 $P(x)$, 使得相对熵或 KL 距离 $\mathrm{KL}(P\|P_\theta)$ 接近 0。其中 $P(y)$是待求的隐含变量(概率分布)。

EM 算法首先预设 $P(x|\theta_j)$和 $P(y_j)$。E-step 得到:

$$P(y_j|x)=P(y_j)P(x|\theta_j)/P_\theta(x),\ P_\theta(x)=\sum_k P(y_k)P(x|\theta_k). \tag{21}$$

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

$$P^{+1}(y_j)=\sum_i P(x_i)P(y_j|x_i). \tag{22}$$

和 M2-step：优化似然函数

$$P(x|\theta_j^{+1})=P(x)P(y_j|x)/P^{+1}(y_j). \tag{23}$$

如果是高斯混合模型，就可以用 $P(x)P(y_j \mid x) / P^{+1}(y_j)$ 的期望和标准偏差作为 $P(x \mid \theta_j^{+1})$ 的期望和标准偏差。$P^{+1}(y)$就是前面的 $P^*(y)$。

从 G 理论看，M2-step 是让语义信道匹配香农信道，E-step 是让香农信道匹配语义信道，M1-step 是让信宿 $P(y)$匹配信源 $P(x)$. 重复上面三个步骤可以使混合模型收敛。收敛的 $P(y)$就是要求的隐含变量。根据 $R(G)$函数的推导过程，可见 E-step 和 M1-step 都最小化信息差 $R–G$； M2-step 最大化语义互信息 G。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)$P(y)$收敛比较慢；2)如果似然函数也是固定的，如何求解 $P(y)$?。基于 $R(G)$函数的分析，作者把 EM 算法改进为 EnM 算法【25,29】。EnM 算法中 E-step 不变，M-step 就是 EM 算法中的 M2-step。 另外在两者之间添加 n-step——重复式(21)和(22)求 $P(y)$ 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 $P(y)$匹配 $P(x)$。M-step 只优化似然函数。因为 n-step 之后，$P(y_j)/P^{+1}(y_j)$约等于 1，我们可用下式优化模型参数：

$$P(x|\theta_j^{+1}) = P(x)P(x|\theta_j)/P_\theta(x). \tag{24}$$

如果没有 n-step, 会有 $P(y_j) \neq P^{+1}(y_j)$，$\sum_i P(x_i)P(x \mid \theta_j)/P_\theta(x_i) \neq 1$. 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，我们可以选取较大的 n，直至 $P(y)$收敛。当 n=1 时， EnM 算法就变为 EM 算法。

下面用数学公式证明 EnM 算法收敛。E-step 之后，香农互信息变为：

$$R = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \frac{P(y_j \mid x_i)}{P^{+1}(y_j)}. \tag{25}$$

我们定义：

$$R'' = \sum_i \sum_j P(x_i) \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)} P(y_j) \log \frac{P(x_i \mid \theta_j)}{P_\theta(x_i)}. \tag{26}$$

然后我们能推导出： 在 E-step 之后，有

$$KL(P \| P_\theta) = R'' - G = R - G + KL(P_Y^{+1} \| P_Y), \tag{27}$$

其中 $KL(P||P_\theta)$是 $P(x)$和 $P_\theta(x)$的 KL 离散度，迭代收敛后它接近 0； $KL(P_{Y^{+1}}||P_Y)$即 $KL(P^{+1}(y))||P(y))$，它在 n-Step 之后接近 0.

式(27)可以用来证明混合模型收敛。因为 M2-step 最大化 G, 且 E-step 和 n-step 最小化 $R–G$ 和 $KL(P^{+1}{}_Y)\|P_Y)$, 所以 $H(P\|P_\theta)$能接近 0。上面方法也能用来证明 EM 算法收敛。实验证明，只要样本足够大，n=2 的高斯混合模型都会全局收敛。

如果似然函数不变，EnM 算法就变成 En 算法。En 算法可以用于求约束函数不变时的隐含变量。

4.2 约束控制信息和主动推断(放羊问题)

前面，我们用 G 测度度量通信系统中的语义信息，要求预测 $P(x \mid \theta_j)$符合事实 $P(x \mid y_j)$，约束控制信息相反，要求事实符合目标(如同主动推断)。我们也称约束控制信息为目标导向(goal-oriented)信息和合目的信息【28】。对于动物来说，快感就包含这种信息。

一个祈使句可看作一个控制指令，我们需要知道控制结果是否符合控制目标。结果越符合，信息就越多。控制目标可用似然函数表示，也可以用真值函数表示。下面主要讨论用真值函数表示的目标。比如下面目标：

- “粮食产量接近或超过 7500 kg/公顷";
- "工人工资最好超过 5000 元";
- "人口死亡年龄最好超过 80 岁";
- "电动车巡航距离最好超过 500 公里";
- "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：

$$I(X;a_j/\theta_j)=\sum_i P(x_i\mid a_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}, \tag{28}$$

其中 θ_j 是一个模糊集合，表示控制目标是一个模糊范围。现在 y_j 表示一个控制任务，而 a_j 表示所选择的动作。用 a_j 代替 y_j 一是因为，对于同样的 y_j 我们可能选择不同的 a_j；二是为了符合流行的主动推断中用法。

如果有数个控制任务 $y_1, y_2,\ldots$ 我们可用下面语义互信息表达合目的信息：

$$I(X;A/\theta)=\sum_j P(a_j)\sum_i P(x_i\mid a_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}. \tag{29}$$

其中 A 是取值 a 或 a_j 的随机变量。用语义变分贝叶斯方法，可以优化控制比例 $P(a)$, 确保在给定 $G=I(X; A/\theta)$的情况下，控制复杂性(即香农互信息 R)最小.

目标导向信息可看做是约束控制和强化学习中累积的奖励函数。但是这里目标是一个模糊范围，它表示一个计划、命令或祈使句。优化任务和使用最小自由能原理的主动推断任务类似。对于多任务，需要最小化的目标函数是:

$$f=I(X;A)-sI(X;A/\theta). \tag{30}$$

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。要进一步增大两种信息, 我们可以使用信息差迭代得到 $P(a_j|x)$和 $P(a_j)$, 然后用贝叶斯公式得到:

$$P(x_i\mid a_j)=P(a_j\mid x_i)P(x_i)/P(a_j)=P(x_i)m_{ij}^s\Big/\sum_k P(x_k)m_{kj}^s. \tag{31}$$

对于赶羊任务的两种目标，我们分别用似然函数 $P(x|\theta_j)$和真值函数 $T(\theta_j|x)$表示, 上面公式参考(19)和(20)作相应改变。和最小自由能原理所用 VB 相比，上面方法更加简单，并且能通过 s 改变约束强度。

因为优化的 $P(x|a_j)$是 θ_j 和 s 的函数, 我们记 $P^*(x|a_j)=P(x|\theta_j, s)$。如果实际的相应函数 $P(x|a_j)$只能是某种类型分布，比如高斯分布，我们可以使用参数化的相应函数 $P(x|\beta_j, s)$代替 $P^*(x|a_j)=P(x|\theta_j, s)$，比如用 $P(x|\theta_j, s)$的期望和标准偏差作为 $P(x|\beta_j s)$的期望和标准偏差。

如果 n 个目标变为一个目标 y_j。就有 $R=I(X; y_j)$, $G = I(X; \theta_j)$。最小互信息迭代就变成简单的香农信道匹配语义信道， 即 $P(y_j|x)\propto T(\theta_j|x)$或 $P(x|y_j)=P(x|\theta_j)$, 公式(31)变为：

$$P(x_i \mid a_j) = P(x_i)T(\theta_j \mid x_i)^s \Big/ \sum_k P(x_i)T(\theta_j \mid x_i)^s . \tag{32}$$

5. 信息与物理学熵和自由能之间的关系

5.1 热力学局域非平衡和平衡系统中的熵、信息和语义信息

为了澄清信息和自由能之间的关系，我们讨论热力学系统中的信息、熵和自由能以及它们之间的关系.

吉布斯建立了热力学熵和后来被称为香农熵的熵的联系，Jaynes【20,21】证明了：使用 Stirling 公式 $\ln N! = N\ln N - N$ (在 $N \to \infty$时)，证明了波尔茨曼微观状态数 W（N 个分子的）和香农熵的联系:

$$S' = \mathrm{k}\ln W = \mathrm{k}\ln\frac{N!}{\prod_{i=1}^{G} N_i!} = -\mathrm{k}N\sum_{i=1}^{G} P(x_i \mid T_0)\ln P(x_i \mid T_0) = \mathrm{k}NH(X \mid T_0) \tag{33}$$

其中 k 是波尔茨曼常数，x_i 是第 i 种微观状态（i=1,2,…,G; G 是一个分子的微观状态数），N 是分子数，T_0 是绝对温度(反映粒子的平动动能)。$P(x_i \mid T_0)=N_i/N$ 表示给定能量约束时状态为 x_i 的分子出现的概率分布。在能量约束下最大化 S'就得到波尔茨曼分布:

$$P(x_i \mid T_0) = \exp(-\frac{e_i}{\mathrm{k}T_0}) / Z', \quad Z' = \sum_i \exp(-\frac{e_i}{\mathrm{k}T_0}), \tag{34}$$

其中 Z' 是配分函数，它使 $P(x_i \mid T_0)$归一化。

为求温度和分子能量之间的信息，我们使用 Maxwell-Boltzmann 统计【45】，用能量 e_i 作为 x_i（i=1,2,…,M），用 G_i 表示含有能量 e_i 的微观状态数，即简并度。N_i 是能量为 ei 的分子个数。则 G_i/G 可看做是 x_i 的先验概率 $P(x_i)$，N_i/N 是 ei 的后验概率。于是等式(33)变成:

$$\begin{aligned} S &= \mathrm{k}\ln(N!\prod_{i=1}^{M}\frac{G_i^{N_i}}{N_i!}) = -\mathrm{k}N\sum_{i=1}^{M} P(x_i \mid T_0)\ln\frac{P(x_i \mid T_0)}{G_i} \\ &= -\mathrm{k}N\sum_i P(x_i \mid T_0)\ln\frac{P(x_i \mid T_0)}{P(x_i)} + \mathrm{k}N\ln G = \mathrm{k}N[\ln G - KL(P(x \mid T_0) \,||\, P(x)]. \end{aligned} \tag{35}$$

在能量约束下,系统达到平衡时，等式(34)变为：

$$P(x_i \mid T_0) = P(x_i)\exp(-\frac{e_i}{kT_0}) / Z, \quad Z = \sum_i P(x_i)\exp(-\frac{e_i}{kT_0}) \tag{36}$$

现在我们能把 $\exp[-e_i/(kT_0)]$解释为真值函数 $T(\theta_j \mid x)$, Z 解释为逻辑概率 $T(\theta_j)$, 等式 (36)解释为语义贝叶斯公式。

S 和 S'相差一个常数 c(它不随温度变化)，即 $S'=S+c$, $c=\sum_i P(x_i)\ln G_i$。 如果我们总是忽略 c， $\ln G=H(X)+c$ 就变成 $H(X)$，并且

$$S / (\mathrm{k}N) = H(X) - KL(P(x \mid T_0) \,||\, P(x)) . \tag{37}$$

考虑局域非平衡系统，系统不同区域 $y_j(j = 1, 2, \ldots)$有不同温度 $T_j(j = 1, 2, \ldots)$，$P(x \mid T_0)$ 就变成 $P(x \mid T_j)=P(x \mid y_j)$。于是有

$$\sum_j P(T_j)KL(P(x_i \mid T_j) \,||\, P(x_i) = \sum_j P(T_j)[H(X) - S_j / (\mathrm{k}N_j)] . \tag{38}$$

因为 $P(y_j)=N_j/N$, 整理得：

$$I(X;T) = H(X) - S/(kN). \tag{39}$$

这就是局域非平衡系统中，香农互信息和物理学熵之间的关系。可见物理学熵 S 就相当于 X 的后验熵 $H(X \mid Y)$，物理学中最大熵定律可以等价地表述为最小互信息定律。

根据(35)，(36)和(39)，局域平衡时，

$$I(X;Y)=\sum_j\sum_i P(x_i,y_j)\ln\frac{\exp[-e_i/(kT_j)]}{Z_j^{'}}$$
$$=-\sum_j P(y_j)\log Z_j-\overline{e}/(\mathrm{k}NT)=H(Y_\theta)-H(Y_\theta\mid X)=I(X;Y_\theta), \tag{40}$$

其中 $\overline{e}$ 是平均能量。可见，局域平衡时，最小互香农信息可用语义互信息公式表达。又因为这时

$$S=\mathrm{k}NH(X\mid Y_\theta)=\mathrm{k}NF\ , \tag{41}$$

所以热力学熵正比于语义后验熵(即变分自由能)。

为什么物理学和 G 理论中有相同形式的熵和信息？比如模糊熵在通信系统重等于平均失真 $\overline{d}$，在热力学系统中是平均能量 $\overline{e}$ /(kNT)。原来两者中的熵和信息在本质上都是以某种约束为条件的熵和信息。物理学中是能量约束，而 G 理论中是外延约束，两者之间有简单联系：$T(\theta_j\mid x_i)=\exp[-e_i/(\mathrm{k}T_j)]$。这样看来，在热力学中使用 P-T 概率框架也能带来某种方便。

5.2 信息、自由能、热功效率和信息效率

Helmholtz(赫尔姆赫兹)的自由能公式是：

$$F"=E-TS, \tag{42}$$

其中 $F"$是自由能，E 是系统内能。在封闭系统重，系统从不不平衡变为平衡时，S 会增大，自由能会减小。但是，在开放系统中，系统在从平衡到不平衡时,自由能可能增加。E 不变时，自由能增量和熵的关系是：

$$\Delta F"=-\Delta(TS)=TS-\sum_j T_jS_j=\mathrm{k}NTH(X)-\mathrm{k}N\sum_j T_jH(X\mid Y). \tag{43}$$

比较上式和(38)和(39)，可见香农互信息就相当于局域非平衡系统中的自由能增量。在局域平衡时，香农互信息变为语义互信息。

热力学中，Exergy(㶲)是系统能做工的能量【46】，状态变化时，它被定义为：

$$Exergy=(E-E_0)+p_0(V-V_0)-T_0(S-S_0), \tag{44}$$

其中 $E-E_0$ 是系统内能的增量，p_0 和 T_0 是环境的压力和温度，$V-V_0$ 是体积的增量。考虑局域非平衡和平衡系统中 Exergy 和自由能，这时每个局域体积和温度不变。局域不平衡时，有

$$\Delta Exergy<\ \Delta F" \tag{45}$$

局域平衡时，有

$$\Delta Exergy=\ \Delta F" \tag{46}$$

可见，语义互信息就相当于局域平衡系统中的自由能增量，也就是 Exergy 的增量。可以说，语义互信息小于或等于香农互信息就像㶲Exergy 小于或等于自由能 $F"$。。

我们也可以把 kNT 和 kNT_j 当做单位信息价值【23】，这样，$\Delta Exergy$ 就相当于语义信息的价值。

一般说来，自由能越大越好。只有在使用自由能做功的时候，我们才希望耗费的自由能较少；类似地，只有在耗费香农信息传递语义信息的时候，我们希望耗费的香农信息较少。

语义信息 G 就相当于功 W。$W/|\Delta F"|$反映做功效率；类似地，G/R 反映信息效率。对于随机事件的约束控制，G 反映控制效果，R 反映控制成本，G/R 反映控制效率。我们统称两种情况下的 G/R 为信息效率。

6. 最小自由能原理及其理论和实践不一致问题

6.1 目标函数自由能和求隐含变量的变分贝叶斯

Hinton 和 Camp 在文献【1】中提供了下面公式：

$$F=\sum_j r_j E_j-\sum_j r_j \log\frac{1}{r_j} \tag{47}$$

其中 r_j 是前面的 $P(y_j)$，E_j 是根据 y_j 为 x 编码的编码成本(即重构成本)。F 之所以被称之为“自由能”是因为该公式和物理学自由能公式在形式上相似。在热力学中，最小化自由能可以得到 Boltamann 分布。类似地，我们可以得到

$$\begin{aligned} r_j &= \exp(-E_j)\Big/\sum_j \exp(-E_j),\\ E_j &= H(X,\theta_j)=-\sum_i P(x_i\mid y_j)\log P(x_i,y_j\mid\theta) \end{aligned} \tag{48}$$

Hinton 和 Camp 的变分方法被发展为更一般的变分贝叶斯方法。变分贝叶斯的目标函数的通常表示为【5】：

$$F=\sum_y g(y)\log\frac{g(y)}{P(x,y\mid\theta)}=-\sum_y g(y)\log P(x\mid y,\theta)+KL(g(y)\,||\,P(y)). \tag{49}$$

其中 $g(y)$就是 $P^{+1}(y)$。负的 F 通常又叫证据下界 **Evidence Lower Bound**，记为 $\mathcal{L}(g)$。其中 x 应该是矢量，而且还和 y 相关，所以用语义信息方法表示 F 是

$$\begin{aligned} F&=\sum_i P(x_i)\sum_i P(y_j\mid x_i)\log\frac{P^{+1}(y_j)}{P(x_i,y_j\mid\theta)}\\ &=\sum_j P^{+1}(y_j)\sum_i P(x_i\mid y_j)=\log\frac{P^{+1}(y_j)}{P(x_i\mid\theta_j)P(y_j)}=H(X\mid Y_\theta)+KL(P_Y^{+1}\,||\,P_Y) \end{aligned} \tag{50}$$

对于 EM 算法，在 M1-step 之后或迭代收敛后，$P^{+1}(y)=P(y)$，于是 $F=H(X|Y_\theta)$。所以前面有“粗略说来，$F=H(X|Y_\theta)$”。我们后面假设计算 F 是在 M1-step 之后计算的，于是 $F=H(X|Y_\theta)$。因为 $R=I(X;Y)=H(X)-H(X|Y)$, $G=I(X;Y_\theta)=H(X)-H(X|Y_\theta)$, 所以信息差 $R-G$ 和变分自由能 F 之间有以下关系：

$$R-G=H(X|Y_\theta)-H(X|Y)=F-H(X|Y). \tag{51}$$

为了优化 $P(y)$，也可以用 VB 常用的平均域近似方法，先优化 $P(y|x)$，然后从 $P(y|x)$和 $P(x)$得到 $P(y)$。那样，就等价于使用 $P(y|x)$做变分，最小化下面 $F^{\#}$：

$$F^{\#}=\sum_x P(x)\sum_y P(y\mid x)\log\frac{P(y\mid x)}{P(x,\ y\mid\theta)}. \tag{52}$$

最小化 $F^{\#}$等价于最小化交叉熵

$$H_\theta(X)=-\sum_i P(x_i)\log P_\theta(x_i), \tag{53}$$

也等价于最小化 $KL(P||P_\theta)$和 $R-G$，可使混合模型收敛。

Neal 和 Hinton 曾使用 VB 改进的 EM 算法【2】用以求解混合模型。他们定义

$$F(P(y),\theta)=\mathrm{E}_{P(x,y)}\log P(x,y\mid\theta)+H(Y) \tag{54}$$

为负的变分自由能。为了方便，我们用 F'表示负的变分自由能，即令 $F'=F(P(y),\theta)=-F$。

Neal 和 Hinton 说明，使用增量算法(见【2】中等式(7))，在 E-step 和 M-step 都增大 F'，可使混合模型更快收敛。他们的 M-step 和 EnM 算法中的 M-step 相同，但是 E-step 每次只更新一个 $P(y_j|x)$, 保留 $P(y_k|x)(k\neq j)$不变，类似于 Beal 用的平均域近似法【3】。这实际上也是最小化 $H_\theta(X)$或 $KL(P||P_\theta)$，也是最小化信息差 R-G。

实验表明，EM 算法，EnM 算法，增量算法和 Beal 的 VB-EM 都能使混合模型收敛，但是不幸的是，在混合模型收敛过程中，F'未必持续增大，或者说 F 未必持续减小。这也就是说，VB 的计算结果是正确的，但是理论是不正确的。

有人用完整数据对数似然度 $Q=-H(X,Y_\theta)$的不断增大解释或证明 EM 算法收敛【34,35】，问题同样。Q 和 F'一样，在混合模型收敛过程中可能会减小。

6.2 最小自由能原理

Friston 等人把 VB 首先应用到脑科学，后来推广到生物行为科学， 从而把最小自由能准则发展为(最小)自由能原理【6,7】。

自由能原理使用 μ，a，s 和 η 分别表示四种状态：

- μ: 内部状态; 在 SVB 中它就是似然函数中的 θ。
- a: 主观动作; 在 SVB 中它是 y(用于预测)和 a(用于约束控制)。
- s：感知；就是前面的观察数据 x。
- η：外部状态，是前面的 y 或 $P(x|y)$; 在 SVB 用于约束控制时，对应 η 的是外部相应 $P(x|\beta_j)$, 希望它等于 $P(x|\theta_j)$。

  根据 MFE 原理，优化有两种【6】:

$$\begin{aligned}\mu^* &= \arg\min_{\mu}\{F(\mu,a;s)\},\\ a^* &= \arg\min_{a}\{F(\mu^*,a;s)\}.\end{aligned} \tag{55}$$

前一个等式是优化似然函数 $P(s|\mu)$. 后一个是优化动作选择，即 $P(a)$和 $P(a|s)$。后者又叫主动推断(Active Inference)。这和使用 SVB 优化约束控制时求 $P(a)$和 $P(a|x)$类似。

上面两种或三种状态的组合也成为一种状态，比如 $b=\{s,a\}$表示 Markov blanket 毯，$\pi=\{b,\mu\}$表示特殊状态(particular state)。马尔科夫毯把 μ 和 η 隔离开来，使得我们可以固定一个改变另一个，实现内部状态与外部环境的动态平衡

Friston 有时把 F 解释为意外、惊奇和不确定性，有时把 F 解释为误差。减小 F 就是减少意外和误差。这从 G 理论的角度容易理解，因为 F 是语义后验熵——反映预测后为残差编码的平均码长【1】，当事件符合预测或目的时，不确定性和意外减小，语义信息量增大。又因为 $G=H_\theta(Y)-\bar{d}=H(X)-F$, 在 $H_\theta(Y)$和 $H(X)$固定时，F 和 $\bar{d}$ 同时增大或减小。所以减小 F 就是减小误差。

最有意义的是，Friston 等人试图用最小自由能原理解释生物体如何通过与外界协调，创造有序，对抗熵增大。

6.3 从最大熵原理到最小自由能原理的进步——联系进化系统论

1957 年，Jaynes 提出最大熵原理【20,21】。该理论把物理学熵看作是信息熵的特例，并提供了求解最大熵分布的方法。该方法不仅可用于一定约束条件下系统状态的概率预测，也可用于随机事件的约束控制。和熵增大定律(即热力学第二定律)相比，Jaynes 的最大熵原理是一个重要进步。因为在物理学熵增大定律中，系统是封闭的，约束是固定的。而在 Jaynes 的最大熵原理中，约束是主动的，所以它可以用于人类对自然的干预。但是，最大熵原理还是不足以解释生物系统。一个重要原因是，生物有目的，存在预测和控制的符合和不符合问

题。香农熵和信息都不反映预测或约束控制的符合程度。

而根据最小自由能原，F 越小，反映主观预测 $P(x|\theta_j)$越符合客观事实 $P(x|y_j)$; 另一方面，最小自由能原理中的主动推断(active inference) 使得客观事实 $P(x|y_j)$接近主观目的 $P(x|\theta_j)$。和最大熵原理相比，两者都最大化香农后验熵 $H(X|Y)$, 但是最小自由能原理同时用机器学习中常用的最大似然准则优化目标函数，

自从 Clausius 提出熵增大定律，一直存在两种世界观。一种是消极的基于熵的世界观，如里夫金和霍华德的书《熵：一个新的世界观》【10】所代表的所言，地球是个封闭系统(见的后记)，生物发展和科技进步都带来地球的熵增大（见第五章），人类所能做的是减小自由能和资源的消耗，减缓熵增大过程。这种观点只强调熵增大的必然性，忽视生命就是就是在对抗熵增大的过程中发展的；忽视了人类可以主动控制自然，即在增加自由能的同时，减缓熵增大或减小熵。

另一种是进化系统论(Evolutionary Systems Theory, EST)持有的积极的基于熵的世界观。薛定谔(Schrödinger)【12】早在 1944 年的著作“生命是生命”提问：生命是什么？它如何对抗熵增大从而在物理系统中实现？他因此开启了一种新的研究方向, 这种研究集中于 EST 中。如【11】所言，EST 是一个跨学科领域，建立在费舍尔(Fisher)、赖特(Wright)、霍尔丹(Haldane)和普利高津(Prigogine)等人的研究基础上【47-49】, 并经哈肯，艾根(Eigen)和舒斯特(Schuster)，Kelso 等人得到发展【50-52】。EST 通过一般选择与自组织之间的相互关系来解释动态的、不断进化的系统。

最小自由能原理的重要意义是：它支持建立在 EST 上的积极的 Entropy-based 世界观，并且提供了一个综合的优化准则(包括最大似然准则和最大熵准则)和主动推断方法，用以解释和改进生物对环境的预测和适应。

Haken 的协同论【50】能很好解释鸟和浆果或昆虫和花（双方是互惠的，后者提供营养， 前者帮助后者传播种子或授粉）的协同进化。然而，说这些动物的行为遵循 Jaynes 的最大熵原理是不合适的。由于 MFE 原理蕴含了最大熵（ME）原理，并且也包含了最大似然准则，因此它能够反映动物在识别植物和觅食过程中的不一致程度。动物可以利用这种不一致程度作为反馈信号来调整自身行为。这也是为什么 Haken 和 Portugali[16]肯定 MFE 原理在理解生物行为和生态位巩固中的重要意义。

8.3 和 8.4 节继续讨论这一话题。

6.4 为什么变分自由能 F 在混合模型收敛过程中可能增大？

作者 30 年前就考虑过交叉熵

$$H(X|\theta_j) = -\sum_i P(x_i|y_j)\log P(x_i|\theta_j)\,. \tag{56}$$

的性质。当 $P(x|\theta_j)$接近固定的 $P(x|y_j)$时，$H(X|\theta_j)$会减小。但是反过来, 当 $P(x|y_j)$接近固定的 $P(x|\theta_j)$,时，$I(X;\theta_j)$会增大还是会减小？结论是可能增大，也可能减小。确定的是：KL 离散度 $KL(P(x|y_j)||P(x|\theta_j) = I(X;y_j) - I(X;\theta_j)$会减小。比如 $P(x|y_j)$和 $P(x|\theta_j)$是两个期望相同的高斯分布，但是前者的标准差 d_1 小于后者的标准偏差 d_2，d_1 在迭代过程中会变大，交叉熵也会增大。

表 1 显示的是一个更简单的例子。设 x 有 4 个能取值(参看表 1)。当 $P(x|y_j)$从集中变为分散的时候，$H(X|\theta_j)$会增大。

Table 2. 一个例子说明交叉熵 $H(X|\theta_j)$在 $P(x|y_j)$接近 $P(x|\theta_j)$的时候会增大。

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

$P(x\|y_j)=P(x\|\theta_j)$	0.1	0.4	0.4	0.1	0.2log(10)+0.8log(10/4)=1.72

这个结论推广到语义互信息公式就是：香农信道匹配语义信道时，语义互信息 $I(X; Y_\theta)$ 和语义后验熵 $H(X|Y_\theta) = F$ 不确定增大或减小；确定的是 $R - G$ 会减小。VB 的算法就是让香农信道匹配语义信道，所以在匹配过程中 $R - G = F - H(X|Y)$而不是 F 会持续减小。

在高斯混合模型迭代过程中，影响 F 和 $H(X|Y_\theta)$增大或减小有两个原因：

1) 迭代开始时，$P(x|\theta_j)$和 $P(x|y_j)$的分布范围差别较大，在 $P(x|\theta_j)$接近 $P(x|y_j)$后，F 和 $H(X|Y_\theta)$会减小。
2) 相应的真模型的 $F^\#=H(X|Y)$(即香农条件熵)较大，迭代过程中，F 和 $H(X|Y_\theta)$会变大。

当原因 1 占主导地位时，F 减小；当原因 2 占主导地位时，F 增大。

根据前面分析，假设有一个高斯混合模型有两个部件，初始化参数和真模型的参数比，只有两个标准差较小(参看图 6a)，迭代过程中，F 会持续增大。

不对称的标准差和混合比例也会引起 F 有时增大，见 7.1.1 节。

另外，初始的参数 $\mu_1, \mu_2, \ldots$偏向一边(参看图 7a)，会间接造成原因 2)，导致 F 有时增大，见 7.1.3 节。

7. 实验结果

7.1 证明混合模型收敛过程中是信息差而不是 VFE 单调减小

7.1.1 Neal 和 Hinton 例子——混合比例引起 F' 和 Q 减小

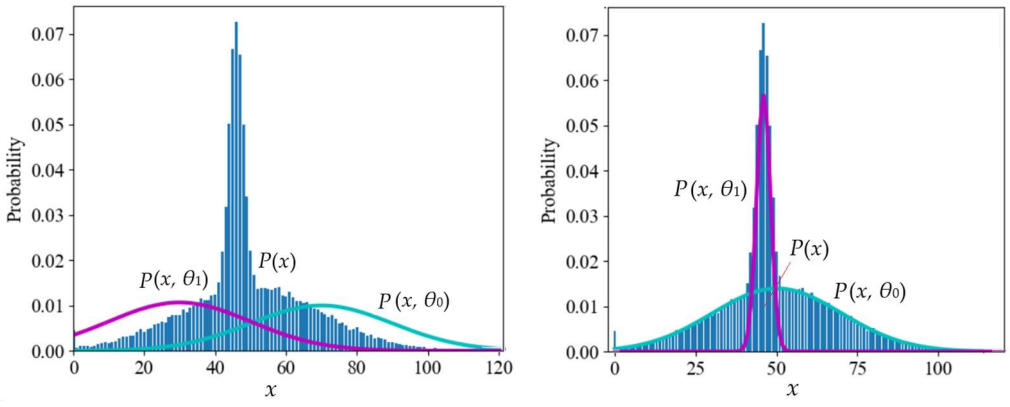
按照流行的观点，在混合模型收敛过程中，$F'= -H(X|Y_\theta)= -F$ 和 $Q = -H(X, Y_\theta)$是持续增大的。但是，实验中经常看到反例。首先我们看 Neal 和 Hinton 的例子【2】(见表 2 和图 4)。表 2 显示了真的和初始的模型参数和混合比例(下面 x 的坐标有所放大，放大公式是 $x=20(x'-50)$ (x'是【2】中的实例坐标)。

表 2. Neal 和 Hinton 的混合模型例子

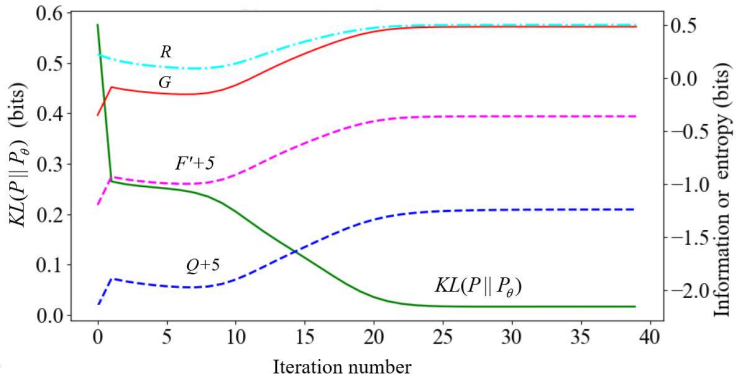
Table 2. Neal and Hinton's example of the mixture model.

	Real Parameters			**Starting parameters**		
	μ^*	σ^*	$P^*(y)$	μ	σ	$P(y)$
y_1	46	2	0.7	30	20	0.5
y_2	50	20	0.3	70	20	0.5

按照上面真模型参数，Q 和 F'在每一轮迭代后都是增大的，只是在某些 E-tep 或 n-step 可能减小【25】。但是，如果把混合比例由 0.7:0.3 换成 0.3:0.7. 就会看到有好几次迭代中，Q 和 F'减小了(参看图 4c)。



(a)迭代开始　　(b)迭代收敛



(c)迭代过程中各测度变化

图 4. 一个混合模型收敛过程(Neal and Hinton　用过的例子，混合比例从 0.3:0.7 变为 0.7:0.3)。F'和 Q 有时减小。

原因是第二个部件的真模型的熵 $H(X|y_2)$比较大，$P(y_2)$增大后，导致 $H(X|Y_\theta)$增大。后来 $H(X|Y_\theta)$还是减小了，也就是 F'增大了，这是因为 $P(x|\theta_j)$接近 $P(x|y_j)$后交叉熵减小。

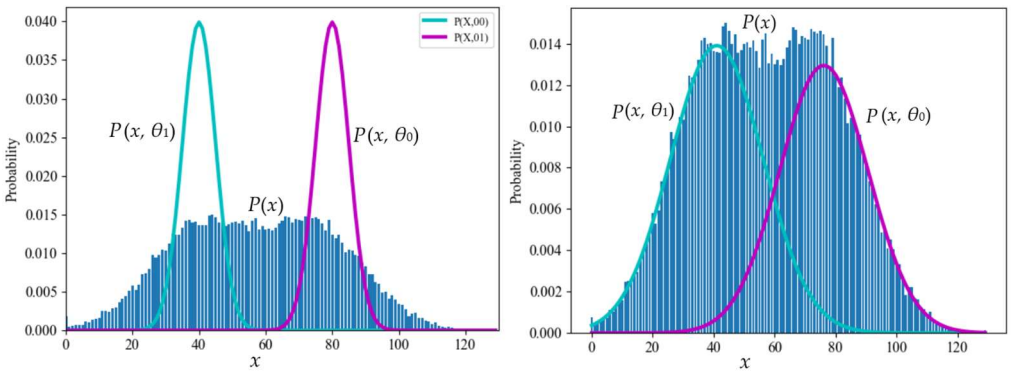
7.1.2 标准偏差引起 F'和 Q 持续减小

表 3 和图 5 显示一个混合模型，初始的两个标准差比真模型的两个标准差小。在迭代收敛过程中，F'和 Q 一直减小(除了开始)。

表 3　一个混合模型的 F' 和 Q 随迭代过程减小

Table 3. A mixture model whose F' and Q decrease with the convergent process.

	Real Parameters			**Starting parameters**		
	μ^*	σ^*	$P^*(Y)$	μ	σ	$P(Y)$
y_1	40	15	0.5	40	5	0.5
y_2	75	15	0.5	80	5	0.5



(a)迭代开始　　(b)迭代收敛

图 5. 一个混合模型的 F' 和 Q 随迭代过程减小

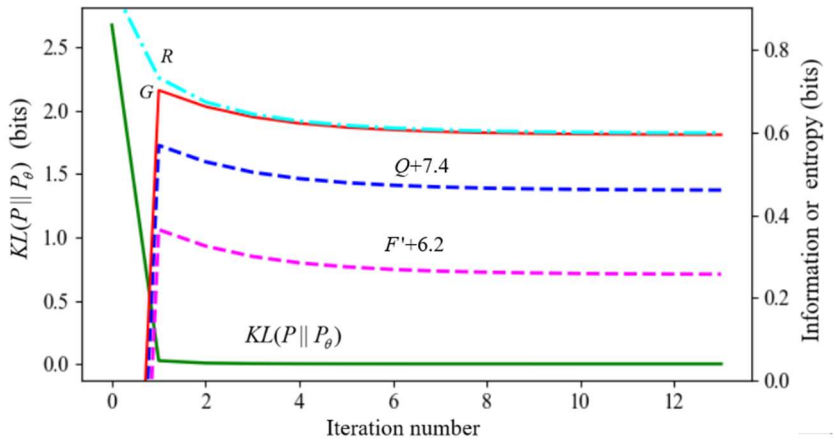
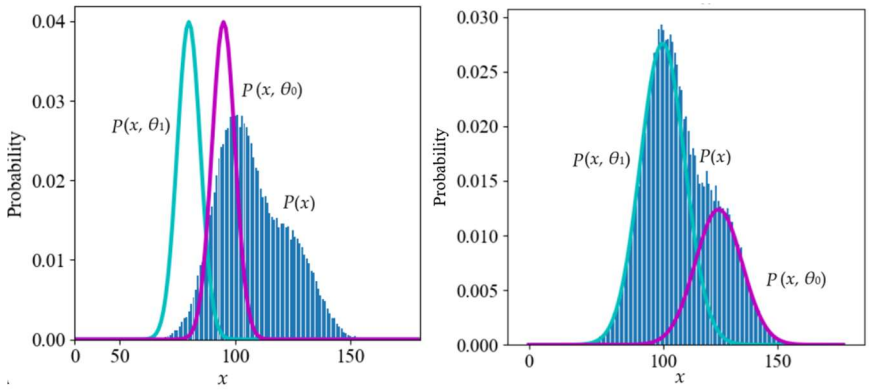


图 5. 一个混合模型，在迭代收敛过程中，F'和 Q 是持续减小的(Against 最小自由能原理)。

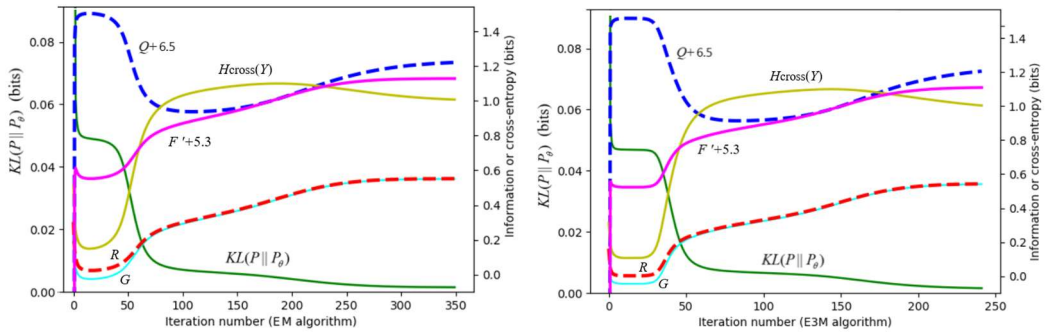
这是一个典型反例——证伪 undermine 最小自由能原理。This is an typical counterexample against the MFE principle.

7.1.3 一个不易收敛的混合模型

图 6 显示的是一个不易收敛的例子，来自【38】。真的模型参数是$(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1))$= (100, 125, 10, 10, 0.7)。为了使收敛更困难，我们设初始的模型参数为$(\mu_1, \mu_2, \sigma_1, \sigma_2, P(y_1))$= (80, 95, 5, 5, 0.5)。实验证明，只要样本足够大，EM, E3M, 增量算法，和 VBEM 都能收敛。但是收敛过程中，只有 R–G 和 KL(P||P_θ)持续减小，F'和 Q 并不持续增大。这个例子说明初始化的 μ_1, μ_2 不适当时，F'和 Q 在迭代过程中也可能会减小。Q 的减小就更明显。



(a) 初始化时的两个部件　(b) 全局收敛时的两个部件



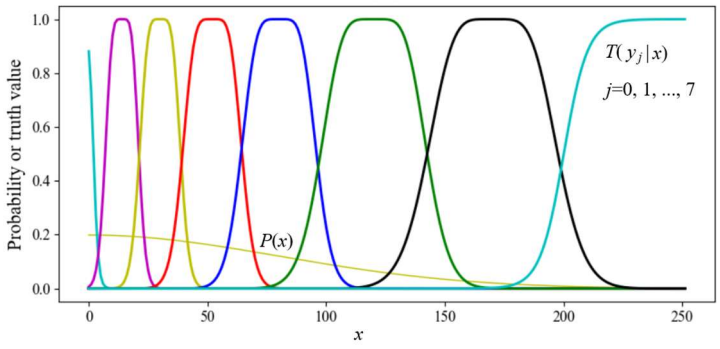
(c)EM 算法迭代过程　　(d)E3M 算法迭代过程

图 6. 一个不易收敛的例子。(a)迭代开始;(b)迭代结束；(c)EM 算法迭代过程；(d)E3M 算法迭代过程，其中 $H_\theta(Y)$=–$\sum_j P^{+1}(y_j)\log P(y_j)$ and F'=Q+ $H_\theta(Y)$= –$H(X|Y_\theta)$= –F.

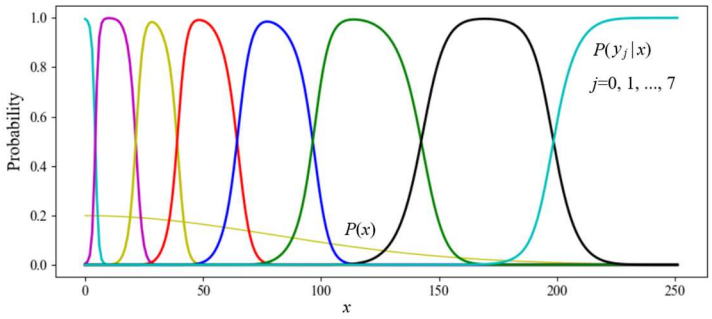
这个例子还说明 E3M 算法比 EM 算法需要较少迭代。但是 E3M 算法每一步花费时间略多。实验表明，前 50 次迭代使用 E3M 算法，其余步骤使用 EM 算法，用时最短。更多讨论见【28】。

7.2 简化的 SVB(En 算法)用于数据压缩

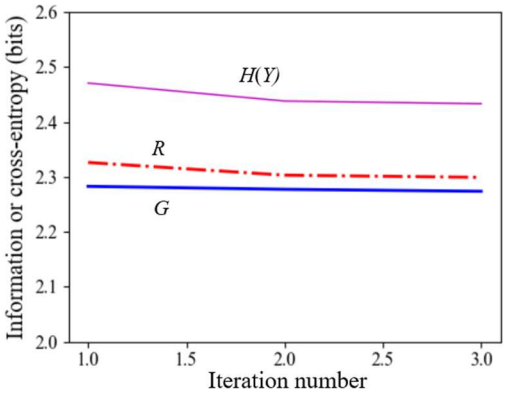
假设要把 8 比特灰度像素(256 个灰度等级)压缩成 3 比特像素(8 个灰度等级)。考虑到亮度低时眼睛灰度分辨率更高，我们使用图 7a 所示真值函数作为约束函数。给定 $P(x)$和 $T(y|x)$, 求信息效率最大的香农信道 $P(y|x)$.



(a)



(b)



(c)

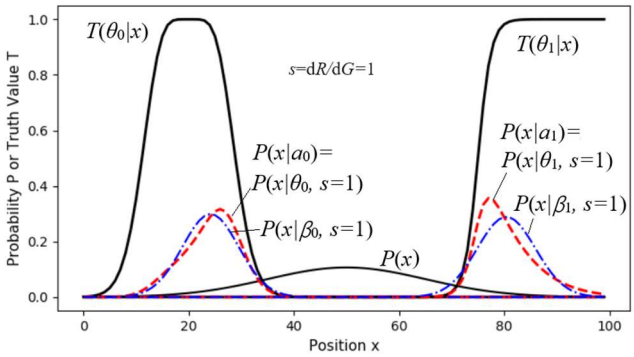
图 7. 用 En 算法求信息效率最大香农信道 $P(y|x)$. (a)给定 8 个范围约束；(b)优化的香农信道；(c)迭代过程中 $H(Y)$和 R–G 持续减小。

重复最小互信息迭代三次， $P(y)$就基本收敛。迭代开始假设 $P(y)$=1/8，熵 $H(Y)$是 3 比特。迭代收敛时 R=2.299 比特，G=2.274 比特， G/R=0.989。说明约 2.3 比特就能传递 3 比特像素信息，信息效率很高。

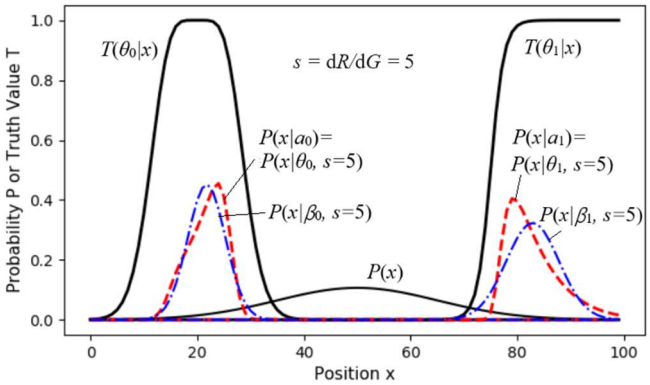
7.3 约束控制实验结果

我们把放羊空间简化为一维空间，其中只有两个牧场(参看图 8)，以便显示控制结果(高斯分布)和目标(范围)之间的符合关系，以及控制结果如何随 s 变化。我们需要根据 $P(x)$和两个真值函数求解隐含变量 $P(a)$.

Figure 9 shows 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**)s=1 时的约束结果



(**b**)s=5 时的约束结果。

图 8. 双任务约束控制优化。

对不同的 s，初始比例 $P(a_0)$=$P(a_1)$=0.5. 使用最小互信息迭代，即 En 算法，可以得到优化的 $P(a_j|x,s)$. 再用下式得到 $P(x|a_j,s)$:

$$P(x_i \mid a_j, s) = P(a_j \mid x_i, s)P(x_i) / P(a_j) = P(x_i)m_{ij}^s \Big/ \sum_k P(x_k)m_{kj}^s . \tag{56}$$

然后使用 $R(G)$函数的参数解求得 $G(s)$, $R(s)$和 $R(G(s))$。图 9a 和图 9b 分别显示了 s=1 和 s=5 时的 $P(x|\theta_j, s)$和 $P(x|\beta_j, s)$。可见 s=5 时，约束更严格，部分模糊边界处的羊群移到更理想位置。图 10 显示了，$s>5$ 时， G 变化很小，说明我们需要在最大合目的性 G 和最大信息效率 G/R 之间权衡，更大的 s 会降低信息效率，是不必要的。

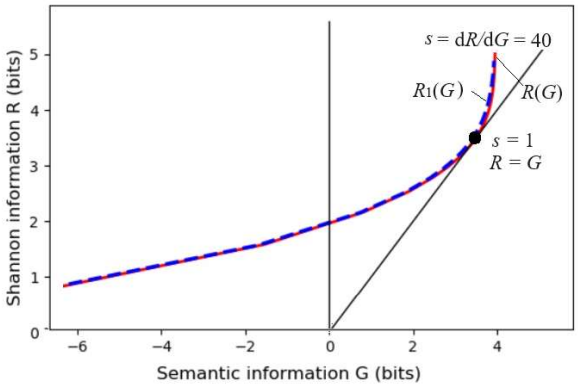


图 9. $R(G)$用于约束控制时权衡效用和信息效率。在 s 从 5 增大时，G 增大很少，说明 s 约等于 5 是较好选择。

其中虚线是 $R_1(G)$函数，它表示，如果控制效果只能是某种分布(比如说高斯分布)时，我们可以用 $P(x|\beta_j, s)$代替 $P(x|\theta_j, s)$， 得到 G and G/R_1。$R_1(G)$和 $R(G)$很接近。

下面假设只单任务约束控制。我们可以令 $R = I(X;y_j)$, $G=I(X; \theta_j)$，优化方法类似，但是不需要求隐含变量。

我们用成年人死亡年龄控制(通过医药条件)作为例子。假设成人死亡年龄的先验概率分布是 $P(x)$，它是正态分布，期望是 $\mu=70$ 标准偏差是 $\sigma=10$. 表示目标的约束函数是 $T(\theta_j|x) = 1/[1+\exp[-0.8(x-80)]$. 任务是：首先找到产生控制结果的分布 $P^*(x|a_j)=P(x|\theta_j, s=1)$, 使得信息效率最大，为 1，然后通过增大 s 增大语义信息 $I(X; \theta_j)$.

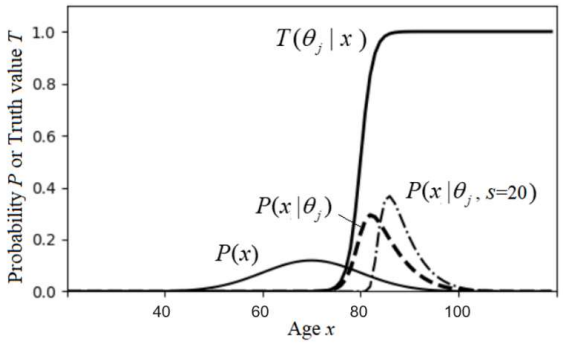


图 10. 认购死亡年龄控制的信息效率。约束函数 $T(\theta_j|x)$表示控制目标。似然函数表示控制结果. $P(x|\theta_j) = P(x|\theta_j, s=1)$ (见虚线)是为了 $G = R$; $P(x|\theta_j, s = 20)$ (见点划线)使得 G 接近其最大值.

实验表明这个例子的 $R(G)$函数和图 9 类似，但是 s 增大时 G/R 下降慢点。s=20 时，平均死亡年龄在 85 左右。这时候 G 就足够大了，s 再大信息效率就会明显降低。

关于本节例子更详细数据，见【28】。

8. 讨论

8.1 VB 和最小自由能原理的两个问题和解决方法

现在我们从前面分析和实验结果看 VB 和最小自由能原理的两个问题。

第一个问题：VB 算法存在理论和实践不一致问题，就是虽然计算结果是对的，但是在混合模型的迭代过程中，变分自由能 F 未必是持续减小的。7.1 节的几个实验结果(见图 5-7)显示：混合模型收敛过程中，只有 $R-G$ 是持续减小的，负的变分自由能 F'和完整数据的对数似然度 Q 也不是持续增大的。

第二个问题是，VB 中“自由能”的用法和物理学有冲突。理由是：5.1 节中分析显示：最小化的 $H(X|Y_\theta) = F$ 正比于局域非平衡系统中的物理学熵(分子能量密度分布决定的熵)而不是物理学自由能 F“. 因为 $H(r)$中 r_j 相当于不同局域中分子比例 $P(y_j)$, 所以 $H(r)$不是分子能量密度分布决定的熵。把 $H(X|Y_\theta)$当作自由能，会引起概念混乱。另外，物理学中，自由能是能做功的能量，越大越好；而 VB 中，自由能越小越好。这是矛盾的。

3.3 和 3.4 节提供了 SVB，它的理论和实践是一致的。6.1 节的实验表明，在所有求隐含变量的迭代过程中，信息差 $R-G$ 都是持续减小的，或者说信息效率 G/R 都是持续增大的。

按第 5 节分析，我们可以把信息看作自由能，这样，我们就好说，生物，特别是人类，通过获取更多信息，保存更多自由能，从而抵抗地球熵增大。

8.2 SVB 和 VB 算法的异同

SVB 和 VB 的主要任务一样：1)使用变分方法，根据观察数据和约束求解隐含变量 $P(y)$；2)优化模型参数或似然函数。不同的是：

- **准则:** VB 和 SVB 在优化模型参数时同样使用最大似然准则，减小语义后验熵；而在优化 $P(y)$时，VB 名义上使用最小自由能 MFE 准则，实际上是使用使混合模型收敛的最小 KL 距离准则(最小化 $KL(P(x)||P_\theta(x))$。和 SVB 使用的最小信息差准则是等价的。
- **变分方法：**VB 只用 $P(y)$或 $P(y|x)$作为变分，而 SVB 轮流使用 $P(y|x)$和 $P(y)$作为变分。
- **计算复杂性：**VB 求 $P(y|x)$用到对数和指数函数【3,5】，计算比较复杂； SVB 中的 $P(y|x)$计算比较简单(对于同样任务，即 s=1 时)。
- **约束方式：**VB 只能用似然函数的对数作为约束(如果用失真函数、相似函数等代替似然函数的对数作为约束，约束项的量纲就不再是比特，因而使用的是两种自由能中的另外一种自由能【18】)。而 SVB 用各种学习函数(包括似然、真值、隶属、相似、失真函数)作为约束时，都转换为语义信息量约束(量纲都是比特)，并且允许用参数 s 增强约束(参看图 9-11)。

因为 SVB 更加兼容最大似然准则和最大熵原理，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。关于语义变分贝叶斯的详细讨论见【28】。

8.3 怎样更好用自由能解释人类对抗地球的熵增大？

最小自由能原理继承了进化系统论的积极世界观，并且强调强调生命体和环境相互协调，增加有序，抵抗熵增大。但是，用最小自由能原理解释人类抵抗地球的熵增大会引起概念混乱，因为在物理世界中，熵减小的同时自由能通常会增大。

《熵：一种新的世界观》得出非常极端的结论：地球是封闭的(见【10】的后记)，生产发展和科技进步总是以环境的更大混乱(熵增大)为代价(见【10】第 5 章)，人类应当控制需求，减少能源和资源消耗，减缓经济发展，甚至恢复农耕文明… 这在理论上是极大的倒退。

其实地球并不是封闭系统，它不断接收阳光带来的自由能。地球通过水、空气或风、植物…保留了太阳带来的自由能，使得地球上生物多样化有序化。人类通过信息效率最大化和做功效率最大化，建造各种电站，种植和养殖各种动植物，从而保留了更多自由能，使地球更加有序。当然，也不排除人类滥用资源、对能源和信息的低效率使用，并使得空气污染导致地球散热困难，从而造成地球熵增大。

造电站和赶羊进牧场是类似的。一方面要求达到目的，另一方面要求成本尽可能低。

人类造水电、风电和光伏电站能大量保存太阳送来的自由能，虽然造电站也要耗费自由能，但是相对于它节省的自由能或者是所做的功，耗费的自由能只占是很小的比例。如果不

造这些电站，这些自由能最终也会以热辐射的形式消失。保留更多自由能是人类能抵抗地球熵增大的原因。

如果使用最小自由能原理，我们就不好说人类因为更多保留更多自由能而增大地球的有序性。

作者提出的最大信息效率原理也继承了 EST 的积极的 Entropy-based 世界观. 除了赞成用最大化自由能 F"、和做功效率($W/\Delta F$")抵抗熵增大，作者还主张用最大化语义信息(G)、信息效率(G/R)抵抗熵增大。另外，作者除了提出不同形式的熵(用于语义通信和约束控制)，还提出资本增值熵(Capital Growth Entropy)用于优化投资组合和信息价值【53】。增值熵有利于解释不确定事件的正反馈。

8.4 美感和色觉机制的进化如何遵循最大信息效率原理？

色觉传递的信息是典型的通信信息，而快感(包括美感)传递的是典型的控制信息。Friston 用最小自由能原理解释人脑和行为如何和环境协调。仿照 Friston 的努力，下面我们讨论人和动物的审美机制和色觉机制如何遵循最大信息效率原理。

8.4.1. 最大信息效率原理和生物学的功利主义如何相互支持

达尔文在物种起源一书中肯定生物学的功利主义【55】（Section 6.7），这种功利主义认为：生物的一切构造都因为对生存有用而存在。他说，如果否定功利主义的是对的，自然选择理论就完全不能立足。

根据生物学功利主义观点，快感和不快感是激励人求利避害的反馈信号，快感功能因为对生存有利而存在。据此，人类觉得苹果好吃，并不是因为苹果本身，而是因为苹果符合人的生存需求，吃苹果的快感含有正反馈信息。从强化学习和主动推断的角度看，快感不快感传递的就是约束控制信息或合目的信息。老虎和狼就不喜欢吃苹果，它们更喜欢血腥味。这说明快感不快感作为反馈信息服务于不同动物的不同生存需要，这符合最大信息效率原理。另外，人的需求满足到一定程度后，快感就会减弱，这促使人在充分满足和降低劳动成本之间权衡，如同 7.3 节的例子。

但是，这个观点收到孔雀华丽羽毛和烟酒和毒品的挑战（它们能引起快感，但是对生存不利）。根据 Popper 的证伪原理，一个反例就能证伪一个全称假说。由于上面反例存在，有人因此断言生物学的功利主义是错的。

现在我们用最大信息效率原理改进生物学功利主义，得出这样的用以辩护的结论：

- 因为人和动物的各种构造总体上对自身生存有利，所以存在（这是模糊大前提）；
- 但是由于通信和控制成本的原因，错误在所难免；实际的构造是基因在最大效用和最大信息效率之间权衡的结果。这就像是说，法律是为了公正，但是由于成本和效率的原因，不能保证它在任何情况下公正。

要检验这一结论，我们可以使用合适的 Bayes 确证测度，通过统计数据得到模糊大前提的确证度。笔者曾提出两个确证测度【26】分别模糊确证大前提和其后件，它们兼容 Popper 的证伪思想，因为它们能确保反例对确证度影响更大。

即使如此改进，雌孔雀审美趣味的起源仍然令人困惑。

8.4.2. 快感机制的信息效率解释和华丽鸟类审美趣味的进化

美感是一种特殊快感，它可以解释为促使人接近对生存有利的对象的反馈信号。怎样解释雄鸟（比如雄孔雀）的华丽羽毛和雌鸟审美趣味的进化？达尔文提出用性选择(即按美选择)理论【55】，而华莱士强烈反对【55】。Fisher(最大似然准则倡导者和 EST 奠基者之一)提出一种后来被称为军备竞赛解释【56】，但是它不能解释军备竞赛的各种奇特方向，比如雄孔雀带有球形的长尾巴。争论仍在继续【57】。

为解释人类和鸟类审美趣味的进化，作者提出需求美学——把美感看作激励人在空间接近对象的正反馈信号（含有控制信息），并据此发现鸟雀华丽外表模拟它们喜爱的食物和环

境，比如雄孔雀的长尾巴模拟蓝莓或浆果树。现在我们可以这样解释鸟类性选择：需求关系选择了鸟类的审美趣味，雌鸟的审美趣味选择了雄鸟的华丽羽毛。由于控制成本或信息效率的原因，我们不能要求雌孔雀的审美机制只对真的浆果树敏感，而对类似的图案不敏感。现在我们可以说，进化论、需求美学和最大信息效率原理相互支持。

Appendix C 包含两个图片（用以说明鸟雀华丽羽毛反映它们喜爱的食物）和相关网页链接。

8.4.3. 色觉机制的信息效率解释和进化

朴素的反映论认为，色觉机制是为了正确反映色光的属性。而根据生物学的功利主义和 Holmholtz 的符号论，色觉只是符号，色觉机制只是为了获取色光中和自己生存有关的信息。最大信息效率原理支持后者。

要精确地用数学表示一种色光，我们需要用无限维矢量表示其频谱分布。Young-Helmholtz 的三色素理论【58】表明，人眼色觉机制把无限维矢量信息转换为三维矢量信息(电视机和显示器就采用三原色信号显示彩色图像)。原因是，人类为了分辨物体，比如山水树木，花果鸟兽，三维矢量信息就足够了。从多维到三维就是为了压缩数据，提高信息效率。并且，不同的动物色觉分辨范围不同。人类和喜欢以水果和花蜜为食的鸟类和昆虫具有和三原色或多原色机制；而捕猎动物大多是色盲，但是它们(比如鹰和青蛙)的视觉对运动物体特别敏感。有些昆虫能看到紫外光，而有些鱼类能看出红外光。有的鸟、昆虫和鱼具有四原色机制。这些都说明，色觉不是简单地反映客观的色光性质，而是高效地获取主体所需要信息。一方面，感官尽可能获取足够的对生存有重要影响的信息(语义或语用信息)，另一方面，感官尽可能少地传递色光的信息(香农信息)。

然而，下面两个问题是 EST 没有很好解决的。

1） 色觉机制非常精致而且巧妙，它是如何进化成这样的？

2） 怎样说明它在进化的每个阶段都比以前有更好的适应性？

为解释色觉机制进化，作者提出 3-8 译码模型，它能方便解释色觉机制进化，并说明进化的每一步都比以前能更多获得色光信息，另外，它还能方便解释从三原色和到心理色（包含色调）的转换。3-8 译码模型贯彻了生物学功利主义思想，和最大信息效率原理相互支持。

Appendix D 包含用以说明色觉进化的图片和相关网页链接。

9. 结论

Friston 的最小自由能原理继承了进化系统论关于自组织导致有序从而抵抗熵增大的积极思想，和里夫金的消极世界观(仅仅基于封闭系统中熵增大原理)不同，它用开放的非平衡系统的熵和信息变化解释生物系统的自组织和有序性。和前人方法不同的是，MFE 原理使用 VB 作为核心数学方法，VB 使用一种新的和信息相关的测度(变分自由能 $F=H(X|Y_\theta)$)作为目标函数，替代了常用的效用(或误差)和熵的线性组合。本文通过 VB 的具体算法得到结论：最小自由能准则其实包含最大似然准则(用于优化模型参数)和最大熵(即 $H(X|Y)$)准则。这一原理有助于解释生物怎样预测和适应(包括改变)环境，其中优化方法也能用于促进生态系统稳定发展。

然而，最小自的由能原理有两个缺点，这两个缺点都来自 VB。缺点之一是： 作者声称在两种优化过程中都最小化 F, 但是在实践中，最小化的是 $F-H(X|Y)$。其实践是正确的，理论是不完善。因为在某些情况下， $F-H(X|Y)$减小时，F 会增大。缺点之二是：最小化自由能的说法和物理学中的自由能概念矛盾，因为在物理学中，自由能越大越好，只是在封闭系统中，自由能会因为熵增大而被动减小。主动减少自由能和抵抗地球熵增大的目的矛盾。

作者提出语义变分贝叶斯和最大信息效率原理，它们可以作为 VB 和 MFE 原理的改进版本。语义变分贝叶斯最小化香农互信息和语义互信息的差 $R-G=F-H(X|Y)$而不是 F，这样，优化隐含变量时，理论就符合实践。而且，使用 SVB，求隐含变量更简单，更好理解。

根据局域非平衡和平衡系统中信息、熵和自由能之间关系的分析，香农信息相当于局域

非平衡系统的自由能，语义信息相当于局域平衡系统中的自由能(即 Exergy)增量，变分自由能 F 相当于局域非平衡系统中的物理熵。这样就好解释：通常我们接收和增大信息就相当于接收和增大自由能；提高信息效率就相当于提高用自由能做功的效率。这样，我们就能得到和 EST 兼容的结论：空气、水，生物，特别是人类，通过保存太阳送来的自由能而抵抗地球的熵增大。

致谢：作者感谢两位审稿人的 comments，本文 Sections 5, 6 and 8 因此做了较大改动，特别是添加了关于进化论和 Exergy 的讨论。作者也感谢雄楚渝博士，因为他五年前就提醒作者关注 Friston 的最小自由能原理。

附录 1. 本文缩写和解释列表

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

附录 B. Python 代码下载地址

除了图 3 和图 10，其他 8 个图的 Python3.6 原代码下载地址：http://www.survivor99.com/lcg/Lu-py2025-2.zip

附录 C. 说明美感是激励鸟类行为的正反馈信号

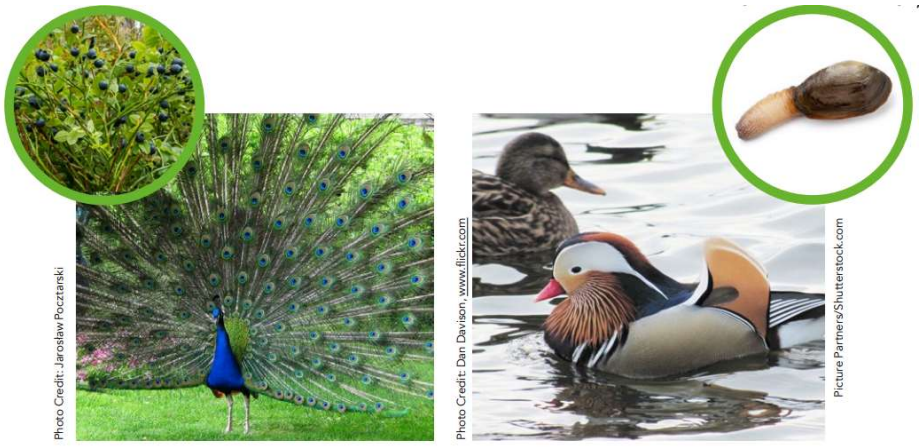


Figure 1: The Peacock's tail looks similar to a blueberry bush.

Figure 2: The Mandarin Duck's tail shape and colour resembles the soft-shell clam.

图解：孔雀和鸳鸯的羽毛模拟它们喜爱的食物。首先，需求关系选择了安娜蜂鸟的审美趣味；后来，雌性的审美趣味选择了雄鸟的羽毛。

上面图片本图片来自 Open Magazine: *Research Features. See* https://researchfeatures.com/needing-aesthetics-explain-birds-beauty-preferences/)。

更多讨论见作者主页：http://survivor99.com/lcg/english.

附录 D. 色觉进化图解

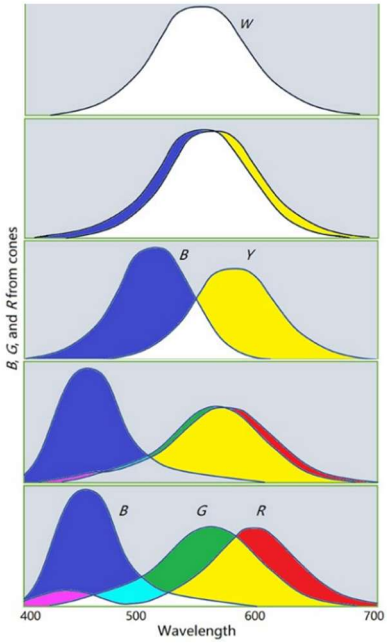


Fig. 3 The evolution of colour vision.

图解：随着人眼视锥细胞敏感曲线从一种分裂为两种和三种时，人眼色觉逐渐从两种进化为 8 种。在每一步进化后，色觉机制都比以前传递更多色光信息。 更多讨论见：https://researchfeatures.com/wp-content/uploads/2021/05/Chenguang-Lu.pdf，http://survivor99.com/lcg/english.

References

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

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

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

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

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

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

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

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

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

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

11 Ramstead, M.J.D.; Badcock, P.B.; Friston, K.J. Answering Schrödinger's question: A free-energy formulation. Phys Life Rev. **2018** *24*, 1-16. doi: 10.1016/j.plrev.2017.09.001.
12 Schrödinger, E. What Is Life? Cambridge: Cambridge University Press, Cambridge, UK, 1944
13 Huang, G.T. Is this a unified theory of the brain? 28 May 2008 From New Scientist Print Edition. Available online: https://www.fil.ion.ucl.ac.uk/~karl/Is%20this%20a%20unified%20theory%20of%20the%20brain.pdf. (accessed on 10 January 2025)
14 Portugali, J. Schrödinger's What is Life?—Complexity, cognition and the city. Entropy **2023,** *25*, 872. https://doi.org/10.3390/e25060872
15 Kim, C.S. Bayesian mechanics of synaptic learning under the free-energy principle. *Entropy* **2024**, *26*, 984. https://doi.org/10.3390/e26110984
16 Carl, M. Models of the translation process and the free energy principle. *Entropy* **2023**, *25*, 928. https://doi.org/10.3390/e25060928
17 Martyushev, L.M. Living systems do not minimize free energy: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Dèsormeau Ramstead et al., Physics of Life Reviews, Volume 24, 2018, Pages 40-41, https://doi.org/10.1016/j.plrev.2017.11.010.
18 Gottwald S.; Braun1, D.A. The two kinds of free energy and the Bayesian revolution，Available online: https://arxiv.org/abs/2004.11763. (accessed on 20 January 2025).
19 Silverstein, S.D.; Pimbley, J.M. Minimum-free-energy method of spectral estimation: autocorrelation-sequence approach, *J. Opt. Soc. Am*. **1990**, *3*, 356-372.
20 Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620.
21 Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev. II* **1957**, *108*, 171.
22 Lu, C. Shannon equations reform and applications. *BUSEFAL* **1990**, *44*, 45–52. Available online: https://www.listic.univ-smb.fr/production-scientifique/revue-busefal/version-electronique/ebusefal-44/ (accessed on 5 March 2019).
23 鲁晨光，广义信息论，中国科技大学出版社，1993.
24 Lu, C. A generalization of Shannon's information theory. *Int. J. Gen. Syst.* **1999**, *28*, 453–490.
25 Lu, C. Semantic Information G Theory and Logical Bayesian Inference for Machine Learning. *Information*, **2019**, *10*, 261.
26 Lu, C. The P–T probability framework for semantic communication, falsification, confirmation, and Bayesian reasoning. *Philosophies* **2020**, *5*, 25.
27 Kolchinsky, A.; Marvian, I.; Gokler, C.; Liu, Z.-W.; Shor, P.; Shtanko, O.; Thompson, K.; Wolpert, D.; Lloyd, S. Maximizing Free Energy Gain. *Entropy* **2025**, *27*, 91. https://doi.org/10.3390/e27010091.
28 Lu C. Semantic Variational Bayes Based on a Semantic Information Theory for Solving Latent Variables, Available online: https://doi.org/10.48550/arXiv.2408.13122. (accessed on 1 January 2025)
29 Lu, C. Using the Semantic Information G Measure to Explain and Extend Rate-Distortion Functions and Maximum Entropy Distributions. *Entropy* **2021**, *23*, 1050. https://doi.org/10.3390/e23081050.
30 Shannon, C.E. Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec. 1959, 4, 142–163.
31 Berger, T. *Rate Distortion Theory*; Prentice-Hall: Enklewood Cliffs, NJ, USA, 1971.
32 周炯槃，信息论基础，人民邮电出版社，1983.
33 Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1997**, *39*, 1–38.
34 Ueda N.; Nakano, R. Deterministic annealing EM algorithm, *Neural Networks*, **1998**, *11*, 271-282, 1998.
35 Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitrechnung*; Ergebnisse Der Mathematik (1933); translated as *Foundations of Probability*; Dover Publications: New York, NY, USA, 1950.
36 von Mises, R. *Probability, Statistics and Truth*, 2nd ed.; George Allen and Unwin Ltd.: London, UK, 1957.
37 Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353.
38 Davidson, D. Truth and meaning. *Synthese* **1967**, *17*, 3, 304-323.
39 Belghazi, M.I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. MINE: Mutual information neural estimation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1–44. https://doi.org/10.48550/arXiv.1801.04062.
40 Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with Contrastive Predictive Coding. Available online: https://arxiv.org/abs/1807.03748 (accessed on 10 January 2023).
41 *Ben-Naim, A. Is Entropy Associated with Time's Arrow?* Available online: https://arxiv.org/abs/1705.01467 (accessed on 20 March 2025).

42 *Ben-Naim, A. Can entropy be defined for and the Second Law applied to the entire universe?* Available online: https://arxiv.org/abs/1705.01100 (accessed on 20 March 2025).
43 *Ben-Naim, A. Can entropy be defined for, and the Second Law applied to living systems?* Available online: https://arxiv.org/abs/1705.02461 (accessed on 20 March 2025).
44 Wikipedia, Boltzmann statistics，Available online: https://en.wikipedia.org/wiki/Maxwell%E2%80%93Boltzmann_statistics (accessed on 20 March 2025).
45 Wikipedia, Maxwell-Boltzmann statistics，Available online: https://en.wikipedia.org/wiki/Maxwell%E2%80%93Boltzmann_statistics (accessed on 20 March 2025).
46 Bahrani, M. Exergy, Available online: https://www.sfu.ca/~mbahrami/ENSC%20461/Notes/Exergy.pdf (accessed on 23 March 2025).
47 Fisher RA. The Genetical Theory of Natural Selection Oxford Clarendon Press 1930.
48 Wright S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: Jones DF, editor. Proceedings of the Sixth International Congress of Genetics. Ithaca, New York: Brooklyn Botanical Garden; 1932.
49 Prigogine I, Stengers I. Order out of chaos: man's new dialogue with nature. New York, NY: Bantam Books 1984.
50 Haken H. Principles of brain functioning: a synergetic approach to brain activity, behaviour and cognition. Berlin: Springer-Verlag; 1996.
51 Eigen M, Schuster P. The hypercycle: a principle of natural self-organisation. Berlin: Springer-Verlag; 1979.
52 Kelso JS. Dynamic patterns: the self-organization of brain and behavior. Cambridge, MA: MIT Press; 1995.
53 鲁晨光，投资组合的熵理论和 价值，中国科技大学出版社, 1997.
54 Darwin， C. *The Origin of Species,* P.F. Collier & Son, New York, USA, 1909. Available online: https://www.bartleby.com/lit-hub/hc/the-origin-of-species/ (accessed on 253 March 2025)
55 Alcock, J. *Animal behaviour: An evolutionary approach*, Sunderland, Massachusetts, Sinauer Associates Inc. 1989.
56 Darwin, C . *The Descent of Man, and Selection in Relation to Sex*. John Murray, New York. 1871.
57 Cronin, H. *The Ant and the Peacock*. Cambridge University Press, London. 1991.
58 Fisher, R.A. The evolution of sexual selection, Eugenics Review, **1915**, *7*, 184-92.
59 Wikipedia, Young-Helmholtz_theory，Available online: https://en.wikipedia.org/wiki/Young%E2%80%93Helmholtz_theory (accessed on 20 March 2025).