

说说Karl Friston的自由能原理



Ruyuan Zhang
心理学话题下的优秀答主

Feitong Yang、滤波插值 等 121 人赞同了该文章

好久不在知乎码字。最近在这个问题下，尝试回答了一下。

如何评价神经科学家Karl Friston的自由能原理和主动推断的时代意义？
385 赞同 · 72 评论 回答

这是个很难懂的理论，看不懂很正常。因为很多人都看不懂，哪怕是专家。比如，曾经有在哥伦比亚大学的各领域专家聚集在一起也看不懂他的理论。lesswrong.com/posts/wpZ...

在知乎上我也没看到有什么比较细致的解释，大部分解释就是把Karl说的英文翻译成中文。

看懂Friston的自由能理论，需要一些预先知识。我个人建议先完全不要去看他的东西，但是需要搞清楚以下几个概念，

变分推断，贝叶斯推断里面的model evidence，生成模型等。

其中最关键的就是变分推断，我发现身边的心理学家或者神经科学家其实很少有人能看懂变分推断的。我大概是2015年看到这个理论，当时读friston的paper也是完全抓瞎，直到我最近理解了变分推断，才算大概明白了自由能原理。

另外，个人不建议看Karl Friston的原文。Karl Friston是个天才，但是他写作有问题，不能把自己的观点用简单的语言表达出来。我个人建议看Sam Gershman的这篇解读，我觉得写得非常好。gershmanlab.webfactional.com...

下面我尝试拆分一下这个理论

我们就按照维基上对free energy principle的介绍，从贝叶斯变分推断讲起，慢慢推过去。

https://en.wikipedia.org/wiki/Free_energy_principle
en.wikipedia.org/wiki/Free_energy_principle

对于一般的贝叶斯推断来说，我们观察到 s 是sensory state, ϕ 是环境中生成 s 的hidden state。这里我选用的是和上面wikipedia中相同的符号，以方便读者理解。我们大脑认识这个世界，基本任务就是根据sensory state来对hidden state做推断，也就是求解后验概率 $p(\phi|s)$ 。这是认知科学中所有贝叶斯模型的最基本思想。

$$p(\phi|s) = \frac{p(s, \phi)}{p(s)} \tag{1}$$

$$p(s) = \frac{p(s, \phi)}{p(\phi|s)} \tag{2}$$

$$\log(p(s)) = \log(p(\phi, s)) - \log(p(\phi|s)) \tag{3}$$

公式(1)是贝叶斯推断的基本格式，然后很容易推到公式(3)。

那么，下面就是变分推断的内容，要求解 $p(\phi|s)$ ，我们引入一个分布 $q(\phi)$

$$\log(p(s)) = \log\left(\frac{p(\phi, s)}{q(\phi)}\right) - \log\left(\frac{p(\phi|s)}{q(\phi)}\right) \tag{4}$$

注意无论 $q(\phi)$ 是什么分布公式(4)都成立。然后两边求对于 $q(\phi)$ 的期望，就有

$$\int \log(p(s))q(\phi)d\phi = \int \log\left(\frac{p(\phi,s)}{q(\phi)}\right)q(\phi)d\phi - \int \log\left(\frac{p(\phi|s)}{q(\phi)}\right)q(\phi)d\phi \quad (5)$$

$$\log(p(s)) = \int \log\left(\frac{p(\phi,s)}{q(\phi)}\right)q(\phi)d\phi - \int \log\left(\frac{p(\phi|s)}{q(\phi)}\right)q(\phi)d\phi \quad (6)$$

由于公式(5)左边的被积分项 $\log(p(s))$ 并不包含 ϕ 所以做完积分不变，就得到公式(6)。然后，我们让 $F(s)$ 等于公式(6)右边第一项的负数。公式(6)右边第二项的负数，就是变分分布 $q(\phi)$ 和所要求解的后验概率 $p(\phi|s)$ 的KL divergence。

$$F(s) = - \int \log\left(\frac{p(\phi,s)}{q(\phi)}\right)q(\phi)d\phi = -E_q\left(\log\left(\frac{p(\phi,s)}{q(\phi)}\right)\right) \quad (7)$$

$$D_{KL}[q(\phi)||p(\phi|s)] = - \int \log\left(\frac{p(\phi|s)}{q(\phi)}\right)q(\phi)d\phi \quad (8)$$

那么公式(6)可以重新表示成

$$\log(p(s)) = -F(s) + D_{KL}[q(\phi)||p(\phi|s)] \quad (9)$$

$$F(s) = -\log(p(s)) + D_{KL}[q(\phi)||p(\phi|s)] \quad (10)$$

好了，记得整个推断的目的是求后验概率 $p(\phi|s)$ ，因为这个太难求了，我们用变分分布 $q(\phi)$ 来近似它，当 $q(\phi)$ 和后验概率 $p(\phi|s)$ 越接近，两者的 $D_{KL}[q(\phi)||p(\phi|s)]$ 就越小(注意 $D_{KL}[q(\phi)||p(\phi|s)] \geq 0$)。同时， $-\log(p(s))$ 虽然我们不知道是多少，但肯定是个常数，因为给定一个 sensory state，外部世界产生它的概率是恒定的，只不过我们求不出来。所以当 $D_{KL}[q(\phi)||p(\phi|s)]$ 越小，公式(10)右边越小，公式左边 $F(s)$ 也就越小。反之也成立，如果我们优化使得 $F(s)$ 越小，那么 $D_{KL}[q(\phi)||p(\phi|s)]$ 也就越小。

这里 $F(s)$ 就是 free energy，minimize free energy 就等价于 minimize $D_{KL}[q(\phi)||p(\phi|s)]$

以上内容，和大脑其实无关，我只不过重复了一遍机器学习里面变分推断的原理，是为了说明，minimize free energy 其实就是贝叶斯推断中变分推断这一种特殊形式，目的就是求后验概率，只不过换了个名字而已。

所以说，Karl Friston 说的 minimize free energy，不是什么玄学。就是认知神经科学当中经常说的贝叶斯推断，也就是说我们人脑在做贝叶斯推断时候，不过是用变分推断来求解后验概率的过程，这就是 minimize free energy。

我们继续来看一下 en.wikipedia.org/wiki/Free_energy 上面给出的公式

$$F(s,u) = -\log(p(s|m)) + D_{KL}[q(\phi|u)||p(\phi|s,m)] \quad (12)$$

这个公式是不是很像我上面写的公式(10)? 但是细心的你也发现了一点区别，在上面公式(10)里面是没有 Internal model u 和 Generative (world) model m 这两项的，这又是什么意思呢?

先解释 Generative (world) model m 。这代表是外部物理世界运行的规律，比如太阳从东边升起，从西边落下，总有一个客观规律存在。比较让人困惑的是，这个 m 和 hidden state ϕ 到底是什么关系? 简单来说， m 是外部世界运行的模型， ϕ 是这个模型的参数。 m 比 ϕ 高一级，联合 sensory state 在一起，数学关系就是：

$$p(s|m) = \int p(s|\phi)p(\phi|m)d\phi \quad (13)$$

换句话说，sensory state 并不是无边无际的，取决于外部物理世界的运行规律。所以严格来说 $-\log(p(s))$ 应该换成 $-\log(p(s|m))$ 。记住，虽然我们永远无法求得 $p(s|m)$ ，但是它肯定还是个定值。

再来说 Internal model u 。这里就是 Karl Friston 理论上最大的贡献，他认为我们大脑是用变分分布 $q(\phi)$ 来不断猜测和逼近实际的后验分布。但是这个 $q(\phi)$ 也不是随便来的，它取决于我们大脑内部对外部世界的认识，比如，实际情况是太阳从东边升起，从西边落下。但是大脑是完全无从得知这一客观真理的，有可能大脑就觉得太阳应该是从东南边升起，从西北边落下。所以 $q(\phi)$ 应该替换成 $q(\phi|u)$ 。

那么，如果

$$\begin{aligned} F(s,u) &= -\int \log\left(\frac{p(\phi|s,m)}{q(\phi|u)}\right) q(\phi|u) d\phi = - \\ &E_q\left(\log\left(\frac{p(\phi|s,m)}{q(\phi|u)}\right)\right) \tag{14} \\ D_{KL}[q(\phi|u) \parallel p(\phi|s,m)] &= - \\ &\int \log\left(\frac{p(\phi|s,m)}{q(\phi|u)}\right) q(\phi|u) d\phi \tag{15} \\ F(s,u) &= -\log(s|m) + D_{KL}[q(\phi|u) \parallel p(\phi|s,m)] \tag{16} \end{aligned}$$

这就是Karl Friston minimize free energy理论的完整表达式，其中公式(16)和Wikipedia上面给出的公式(12)一致。

在Wikipedia上面还讲了minimize free energy理论和predictive coding, optimal control, active inference等多个概念的关系。我当然没有读完所有列出的论文，也绝不敢说理解了他说的每一句话，但是有了上面这个基础，去理解其他概念可能稍微容易点。

编辑于 2020-07-18 10:07

认知神经科学 贝叶斯统计 计算神经科学



欢迎参与讨论

20 条评论

默认 最新



元子 关注我的人

“Karl Friston理论上最大的贡献：大脑是用变分分布 $q(\phi)$ 来不断猜测和逼近实际的后验分布”，这种思路背后隐含了一个小人在大脑里，这是用第三者视角研究意识，这是单向线性思维（还原思维）的必然，不论绕多少个圈圈后面都需要有一个类似“上帝”的源头，除非能说明这个小人的工作原理（那就是意识的工作原理）。实际上，最小自由能原理是大脑的行为动力，而不是大脑的行动方向，而且大脑根本不猜测世界是怎样的，他只是与世界交互，就像河流从来不知道大海是什么样，他能奔向大海只是往低处流（水不是懂得了遵守重力原理然后努力往低处流，而是水被重力直接拉向低处）。所以，不要再追随最小自由能原理了，改变思路寻找基于自然作用的形式化表达（大脑就是这样的）。

2023-10-11

回复 2



轩朗 关注我的人

请问这个自由能理论如何证伪？

2020-07-19

回复 1



Ruyuan Zhang 作者 置顶

这是个非常好的问题。我记得看过一个报道karl自己都说不证伪。我自己的思考是，要么你证明人就不是在做贝叶斯推断，要么证明人做贝叶斯推断不是用的变分的方法，而是用别的算法，比如MCMC，但是后者逻辑上来说并不严格

2020-07-19

回复 6



Hover Xiao

是什么的机缘，居然让我看到这篇文章~~我在搜索mRNA优化算法，居然跳到了心理学和神经科学。热力学理论要统一生物系统研究了啊。

2023-11-01

回复 喜欢



OxAA 关注我的人

读过原文 感觉是民科

2020-07-18

回复 喜欢



OxAA 关注我的人 Nathanwhat

他以前是会在一些小破期刊发关于free energy hypothesis of the brain的文章，直到在2010年在ature review neuroscience上吹unified brain theory。他提的理论涉及很广，但是几乎没法证伪，这是为什么我说是民科（伪科学）。

2020-07-20

回复 1



duskhdm OxAA

好家伙，读科普竟然遇到了raca家人。salute

2022-04-12

回复 喜欢

展开其他 2 条评论



Eliot Alexander

关注我的人

研究下深度学习里的变分自编码器有助于理解变分的意义。

2020-07-18

回复

喜欢



xhb

关注我的人

好文！里面讲到的变分推断，尤其是关于vba的原理那一块和他们介绍的非常接近！可惜数学水平有限，还不能完全理解.....

2020-07-18

回复

喜欢



孤刃

关注我的人

gershman真的是奇才，不仅研究上做得好，是大牛，在写东西上也深入浅出，很多文章可读性极强、读起来极为舒服。。。

2020-07-21

回复

喜欢



腹肌撕裂的猫

你好，Gershman那篇自由能论文的连接打不开了，能提供个论文标题吗

2021-12-09

回复

喜欢



Ruyuan Zhang

作者

What does the free energy principle tell us about the brain?

我试一下链接能打开

gershmanlab.com/pubs/fr...

2021-12-09

回复

1 喜欢



腹肌撕裂的猫

Ruyuan Zhang

可以了，谢谢

2021-12-09

回复

喜欢



张磊

我关注的人

早几年的时候，人们提到free energy基本还是框架在参数估计领域，现在烂大街了😂

2020-07-19

回复

喜欢



anningyu

我关注的人


所以这跟Hinton的wake sleep algorithm算是一脉相承

2020-07-19

回复

喜欢

点击查看全部评论 >

 欢迎参与讨论

文章被以下专栏收录



从科研到脑科学

谢谢科研，心理学和认知神经科学

推荐阅读



Marcus 电荷转移理论

Triborg

Klein-Gordon
Equation

相对论性量子力学（1） — Klein-Gordon方程

Porcupine



统计力学-2 麦克斯韦-玻尔兹曼分布

zhbjc

发表于hjc的物...

朗道能级量子化和整数效应（2）

2021 年北京的国庆又下雨聊。这篇简单聊一下其他相关整数量子霍尔效应的内...关Hofstadter's butterfly。IQHE 是来自于朗道能级量自由空间加磁场取朗道...
Gaple... 发表于物