Friston等人的最小自由能原理的失误和改进

——两种变分自由能与信息效率

关键词：变分贝叶斯，变分自由能，最小自由能原理，香农互信息，语义互信息， 信息效率

本PPT的PDF版下载地址： http://www.survivor99.com/lcg/fep-error.pdf

鲁晨光
2026.5
个人网站： http：//www.survivor99.com

背景1：变分贝叶斯方法(VB)

- 贝叶斯方法——贝叶斯主义推断方法，强调使用先验分布P(z)
- VB就是变分贝叶斯推断方法。
- https://en.wikipedia.org/wiki/Variational_Bayesian_methods
- Hinton等人: 提出VB用于无监督学习；优化准则：最小变分自由能(VFE)准则
- Hinton开始提出赫尔姆霍茨机， 后来又提出有限波尔茨曼机(RBM), 都用这一准则。因为自由能是物理学概念， Hinton获得物理学诺贝尔奖。
- 代表文章，著作：
- Hinton, G.E.; van Camp, D. **Keeping the neural networks simple by minimizing the description length of the weights**. 通过最小化权重编码长度保持神经网络简单性
- Hinton, G.E.; Zemel, R.S. **Autoencoders, minimum description length and Helmholtz free energy.** 变分编码器， 最小编码长度和Helmoholtz自由能
- Neal, R.; Hinton, G. **A view of the EM algorithm** that justifies incremental, sparse, and other variants. 关于EM算法的一个观点...
- 变分自由能(VFE)是什么？权重长度呈概率分布是P(x)，如果按预测的概率分布Pθ(x)编码，xi的码长是 -logPθ(xi),实际编码长度接近交叉熵，大于或等于香农熵：

$$H_\theta(X) = -\sum_i P(x_i)\log P_\theta(x) \geq H(X)$$

- VB任务：怎样让交叉熵接近香农熵？
- 反映香农无记记忆离散信源无失真编码定理。这是信息论方法， 本来和物理学没啥关系。

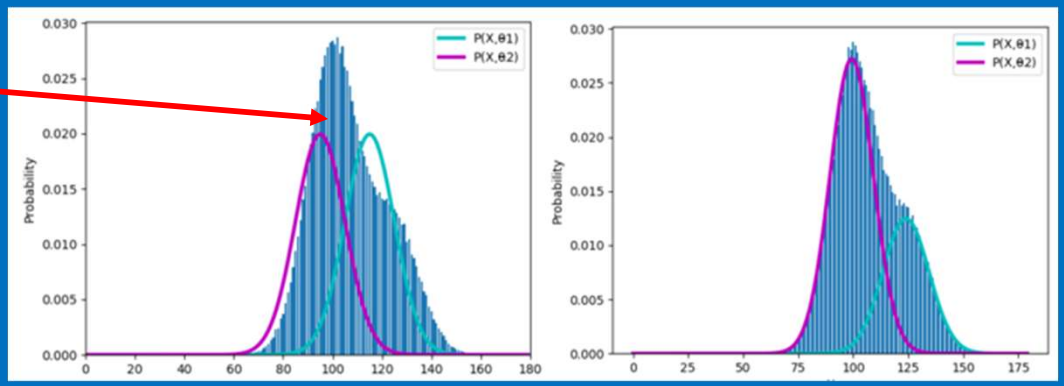
背景2：从VB到最小自由能原理（FEP）

- Friston研究脑神经科学，把VB应用到脑神经和行为科学
- 发展出自小自由能原理（FEP）——对标Jaynes的最大熵原理,
- 核心思想：人脑和环境相互适应，实现自己的目标；而不仅仅是被动服从和适应约束。
- 最近用于解释生命现象和自组织问题。
- Friston等人代表文章和专著：
- Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**
- Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, USA, 2022.
- 物理学界反对的比较多（主要因为质疑：1，自由能概念；2.作为一般原理）
- 我认为他们的方向是对的——因为最大熵理论要发展，但是数学上有问题
- 最近Friston把FEP用于主动推断Active Inference
- 现在强化主动推断， 弱化FEP
- VB、FEP，变分自动编码器VAE，和主动推断有广泛影响
- 核心问题：最小化VFE包含怎样的优化思想？
- 到目前为止还不清楚，VB发明者也没讲清，看网上介绍很难理解

理解变分自由能(VFE)——EM算法用于混合模型

- Neal和Hinton, Beal…很多人都用VB解决混合模型问题——无监督学习。
- 样本分布来自真模型：
- $P(x)=P^*(z_1)P(x|z_1^*)+P^*(z_2)P(x|z_2^*)$.
- 混合模型来自模型预测：
- $P_\theta(x)=P(z_1)P(x|z_j,\theta)+P(z_2)P(x|z_j,\theta)$.

开始猜不准　　收敛后



- 怎样让Hθ(x)→H(x)?
- 已知P(x)和模型类别（比如高斯）, 求解 P(z), 和P(x|θj)=P(x|zj, θ)
- EM算法是最小化联合交叉熵Hθ(X,Z), 采用变分方法
- 它先假设一个分布P(z), 和似然函数P(x|zj,θ)
- 用E步（算期望）得到 $q^{+1}(z_j \mid x_i) <= q(z_j)P(x_i \mid \theta_j)/Z_i,\ \ Z_i = \sum_j q(z_j)[P(x_i \mid \theta_j)],$
- 再用M1步: $q^{+1}(z_j) <= \sum_i P(x_i)q^{+1}(z_j \mid x_i),$
- M2步： P(x|zj,θ)<=P(x|z*j)， 对于高斯混合，求期望和标准偏差
- 重复就能让Hθ(x)→H(x)。
- 任务类似，VB解法有些不同。通常认为VB是比EM算法更一般的方法。

物理学自由能和Boltzmann分布

- 物理学中，赫尔姆霍茨自由能： F = E(ei,pi) - TS(pi)
- Hinton等人文章中：

$$F = \sum_i p_i E_i - H$$

$$H = -\sum_i p_i \log p_i$$

能量　　熵

- H—kTNH(x)
- Pi是变分，令∂F/∂S(pi)=0， 可以求出波尔茨曼分布：

$$p_i = \frac{e^{-E_i}}{\sum_j e^{-E_j}}$$

为什么叫变分自由能（VFE）

- 物理学中，赫尔姆霍茨自由能： F = E(ei,pi) - TS(pi)
- Hinton本意：

  P(x,z)=Pθ(x,z)

  =P(z)P(x|zθj)

$$VFE = E_{q(z|x)} \log \frac{q(z|x)}{P(z|x)} + H_\theta(x)$$

$$= E_{q(z|x)} \log \frac{q(z|x)}{P(x,z)} = E_{q(z|x)} \log \frac{1}{P(x,z)} - H(z|x)$$

能量　　熵

- 为什么不直接最小化Hθ(X)，因为其中没有q(z|x)
- 为什么常见的变分自由能公式是这样的 →

  Wiki Pedia Variational Bayesian Method中， VFE：

  为了更像物理学自由能公式，

  所以用q(z)代替q(z|x)，也带来问题。

$$F = \mathrm{E}_{q(z)} \log \frac{q(z)}{P(x,z)}$$

- 类似地， 令∂VFE/∂q(x)=0， 求出q(z).

VB使用的q(z)是后验分布q(z|x)，证据

- 1. 你可以问大模型，它们都说是；
- 2. 在Wikipedia, 词条Variational Bayesian Mehtods中有：

In variational inference, the posterior distribution over a set of unobserved variables $\mathbf{Z} = \{Z_1 \dots Z_n\}$ given is approximated by a so-called **variational distribution**, $Q(\mathbf{Z})$:

$$P(\mathbf{Z} \mid \mathbf{X}) \approx Q(\mathbf{Z}).$$

$$Q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i \mid \mathbf{X})$$

$$q_j^*(\mathbf{Z}_j \mid \mathbf{X}) = \frac{e^{\mathrm{E}_{q_{-j}^*}[\ln p(\mathbf{Z},\mathbf{X})]}}{\int e^{\mathrm{E}_{q_{-j}^*}[\ln p(\mathbf{Z},\mathbf{X})]}\, d\mathbf{Z}_j}$$

- 3. Hinton等人文章中：

$$p_i = \frac{e^{-E_i}}{\sum_j e^{-E_j}} \qquad (5)$$

This is exactly the posterior probability distribution obtained when fitting a mixture of Gaussians to an input vector.

the single Gaussian posterior, Q, for a noisy weight is

$$G(P, Q) = \int Q(w) \log \frac{Q(w)}{\sum_i \pi_i P_i(w)} dw \qquad (16)$$

- VB难懂， 这是主要原因！

Friston等人的一个错误结论

- 期望变分自由能（VFE）既表示平均惊奇，又表示预测不准确性。
- 但平均惊奇 $H_\theta(X)$ 与预测不准确性 $H_\theta(X|Z) = H(X|Z_\theta)$通常差异很大。

$$H_\theta(X) >> H_\theta(X|Z)$$

- 两种期望变分自由能来自两种变分q(z), q(z|x)

Hinton, VB

Friston, FEP

Hinton 等人用后验分布：
$q(z) = q(z|x)$→ F2

Friston 等人用先验分布：
$q(z) \neq q(z|x)$ → F1

F2 ≈ $H_\theta(X)$ 表示平均惊奇

=?
No!

F1≈ $H(X|Z_\theta)$ 表示预测不准确性或不确定性

$$F_2 = D_{KL}(q(x|z)||P(z)) + H_\theta(X)\ (z) \approx H_\theta(X) \quad >> \quad F_1 = D_{KL}(q(z)||P(z)) + H(X|Z_\theta) \approx H(X|Z_\theta)$$

物理学自由能 vs 变分自由能

- 我发表于Entropy《Improving the Minimum Free Energy Principle to the Maximum Information Efficiency Principle》中讨论过变分自由能概念的问题。

物理学

物理学自由能 **F = E - TS**

对应香农互信息

R = I(X;Z) = H(X)- H(X|Z)

VB和FEP

F2 ≈ Hθ(x) :编码长度

F1 ≈ Hθ(x|z): 残差编码长度

对应物理学约束平衡系统的熵

- 通信中类似的是信息，两者都是好东西，通常增大好；但是用起来要节省。

F1， F2， 熵， 都是不好东西或负担，通常减小好。

- 物理学中熵增大， 自由能减小， 这是被动减小；而VB主动减小VFE。

两种自由能之间关系

- F1≈条件交叉熵 Hθ(X|Z)　　F2≈交叉熵 Hθ(X)

F1	+ R	= F2
预测不确定性 负的效用	香农互信息 成本	总目标：追求效率

- 最小变分自由能准则包含追求最大信息效率思想！
- 两者关系：

  F2=F1+R
- 推导公式

$$F_2 = \mathrm{E}_{q(x)}\mathrm{E}_{q(z|x)}\log\frac{q(z|x)}{P(x,z)}$$

$$= \mathrm{E}_{q(x,z)}\log\frac{q(z|x)}{P(z)} - \mathrm{E}_{q(x,z)}\log P_\theta(x|z)$$

$$= D_{\mathrm{KL}}(q(z|x)\|q(z)) + D_{KL}(q(z)\|P(z)) - \mathrm{E}_{q(x,z)}\log P_\theta(x|z)$$

$$= R + F_1$$

香农互信息R　　F1

- VB问题：怎么计算信息效率？ 如何权衡最大语义信息和最大信息效率？

语义信息G理论

- 1993， 1997：

- 1999: A generalization of Shannon's information theory" , Int. J. of General Systems, 28: (6) 453-490 1999
- 2025： A Semantic Generalization of Shannon’s Information Theory and Applications. *Entropy* **2025**, *27*, 461. https://doi.org/10.3390/e27050461
- 基本公式1: 真值函数(反映语义）和似然函数之间的关系：
- 语义贝叶斯公式

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

语义互信息公式的性质及G和F1的关系

- 基本公式2：语义互信息公式：

- G和F1之间的关系：

- G和平均失真之间的关系：

$$G=I(X;Z_\theta)=\sum_j\sum_i P(x_i\mid z_j)P(z_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}\approx H(X)-F_1$$

$$=\sum_j\sum_i P(x_i)P(z_j\mid x_i)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)}=H_\theta(Z)-\bar{d}$$

- 公式的性质：
- 1）当语义信道匹配香农信道，即P(x|θj)=P(x|z)或T(θj |x)正比于q(z|x)时， 语义信息量最大达， 达到其上界R。
- 2）当香农信道匹配语义信道时，即P(x|z)=P(x|θj)或q(z|x)正比于T(θj |x)i时， 信息差R-G最小， 或信息效率G/R最大.

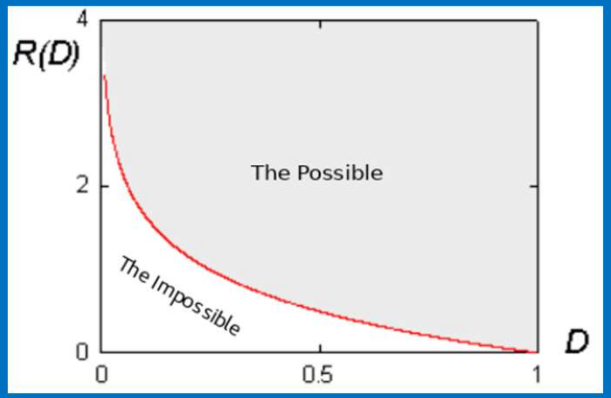
语义变分贝叶斯——来自R(D)函数推广

(8)

- 信息率失真函数R(D)推广到信息率逼真函数R(G)
- 目标函数： R-sG， s=1, 就是追求最大信息效率1
- 用变分方法得到优化香农信道迭代公式：

$$q^{+1}(z_j \mid x_i) = q(z_j)[P(x_i \mid \theta_j)]^s / Z_i,\ Z_i = \sum_j q(z_j)[P(x_i \mid \theta_j)]^s,$$

$$q^{+1}(z_j) = \sum_i P(x_i) q(z_j \mid x_i),$$

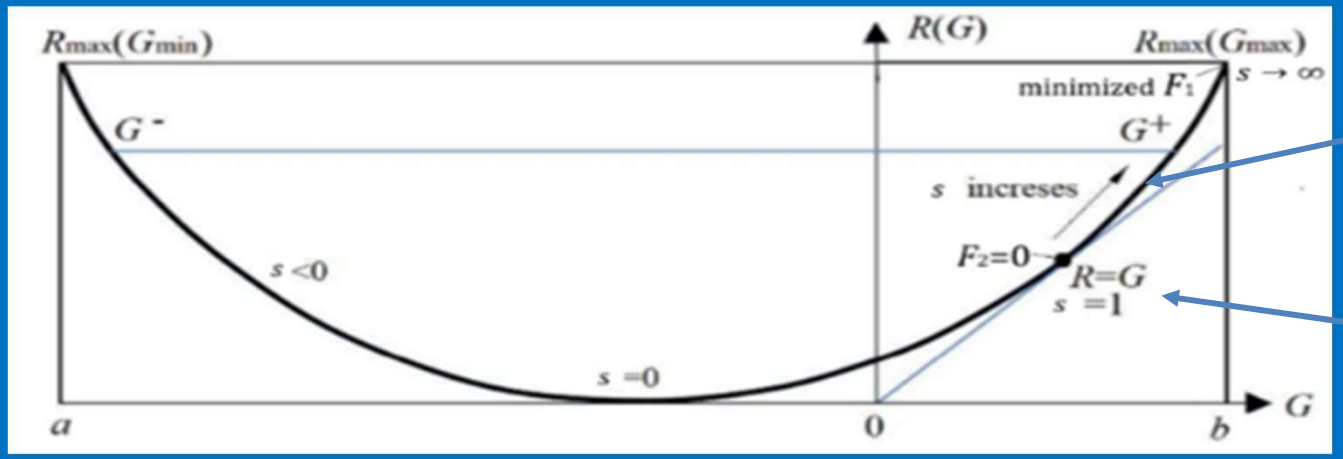


- 增大s，可在最大语义信息和最大信息效率之间权衡

S=1,最小化F2
最大信息效率

最小化R-sG
权衡

S—> ∞, 最小化F1
最大语义信息



s从1增大，牺牲效率G/R，增加效用G

最小化F2, 使得R=G， s=1, G/R=1

F2 和香农互信息，语义互信息及信息效率之间的关系

- F1：预测不确定性
- F2：预测 不确定性+ 信息成本
- IND: 信息差 R- G
- 关系推导：

$$F_2 = E_{q(x,z)}[\log \frac{q(z|x)}{P(z)} - \log P_\theta(x|z)]$$

$$= \mathrm{E}_{q(x,z)}[\log \frac{q(z|x)}{q(z)} + \log \frac{q(z)}{P(z)}] + H_\theta(x|z)$$

$$= R - G + D_{\mathrm{KL}}(q(z) \| P(z)) + H(x) \approx IND + H(x).$$

信息差　　0　　香农熵=常数

- 所以最小化F2→最小化信息差→最大化信息效率
- 这就是最小化VFE包含的优化思想

最小化F1和F2 的不同目标和优缺点

- F1：预测不确定性, 最小化F1等价于最大化语义互信息和对数似然度
- 机器学习中用于优化模型参数 θ;
- 分类，主动推断中用于最小化不确定性或误差

- F2：用于发现隐含变量，混合模型（聚类），无监督学习。

VB：F2
混合模型/求解隐含变量

FEP：最小化F1
预测/分类/主动推断

- 缺点：
- Friston等人错在没用F2, 没有区分两种任务，在数学上漏掉了VB提高通信效率的努力；
- Hinton等人用q(z)代替q(z|x), 容易引起混乱，也不便于表达香农互信息。

- 改进：用IND= R – sG = R - s [H(x) – F2] 代替F2和F1。
- S=1 时， min IND 等价于 Min F2
- S=∞时， min IND 等价于 min F1.

主动推断误用F1的后果——变成误差控制

- 预测模型推断：优化模型参数，使主观预测符合客观事实；
- 主动推断：固定预测模型，改变客观事实，使客观事实符合主观期望——约束控制。

$$G=I(X;Z_\theta)=\sum_j\sum_i P(x_i\mid z_j)P(z_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)}$$

- 如果最小化F2, 两者相等VFE最小；而最小化F1，P(x|z)集中分布F1最小。

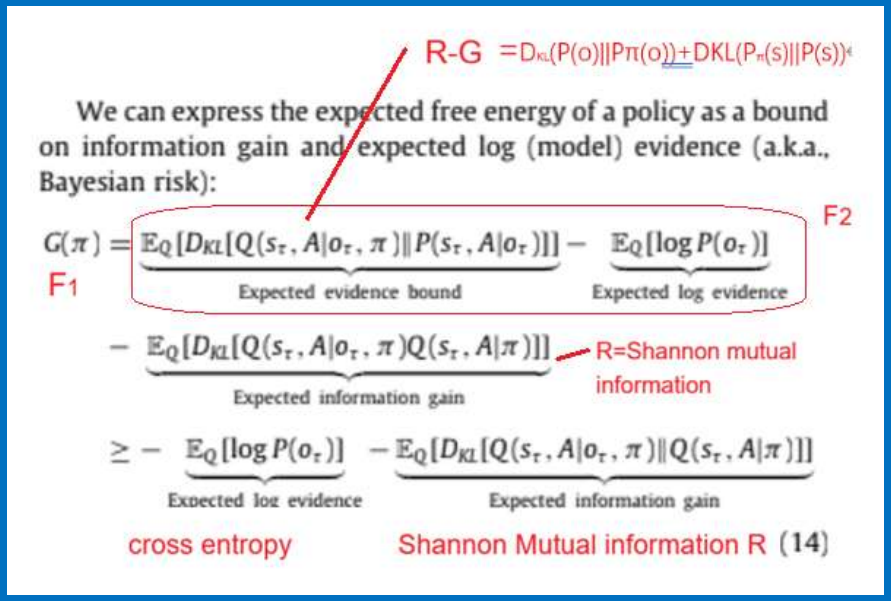
- 例如**主动推断**任务之一：想象公交车一路停靠很多站点。
- 给定初始状态x0和一系列目标约束——用似然函数表示。因为有随机干扰，控制可能不准。优化准则是最小化F1 = H(x|z,θ)= - E logP(x|z,θ).
- 和混合模型不同，**不要求** q(x)=P(x)。
- 假设似然函数是高斯分布，

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- -Logf(x)=相对误差。
- **最小化F1**就等价于最小化平均相对误差， 主动推断就变成误差控制。
  控制信息成本的努力就漏掉了。好像新手频繁打方向盘（用F1）, 老手转得少（用F2）
- 如果用F2, 我们可能使用较小的控制量，近似满足一系列约束。

Friston等人真地误用先验分布作为变分？证据？

1. 在他们很多文章中，q(z)和q(z|x)同时出现在一个公式中。而在VB中不会有(同时出现好表示香农互信息）



2. 我和Friston先生本人有过交流，
   他说他用的期望自由能是F1。
3. 他也说从F1公式能推导出F2公式，
   但是他推荐的文章中，实际推导的是

$$F1 = H_\theta(x) - I(X;Y) = F2 - R,$$

   这等价于我推导的F2 = R + F1.

4. 他赞成语义互信息G=H(x)-F1,　　所以F1接近$H_\theta(x|z)$。

语义变分贝叶斯（SVB）与β-VAE比较

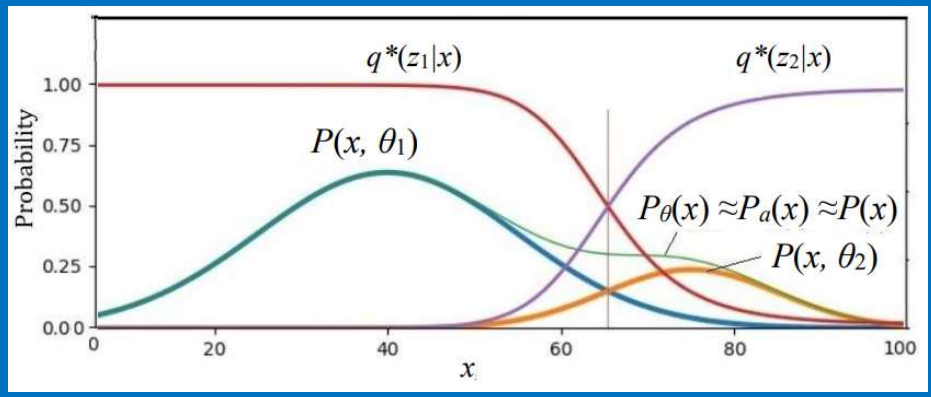
- 新潮流， 改进VB到β-VAE， 涉及信息效率。
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proceedings of the International Conference on Learning Representations
- 目标函数（最大化）：
  - $L = \mathrm{E}_{q\varphi(z|x)}\left[\log p_\theta(x|z)\right] - \beta D_{\mathrm{KL}}(q_\varphi(z|x) \,\|\, p(z)),$
- 等价于（最小化）：
  - F2 = F1 + βR’
- 比较SVB中目标函数：
  - IND=R - sG

- 但其信息项不是真正香农互信息。
- 我的SVB可直接计算R、G与G/R， 允许约束函数是真值，隶属，相似，或失真函数。

实验1：三种任务优化的香农信道q(z|x)的表现

- 1. 混合模型，最小化F2时：q(z|x)≈P(z|x)

最小化 **F2**

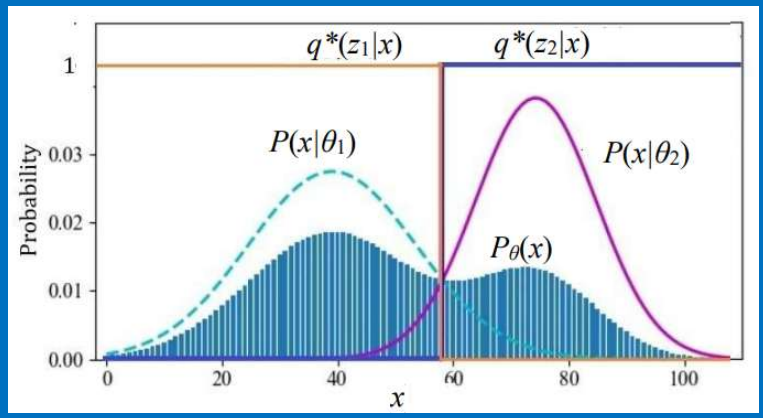


这时, F_2=0， G/R=1 ， s=1 ，

$q(x) \approx P_\theta(x) \approx P(x)$

- 2. **最大互信息分类**

最小化 **F1**

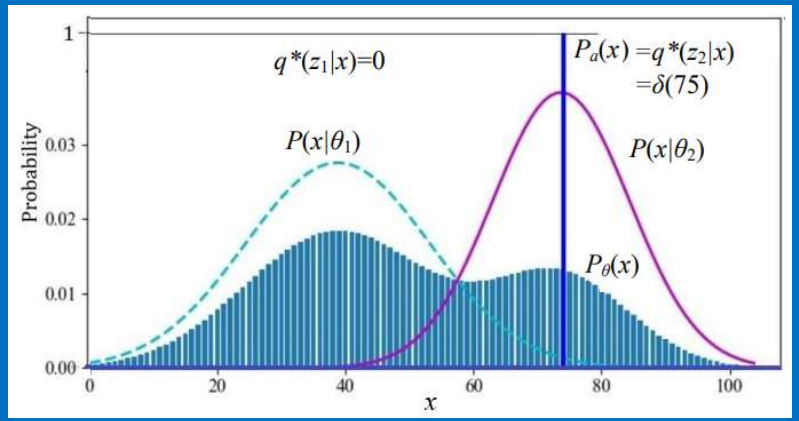


G/R 下降

- 3. 主动推断或约束控制时，最大F1或最大化语义信息

$P_a^*(x) = g^*(z_2|x) = \delta(75)$, 不需要 $q(x) \approx P(x)$.

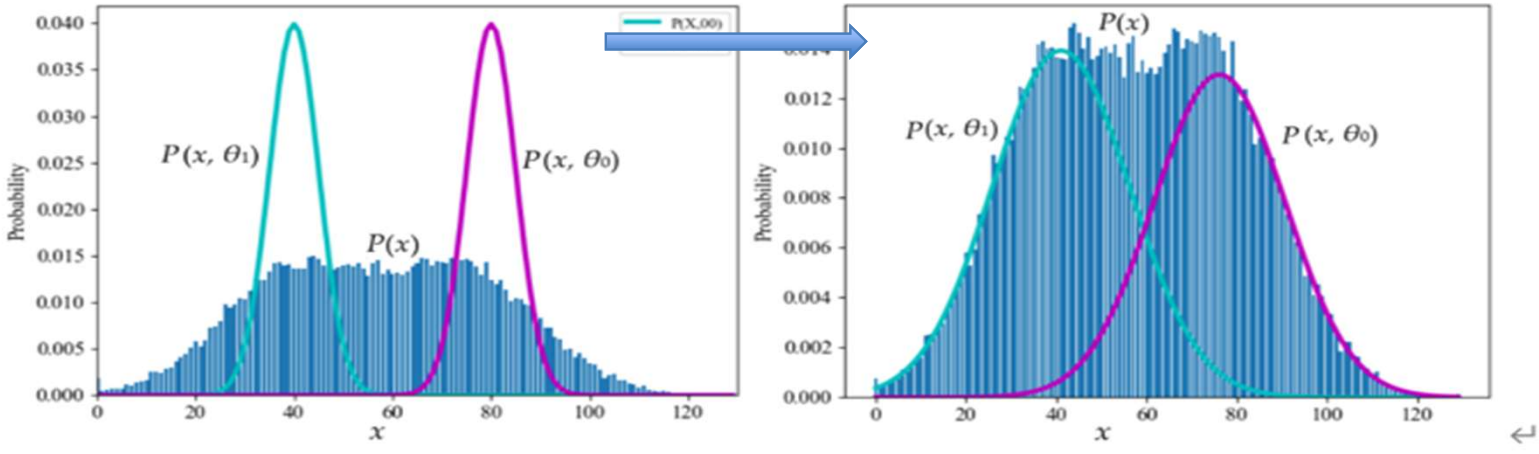
最小化 **F1**



实验2：F2下降而F1上升的例子

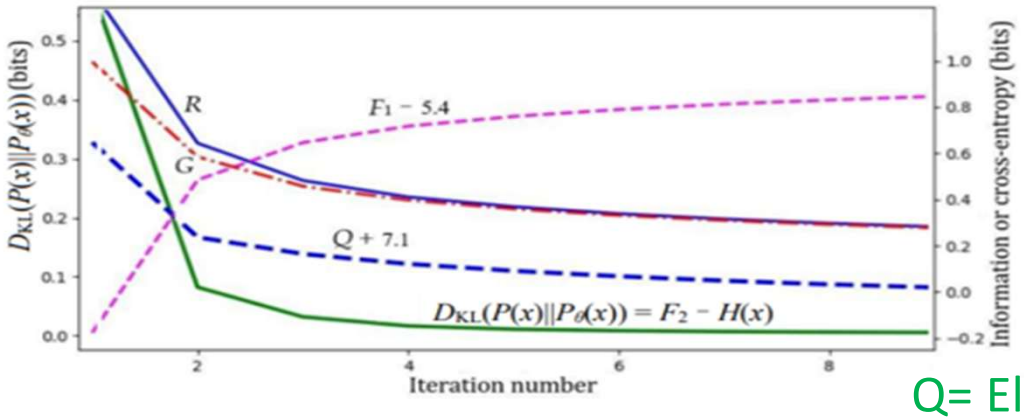
- 存在反例，F2下降时F1可能增加，说明两种自由能并不等价。

F2减小，F1增大



(a) The iteration starts

(b) The iteration converges



Q= ElogPθ(x,z)

(c) F_2, $R - G$, F_1, and Q change in the iterative process.

Fig. 5. During the convergence of the mixture model, $D_{KL}(P(s) \parallel P_\mu(s))$, $R - G$, and F_2 decreased, while F_1 increased, and Q decreased.

主要参考文献

1. Hinton, G.E.; van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proc. 6th Annu. Conf. Comput. Learn. Theory*; 1993; pp. 5–13.
2. Hinton, G.E.; Zemel, R.S. Autoencoders, minimum description length and Helmholtz free energy. In *Proc. 6th Int. Conf. Neural Inf. Process. Syst.*; 1993; pp. 3–10.
3. Neal, R.; Hinton, G. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*; Jordan, M.I., Ed.; MIT Press: Cambridge, MA, USA, 1999; pp. 355–368.
4. Wikipedia. Variational Bayesian methods. Available online: https://en.wikipedia.org/wiki/Variational_Bayesian_methods
5. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. doi:10.1038/nrn2787.
6. Friston, K.J.; Parr, T.; de Vries, B. The graphical brain: Belief propagation and active inference. *Netw. Neurosci.* **2017**, *1*, 381–414. doi:10.1162/netn_a_00018.
7. Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, USA, 2022.
8. Lu, C. Improving the minimum free energy principle to the maximum information efficiency principle. *Entropy* **2025**, *27*, 684. doi:10.3390/e27070684.
9. Lu, C. A semantic generalization of Shannon's information theory and applications. *Entropy* **2025**, *27*, 461. doi:10.3390/e27050461.
10. Higgins, I.; et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*; 2017.

总结1：核心逻辑链

- 两种变分q(z)=q(z|x)，q(z)<>q(z|x)

↓

- 两种期望自由能 F2，F1

↓

- 两种优化目标：最小化F2即最小化R-G， 最小化F1 ≈ H(x|z, θ)即最大化语义信息G

↓

混合模型，无监督学习需要F2；主动推断实际用F1

↓

- 改进VB→SVB，F2→R-sG， 重视权衡语义信息和信息效率

总结2：结论

- VB和FEP中变分不同，变分自由能也不同。
- 用q(z)作为q(z|x)的简写，得到期望的自由能F2, 用q(z)<>q(z|x)作为变分，得到F1。
- 最小化F2包含最小化F1， 反之不然。
- 混合模型需要最小化F2，减小不确定性或优化预测模型需要最小化F1。
- Friston等人结论， 变分自由能表示既预测不准确性（H(x|z θ)， 也表示平均惊奇Hθ(x)，这是错的， 错在使用q(z)<>q(z|x)作为变分得到F1≈H(x|z ,θ)，客观上漏掉了VB追求信息效率的努力。
- VB用q(z)作为后验分布q(z|x)的简写不便于表达香农互信息， 也不便于计算信息效率。
- SVB用信息差R- sG作为目标函数代替F2，可以继承VB重视信息效率的思想， 并能在最大语义信息和最大信息效率之间权衡。

感谢熊楚瑜博士，为他五年前提醒我关注Friston的最小自由能原理。

欢迎批评交流，更多讨论见我主页：http://www.survivor99.com