arXiv:2304.02658v1 [cs.NE] 5 Apr 2023

预测编码作为反向传播的神经形态替代方案：批判性评估

Umais Zahid
华为技术有限公司研发部门，英国伦敦
umais.zahid@huawei.com

Qinghai Guo
华为技术有限公司研发部门，中国深圳
guoqinghai@huawei.com

Zafeirios Fountas
华为技术有限公司，英国伦敦
zafeirios.fountas@huawei.com

摘要

反向传播已迅速成为现代深度学习方法中主要的信用分配算法。近年来，研究表明，预测编码（PC）的一些改进形式——这种算法起源于计算神经科学——能够产生与反向传播下近似相等或完全相等的参数更新。鉴于这一联系，有人提出，PC可以作为反向传播的替代方案，并具备一些理想特性，可能有助于其在神经形态系统中的实现。在本文中，我们利用文献中提出的不同当代预测编码变体来探讨这些主张。我们获得了这些预测编码变体的时间复杂度界，并证明其下界由反向传播给出。我们还介绍了这些变体的关键特性，说明了它们对神经生物学合理性及其解释的影响，尤其是从标准预测编码作为潜在概率模型的变分贝叶斯算法这一视角来看。我们的研究为这两种学习框架之间的联系带来了新的认识，并表明，与此前设想的相比，预测编码以其当前形式作为反向传播直接替代方案的潜力可能更为有限。

关键词 反向传播 · 预测编码 · 神经形态计算

1 引言

预测编码（预测编码）是过去二十年间兴起的一种重要神经科学理论，已成为一种极具说服力的大脑感知与行动计算模型（Friston，2005年；Rao&Ballard，1999）。从理论角度看，预测编码在其标准表述中，因被解释为一种用于潜在层级模型学习与推理的变分贝叶斯算法而受益（Bogacz，2017年；Friston，2003，2005年；Friston&Kiebel，2009），这使其成为贝叶斯大脑假说和自由能原理等规范性理论的合理实例。更具实证意义的是，其候选神经生物学实现与大脑皮层的经典回路展现出了令人印象深刻的高度对应关系（Bastosetal.，2012；Shipp，2016），同时还成功解释或再现了许多神经生理学和认知现象，例如端部停止（Rao & Ballard，1999）、双目竞争（Hohwy，Roepstorff，& Friston，2008）、注意（Feldman &Friston，2010；Kanai，Ko mura，Shipp，& Friston，2015），以及对事件持续时间知觉的偏差（Fountas et al.，2022）。
近期研究提出，修订版预测编码与反向传播之间存在联系，并将前者视为后者的替代方案，同时具备人们自然会期待一种在制定时考虑神经生物学约束的理论所拥有的各种优势（Millidge、Salvatori、Song、Bogacz 和 Lukasiewicz，2022年；Millidge、Tschantz、Seth 和 Buckley，2020a；Song、Lukasiewicz、Xu 和 Bogacz，2020年）。这些优势包括计算的局部性和并行化能力，从而可能带来更佳性能，并使其更易于在神经形态硬件上实现。

这项工作的目的是调查这些建议，并更精确地阐明反向传播与预测编码之间的关系，特别是关于进行此类比较所需的假设类型，此外，还要准确定义基于PC的表述在并行化能力、局部性以及最终实际性能方面可能具有的优势。在下文中，我们重新回顾了一些

(Rosenbaum, 2022) 所涵盖的内容，同时将其扩展以讨论预测编码在计算与内存复杂度以及计算与记忆局部性方面的一般影响。我们还给出了当前预测编码变体相对于反向传播的时间复杂度下界证明，并实证验证了这些证明。我们的工作发现，虽然所考察的算法提供了有趣的、受生物学启发的反向传播替代实现，但在相同问题设置下它们被证明保证更慢，并且不清楚它们在局部性方面改进的程度，特别是与现代反向传播候选方案相比。最后，我们注意到，当采用进行此类比较所需的修改时，预测编码模型偏离了生成变分贝叶斯框架，该框架经常被用作该理论的动机基础。

2 反向传播

反向传播最初由（Linnainmaa，1976）在累积舍入误差的背景下提出，后来由（Rumelhart、Hinton和Williams，1986）在神经网络的背景下推广，如今已迅速成为现代深度学习任务中的主力信用分配算法。最简单地说，反向传播可以被视为微积分中链式法则的一种算法高效实现，它能够针对标量或低维输出，计算深度嵌套函数的高维梯度。它通过用嵌套函数内部深层数值的子项（其函数依赖项）的梯度来定义这些数值的梯度，从而实现这一点。随后，对于最接近输出的数值，按顺序向后计算这种递归关系。这种递归关系使反向传播能够通过一次前向和反向计算过程，高效地计算函数链中相对于所有参数的梯度。

为了说明其运作方式，我们给出以下简单但具有高度通用性的问题设定。考虑一个任意函数 F，它由多个组成函数 f_i 复合而成，每个函数都可以由一组参数 $\theta_{\mathbf{i}}$ 参数化。于是我们可以写出：

$$\begin{aligned} f_0(x_0, \theta_0) &= \mu_1 \\ f_1(\mu_1, \theta_1) &= \mu_2 \\ &... \\ f_{(L-1)}(\mu_{L-1}, \theta_{L-1}) &= \mu_L \\ f_L(\mu_L, x_L) &= E \end{aligned} \tag{1}$$

其中，我们使用 μ_i 来表示复合函数的中间输出，并使用 x_0, x_L 来表示分别作为我们函数输入和输出提供的已知值。 x_0 和 x_L 例如可以分别是输入和分类标签， μ_ℓ 可以是多层感知器的中间激活向量，而 f_L 可以是分类损失。

如果我们希望计算任意中间参数相对于函数输出 E 的梯度，那么可以先定义以下简单的递归关系。

$$e_\ell = \left\{ \begin{array}{ll} \frac{\partial f_\ell}{\partial \mu_\ell}^T e_{\ell+1} & \text{for } \ell = 1, \ldots, L-1 \\ \frac{\partial f_L}{\partial \mu_L} & \text{for } \ell = L \end{array} \right\} \tag{2}$$

对于多元初等函数， $\frac{\partial f_\ell}{\partial \mu_L}$ 项对应于初等函数 f_i 相对于$\mathbf{x_i}$的雅可比矩阵。并且根据链式法则，误差项 e_ℓ 因此将等于 E 相对于 $x_{\ell+1}$的梯度。

在确定了与 x_ℓ相关的梯度相对应的误差向量 e_ℓ 后，便可以通过将下一层的误差向量与函数相对于参数的雅可比矩阵进行矩阵乘法，来计算输出 E 相对于任意中间参数的梯度。

$$\frac{\partial E}{\partial \theta_\ell} = -\frac{\partial \mu_{\ell+1}}{\theta_\ell}^T e_{\ell+1} \tag{3}$$

尽管在这一情形下这点显而易见，但为了便于我们后续进行比较，有必要在此指出，上文提到的所有雅可比矩阵都是在我们的函数的前馈输出值处进行计算的；正如我们将要讨论的那样，对于贝叶斯预测编码而言，情况未必如此。我们还要指出，在实践中，公式（3）和（2）中规定的雅可比矩阵几乎从未被显式实例化。相反，实现反向传播的自动微分程序会将每个基本函数或组成函数与一个相应的向量-雅可比积（VJP）函数关联起来，而这种方式通常在计算和内存方面的效率要高得多（Bradbury 等，2018年；Paszke 等，2017年）。

2.1 计算与时间复杂度

计算前向传播（公式 1）以及后续反向递归（公式 2 和 3）的成本，可以证明其计算和时间成本上界为通过相应初等函数执行一次前向传播成本的一个较小常数倍。这一自动微分文献中的众所周知的结果有时被称为廉价梯度原理，也是反向传播能够在现代大数据场景中作为信用分配算法取得如此成功的原因之一。在（Baur & Strassen，1983）的开创性工作中，该常数被证明对于有理函数为 3；而对于由初等算术函数和三角函数构成的任意函数，该常数通常为 5（Griewank，1997，2009；Griewank & Walther，2008）。该常数的精确值通常还取决于所考虑的具体硬件上各种基本数学运算（例如加法、乘法、内存访问和非线性运算）的相对成本。对于现代深度神经网络运算及其相关硬件，该常数通常取 3（Hoffmann 等人，2022；Kaplan 等人，2020），我们在比较中也采用这一数值。

正如我们将看到的，目前预测编码和反向传播算法的实现所采用的基本计算单元相同：向量-雅可比积。根据廉价梯度原理，计算该向量-雅可比积的成本可由一次前向求值的成本限定。因此，为了比较这些算法，我们可以报告相对于计算一次前向传播的成本而言的计算复杂度；对于初等组成函数 f_i和整体复合函数 F，分别记作 $\mathcal{C}_i$ 和 $\mathcal{C}_F$。
遵循 (Griewank& Walther, 2008) 中的记法，我们将与函数F相关联的计算任务记为 $task(F)$，相应的计算与时间复杂度分别记为WORK$\{task(F)\}$和TIME$\{task(F)\}$。为简单起见，我们假设工作量和时间采用标量复杂度度量，并使用某个正值常数 w将单个基本函数 f_i 的时间复杂度与工作复杂度联系起来，使得 TIME$\{task(f_i)\} = w$WORK$\{task(f_i)\}$。请注意，由于并行计算，对于 F上的其他高级计算任务（例如预测编码推断），这种关系可能并不成立；我们将在后续章节中对此进行讨论并加以处理。因此，利用廉价梯度原理，我们可以写出反向传播时间复杂度的如下分解式和界。

$$\begin{aligned}
\text{TIME}\{backprop(F)\} &= \text{TIME}\{forward(F)\} + \text{TIME}\{backward(F)\} && (4)\\
&= \sum_i \text{TIME}\{forward(f_i)\} + \sum_i \text{TIME}\{vjp(f_i)\} && (5)\\
&\leq w\sum_i \mathcal{C}_i + 2\mathcal{C}_i = 3w\sum_i \mathcal{C}_i && (6)
\end{aligned}$$

3 预测编码

3.1 变分预测编码

我们首先将预测编码的标准表述描述为一种变分贝叶斯算法（Bogacz，2017年；Buckley、Kim、McGregor和 Seth，2017年；Friston，2003年、2005年）。在这一背景下，预测编码算法可以被解释为一种针对潜层次生成模型执行推理（针对潜在状态）和学习（针对参数）的方法。与近期大量将其与反向传播进行比较的研究保持一致，我们将重点限定于针对静态（即非动态/基于时间序列的）观测和状态的预测编码。我们将把这一表述与近期的修改版本区分开来，例如 FPA-PC（Millidge、Tschantz 和 Buckley，2020年）和Z-IL（Salvatori、Song、Lukasiewicz、Bogacz和 Xu，2021年；Song 等，2020年）；我们将其称为变分预测编码（VPC），以强调修改后的表述未必对应于变分贝叶斯推断和学习算法。

为了便于后续比较预测编码与反向传播，本节将沿用 f_i 中首次定义的基本函数，用于定义层级模型中的条件关系。

特别是，我们将把我们的组成函数 f_i 的输出视为高斯潜在随机变量的均值，其中每个高斯潜在随机变量都以其父节点为条件。然后，当 f_L 是一个与特定输出分布 P_L的对数似然相对应的损失函数时（这是一种常见的问题设定），这将为我们的概率图模型带来如下对数联合概率：

$$\log p(x_0, ..., x_{L-1}, x_L|\theta_0, ...\theta_L) = \log p(x_L|x_{L-1}, \theta_{L-1}) + ... + \log p(x_1|x_0, \theta_0) \quad (7)$$

with

$$x_\ell|x_{\ell-1} \sim \begin{cases} \mathcal{N}(f_{\ell-1}(x_{\ell-1},\theta_{\ell-1}),\Sigma_\ell)) & \text{for } \ell = 1,...,L-1 \\ P_L(f(x_{L-1,t},\theta_L)) & \text{for } \ell = L \end{cases} \tag{8}$$

基于这一模型，PC 回答了这样一个问题：如何学习参数 θ，使其模型证据 $\log\ p(\mathbf{x}^{(1)}{}_{\text{OBS}}\ \cdots\mathbf{x}^{(D)}{}_{\text{OBS}}|\boldsymbol{\theta})$ 达到最大值。为简化符号，我们将所有观测（x_0 和 x_L）、潜在状态（$x_1,...,x_{L-1}$）和参数（$\theta_0,...,\theta_L$）分别合并为向量 $\mathbf{x}_{\text{OBS}}$、$\mathbf{x}_{\text{LAT}}$ 和 $\boldsymbol{\theta}$，上标索引表示来自数据生成（观测）分布 $\mathcal{X}$ 的样本。

通过优化模型参数以最大化模型证据，通常会遇到一个难题：对潜在状态进行边缘化的计算难以处理（$\mathbf{x}_{\text{LAT}}$）。对此，变分贝叶斯的解决方案是优化证据下界（ELBO），它通常也称为（负）自由能。该下界依赖于一种可处理的潜在状态近似后验密度形式

$$\text{ELBO} = E_{q(\mathbf{x}_{\text{LAT}})}\left[\log p(\mathbf{x}_{\text{OBS}},\mathbf{x}_{\text{LAT}}|\boldsymbol{\theta})\right] - E_{q(\mathbf{x}_{\text{LAT}})}\left[\log q(\mathbf{x}_{\text{LAT}})\right] \tag{9}$$

对于近似后验密度 $q(\mathbf{x}_{\text{LAT}})$，预测编码采用点质量（狄拉克 δ）分布，可以是隐式采用（Bogacz，2017年；Buckley 等，2017年；Millidge、Tschantz 和 Buckley，2020），也可以是显式采用（Friston，2003年、2005年）。将该狄拉克 δ 分布的中心记为 $\boldsymbol{\phi}$，上述下界可简化为目标函数

$$\text{ELBO}_{\mathbf{PC}}(\boldsymbol{\phi},\boldsymbol{\theta}) = \log p(\mathbf{x}_{\text{OBS}},\boldsymbol{\phi}|\boldsymbol{\theta}) \tag{10}$$

预测编码通过先对狄拉克 δ 参数 ϕ 执行上升步骤来优化此函数，这些参数对应于潜在状态近似后验分布的众数。获得相对于 ϕ 的联合对数概率最大值后，我们再通过计算在这些 ϕ取值处评估的联合对数概率的梯度，来更新参数 θ。

视角不同，这一初始步骤可以有多种解释：由于对对数联合概率进行最大化，它可以被视为对潜在状态求取最大后验估计；也可以被视为期望最大化方案中的期望步骤（Friston，2005年）；或者从变分贝叶斯的角度看，被视为收紧变分下界的步骤。随后，最大化步骤（对 θ进行优化）可以通过小批量或随机梯度下降（SGD）来实现，从而能够针对大型观测数据集进行可处理的优化。最终算法可简要概括如下：

1. 定义一个关于潜在状态（$\mathbf{x}_{\text{LAT}}$）和观测（$\mathbf{x}_{\text{OBS}}$）的（可能具有层次结构的）图模型，其参数为 $\boldsymbol{\theta}$（即 $\log\ p(\mathbf{x}_{\text{OBS}},\mathbf{x}_{\text{LAT}}|\boldsymbol{\theta})$） 2. 对于小批量$\mathbf{x}_{\text{OBS}} \sim \mathcal{D}$，其中 $\mathcal{D}$ 是数据生成分布**推理：** 通过对– $\log\ p(\mathbf{x}_{\text{OBS}},\mathbf{x}_{\text{LAT}}|\boldsymbol{\theta})$执行梯度下降，为潜在状态求取最大后验估计（$\mathbf{x}_{\text{MAP}}$）**学习：** 使用随机梯度下降，根据推理结束时求得的最大后验估计处的负对数联合概率更新参数 $\boldsymbol{\theta}$： $-\log\ p(\mathbf{x}_{\text{OBS}},\mathbf{x}_{\text{MAP}}|\boldsymbol{\theta})$

为了后续讨论，查看上述算法中概述的梯度上升（推理）过程中发生的精确函数形式和计算将是有用的。首先，由于方程（7）中描述的对数联合中潜在状态的条件对数似然是高斯的，因此它们采取一系列平方精度加权预测误差的形式，每个对应于一个组成函数 f_i，加上输出对数似然

$$\begin{aligned} \text{ELBO}_{\text{PC}} &= \log p(x_0,x_1,...,x_{L-1},x_L) && (11)\\ &= f_L(x_L,x_{L-1}) && (12)\\ &\quad + ... \\ &\quad + (x_\ell - f_{\ell-1}(x_{\ell-1}))^T\Sigma_\ell^{-1}(x_\ell - f_{\ell-1}(x_{\ell-1})) \\ &\quad + ... \\ &\quad + (x_1 - f_0(x_0))^T\Sigma_1^{-1}(x_1 - f_0(x_0)) && (13) \end{aligned}$$

其中，为简洁起见，我们暂时省略了对参数 θ_ℓ 的依赖。

该目标函数相对于特定潜在状态 x_ℓ 的梯度，就是与从其父节点对 x_ℓ 的预测相对应的精度加权预测误差，以及与之相对应的精度加权预测误差

对其子节点的预测 x_ℓ 所做的。梯度下降（推理）过程因此可以描述为以下离散化梯度流，假设 f_L 采用高斯输出对数似然（即欧几里得输出损失）：

$$x_{\ell,t+1} = x_{\ell,t} - \gamma \left.\frac{\partial F}{\partial x_\ell}\right|_{x_{\ell,t}} \tag{14}$$

$$= x_{\ell,t} - \gamma \left[\Sigma_\ell^{-1}(x_{\ell,t} - f_{\ell-1}(x_{\ell-1,t})) - \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,t}}^{T} \Sigma_{\ell+1}^{-1}(x_{\ell+1,t} - f(x_{\ell,t}))\right] \tag{15}$$

其中 γ 是某个推理步骤的步长。假设方差为单位矩阵，该方程可简化为

$$= x_{\ell,t} - \gamma \left[(x_{\ell,t} - f_{\ell-1}(x_{\ell-1,t})) - \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,t}}^{T} (x_{\ell+1,t} - f(x_{\ell,t}))\right] \tag{16}$$

更一般地用误差来表述，我们可以将其写成包含输出损失和对数似然的任意定义

$$= x_{\ell,t} - \gamma \left[e_{\ell,t} - \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,t}}^{T} e_{\ell+1,t}\right] \tag{17}$$

with

$$e_{\ell,t} = \begin{Bmatrix} (x_{\ell,t} - f_{\ell-1}(x_{\ell-1,t})) & \text{for } \ell = 1, ..., L-1 \\ \frac{\partial f_L}{\partial f_{L-1}}^T & \text{for } \ell = L \end{Bmatrix} \tag{18}$$

其中我们滥用符号并使用 f_{L-1} 来表示该函数及其在当前值 ofx_{L-1}处评估的输出，即 $f_{L-1}(x_{L-1,t})$。我们这样做是为了将其与 μ_L 以及方程 (2) 中的相关雅可比矩阵区分开来，后者表示使用前馈值 μ_{L-1} 计算的预损失输出预测，即 $f_{L-1}(\mu_{L-1})$。也就是说，在每个推理步骤中，潜在状态被更新，使其最小化与其潜在状态当前预测相对应的误差，以及与其子潜在状态当前预测相对应的误差。

随后，使用在这些收敛的 MAP值处计算的对数联合分布导数来更新模型参数（通过随机梯度下降）。该梯度是与其所参数化的条件分布相关的误差和该函数相对于 θ 的雅可比矩阵的乘积。令 t_c 为变分模态收敛的时间。则对数联合分布相对于模型参数的梯度为

$$\frac{\partial F}{\partial \theta_\ell} = \frac{\partial f_\ell}{\partial \theta_\ell}^T e_{\ell+1,t_c} \tag{19}$$

方程（15-16）描述了标准预测编码公式的动力学。在这些动力学下，对潜在状态的推理可以被视为在相应概率模型上的 MAP 推理，而学习则可以被视为关于其模型参数最大化证据下界。具体而言，该下界对应于在我们的潜在状态上假设的点质量变分分布。

反转配置：这种特定的预测编码表述与认知科学中采用的标准预测编码表述（Buckley 等，2017年；Friston，2003年、2008年；Rao & Ballard，1999年）以及大多数生成式建模方案恰好相反，因为在这种表述中，观测对潜变量（如分类标签）进行参数化，因此层级高于潜变量。以这种方式反转预测编码，以便与反向传播进行比较，会对神经生物学合理性产生不容忽视的影响。历史上，标准预测编码表述的优势之一在于其提出，大脑皮层中层级较高的潜变量具有非线性的混合与调制效应（$f(x_\ell, \theta_\ell)$），这与我们对大脑皮层解剖结构的认识相符；也就是说，自上而下的反馈连接通常具有更多分支并发挥调制作用，而自下而上的影响，或前馈连接，则更多发挥驱动作用（$-e_\ell$）（Bastos 等，2012年；Friston，2005年、2008年；Markov 等，2014年；Shipp，2016年）。

值得注意的是，标准预测编码的有监督生成式（即非反转）表述迄今难以在复杂分类任务中取得有竞争力或接近有竞争力的结果。例如，据我们的经验以及此处观察到的情况，传统的生成式预测编码模型在经过训练以分类和生成 MNIST 手写数字后，测试集准确率始终无法超过 80%（Kinghorn、Millidge 和 Buckley，2022）。作者尚未发现任何传统（生成式）预测编码在标准机器学习任务中取得有竞争力的分类表现的实例。尽管可能的原因超出了本文的讨论范围，但这很可能是由于现有尝试中通常使用的基于高斯的监督信号，或者是狄拉克 δ（点质量）近似后验在训练生成模型时的局限性所致，正如（Zahid、Guo、Friston 和 Fountas，2023）所指出的。

3.2 记忆局部性、权重传输与权重对称性

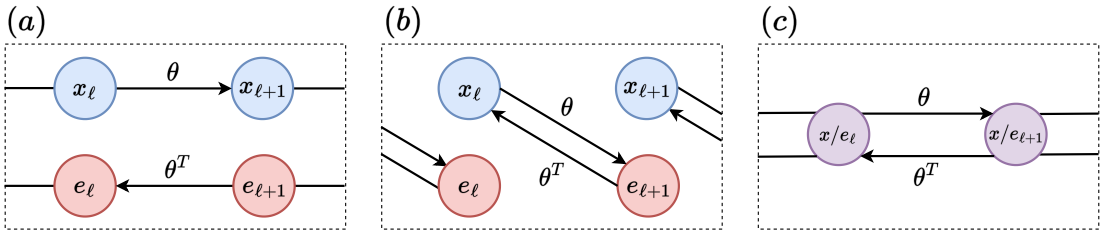

图1：（a）反向传播标准提案实现中的权重传输问题。（b）标准预测编码公式中的权重对称性问题。为清晰起见，我们省略了预测编码所需、但与权重传输问题无关的额外连接。（c）一种旨在解决反向传播权重传输问题的替代提案示例（Amit，2019）

在神经生物学背景下，局部计算通常指局部可塑性，即由其所连接神经元的活动（也就是说，突触前和突触后的神经元活动）所决定的突触权重更新，以及突触局部性。评估学习算法在大脑中的实现是否合理时，常会运用这种局部性的概念。从历史上看，反向传播在生物学合理性方面受到的一项关键批评，源于它存在这种非局部性；这一问题被称为权重传输问题（Crick，1989；Grossberg，1987；Zipser 和 Rumelhart，1993）。

在上述方程（1）所述的问题设定下，权重传输问题可概括为：参与计算前馈激活 $\mu_{\ell+1}$ 的特定权重或参数矩阵（例如 θ_ℓ），需要被传输以用于计算误差 e_ℓ。具体而言，考虑由权重矩阵 θ_ℓ 所表示的仿射变换以及非线性函数 $g(\cdot)$ 组成的函数 f_ℓ，使得

$$\mu_{\ell+1} = f_\ell(\mu_\ell, \theta_\ell) = g(\theta_\ell \mu_\ell) \tag{20}$$

e_ℓ 的计算过程如下

$$e_\ell = \frac{\partial \mu_{\ell+1}}{\partial \mu_\ell}^T e_{\ell+1} = g'(\theta_\ell \mu_\ell)\theta_\ell^T e_{\ell+1} \tag{21}$$

其中，由于存在转置权重项 θ_ℓ^T，我们需要重复使用前向权重。

针对反向传播，已有许多广为人知的尝试，旨在解决权重传输问题在生物学上难以成立这一问题。其中包括颇具影响力的研究（Lillicrap、Cownden、Tweed与Akerman，2016），该研究表明，通过固定随机矩阵 B 替代 θ_ℓ 来计算误差，仍能促进学习，且无需重复使用前向参数。然而，随机反馈权重仅在浅层网络中保持良好性能，在更深层模型中则会失效。后续及近期研究表明，单独训练这些反馈权重能够提升更深层模型的性能（Akrout、Wilson、Humphreys、Lillicrap与Tweed，2019；Amit，2019）。

尽管这些不同的模型以各种方式解决了权重传输问题，但它们往往需要或引入关于实现这些模型所需神经生物学机制的额外假设。例如，Zipser与Rumelhart（1993）的研究要求以前馈矩阵和反馈矩阵相同的方式进行初始化；Kolen与Pollack（1994）对此提出批评，并通过权重衰减解决了这一问题。然而，Kolen与Pollack（1994）和Zipser与Rumelhart（1993）的研究都要求在独立网络之间传递突触权重变化，这一点本身也被认为不符合生物学实际（Akrout等，2019）。Amit（2019）要求计算过程严格遵循固定顺序，并让前馈神经元随后也充当误差神经元，图1以示意图展示了这一点。

最后，为了进一步比较，(Lillicrap等人,2016)和(Akrout等人,2019)都要求前向和反向通路中的激活之间进行交互，这通常通过三因子式学习规则来实现，而这会带来超出简单双因子赫布学习规则的复杂性。不过，我们注意到，从神经生物学合理性的角度来看，鉴于越来越多的文献提供了三因子学习规则的实验证据或机制性提案（Pawlak,Wickens, Kirkwood, &Kerr, 2010; Sjöström &Häusser,2006; Urbanczik & Senn, 2014），这或许并不构成特别严重的问题。

一个与权重传输非常相似但略有不同的问题存在于预测编码中，这是由于方程15和16中雅可比矩阵项($\frac{\partial f_\ell}{\partial x_\ell}$)的存在。对于 f_ℓ(仿射 + 非线性函数)的相同定义，这导致以下

标准预测编码的推断方程

$$x_{\ell,t+1} = x_{\ell,t} - \gamma \left[e_{\ell,t} - \frac{\partial f_\ell}{\partial x_\ell}\Big|_{x_{\ell,t}}^{T} e_{\ell+1,t} \right] \tag{22}$$

$$= x_{\ell,t} - \gamma \left[e_{\ell,t} - g'(\theta_\ell \mu_\ell)\theta_\ell^T e_{\ell+1,t} \right] \tag{23}$$

其中 θ_ℓ 也用于计算 $e_{\ell+1}$

$$e_{\ell+1,t} = x_{\ell+1,t} - g(\theta_\ell x_\ell) \tag{24}$$

因此，介导后续误差单元（$e_{\ell+1}$）对先前潜在状态（x_ℓ）影响的权重的（负）同一值，也同样是从潜在状态到误差单元的互惠连接所必需的。图1对此进行了描绘。这一特性在最早期的一些关于预测编码的论述中就已被指出（Friston，2005年，2008年），不过当时是从预测编码具有神经生物学合理性的优势这一角度出发的，因为据报道，大脑中普遍存在互惠连接（Felleman &Van Essen，1991）。然而，互惠连接的普遍存在并不足够，因为至少从朴素的角度来看，我们还要求其强度完全等价。

我们将这一问题与权重传输问题区分开来，称之为权重对称性问题，因为它要求同一对神经元之间的正向突触和互惠突触具有等效权重；从神经生物学合理性的角度来看，相较于反向传播所要求的权重传输，这无疑是一个更容易处理的问题。其原因在于，可塑性通常被认为是突触前活动和突触后活动的某种简单函数。鉴于正向突触和互惠突触都能够获取这两种活动，因此正向权重和反向权重趋于一致并非不合理。特别是，给定一个简单的双因素赫布学习规则，至少对于简单的赫布可塑性而言，可以预期正向连接和互惠连接都经历相同的权重修改（即突触前活动与突触后活动的乘积）。进一步引入权重衰减后，权重最终会同步，正如(Kolen与Pollack，1994年)针对权重传输情况所展示的那样。在不采用衰减的情况下，这种方法在标准预测编码网络上确实已得到验证，且性能仅有最低程度的下降（Millidge、Tschantz、Seth与Buckley，2020b）。

接下来，我们将讨论近期提出的若干种预测编码改进形式，它们被视为反向传播的神经形态替代方案。

3.3 无穷方差极限下作为预测编码的反向传播

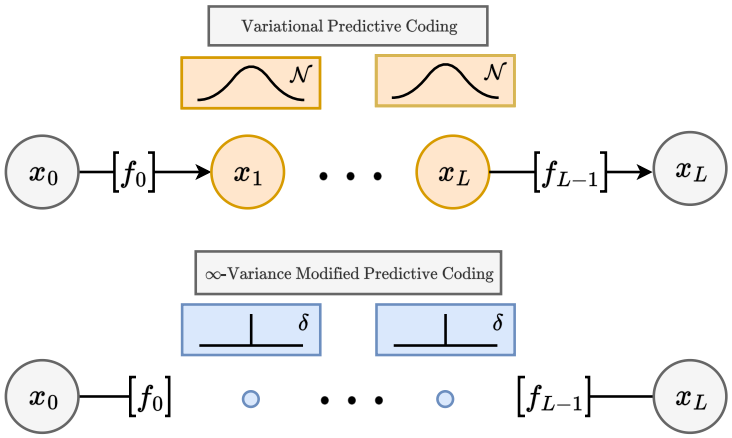

图2：标准预测编码（即变分预测编码，VPC）及近期改进形式所假设的底层概率模型。请注意，VPC包含概率潜在状态，而FPA-PC仅假设存在确定性状态和观测。灰色圆圈表示观测变量，蓝色圆圈表示确定性中间状态，橙色圆圈表示概率潜在状态。

预测编码与反向传播之间的联系最早在（Whittington & Bogacz, 2017）中提出，其中表明，在以下特定的限制性条件下，上述预测编码算法所得到的权重更新将等同于相应反向传播计算图中的权重更新：

要求 1

• 当变分模态或潜在状态 $(x_1, ..., x_{L-1})$ 被初始化为相应基于反向传播的计算链的前馈值 $(\mu_1, ..., \mu_{L-1})$ 时，并且：

要求 2

• **案例1**当前馈预测导致输出损失为零（即负对数似然损失为零）

• **情况2**当输出方差(Σ_L)设置得足够高于剩余方差 $(\Sigma_{\ell<L})$，使得 $\left(\frac{\Sigma_L}{\Sigma_\ell}, \forall \ell < L\right)$ 的比值趋于无穷大并且学习率得到适当缩放。

这两种情况都对应于同一项要求，即在推理过程结束时，潜在状态 $x_1, ..., x_{L-1}$的值仍然等于或接近其前馈值。换言之，这可以看作一项要求：在同时观测输入观测（例如图像）和输出观测（例如分类标签）时，潜在状态的 MAP值，等于仅以输入观测为条件时的 MAP值。理解这项要求的另一种更直观的框架是注意到，在这些情况下所得到的推理和学习过程，等价于针对一个概率图模型进行推理和学习；在该模型中，中间潜在状态具有 0 条件不确定性或方差，因此是输入观测的确定性输出。

在这一框架下，学习退化为对概率图模型实施最大似然学习，其中输入 x_0 通过深度“总”网络函数 $F = f_L \circ f_{L-1} ... \circ f_1 \circ f_0$对输出似然进行参数化，并以 x_L作为观测。这等价于在输出损失函数 f_L （见公式 1）对应于有效对数似然时，使用反向传播实施最大似然训练。关于该方法所假设的概率图模型与标准 VPC的对比示意，请参见图2。
要看出这一点，请注意，对于**情况 1**，这一点显然成立：由于误差为零，上述确定性潜在最大似然目标和采用概率潜在状态的预测编码目标所产生的权重更新相同，均为零。

更有趣的是，如果考虑到将 Σ_L 放大并将所有学习率按其倒数重新缩放，等价于将中间方差 $\Sigma_\ell, \forall \ell < L$缩小，同时保持输出方差不变，并保持学习率不变，那么对于**案例2**也可以看到这一点。这可以说是理解这一过程的一种更自然的视角，因为它使我们能够理解在进行这一操作时所训练的底层概率图模型发生了什么：也就是说，我们正在消除对潜在状态的不确定性，从而从变分贝叶斯（自由能最小化）学习过程转向最大似然学习过程。为了证明这一点，我们可以考察按照（Whittington & Bogacz, 2017）进行**案例2** 缩放过程时，推理和学习方程会发生什么：

考虑将标签方差 Σ_L 按一个较大的常数因子 k进行缩放，随后将参数学习率（α）按 $\frac{1}{k}$ 进行缩放以进行补偿，这会得到以下修改后的推理和学习方程（之前的方程 15和 19），再次假设输出对数似然为高斯分布（平方欧几里得输出损失）

$$x_{\ell,t+1} = x_{\ell,t} - \gamma \left\{ \begin{array}{ll} \Sigma_\ell^{-1}(x_{\ell,t} - f_{\ell-1}(x_{\ell-1,t})) & \\ \quad - \dfrac{\partial f_\ell}{\partial x_\ell}\Big|_{x_{\ell,t}}^T \Sigma_{\ell+1}^{-1}(x_{\ell+1} - f(x_{\ell,t})) & \text{for } \ell = 1, ..., L-2 \\ \Sigma_{L-1}^{-1}(x_{L-1,t} - f_{L-2}(x_{L-2,t})) & \\ \quad - \dfrac{\partial f_{L-1}}{\partial x_{L-1}}\Big|_{x_{L-1,t}}^T (\dfrac{1}{k} \cdot \Sigma_L^{-1})(x_L - f(x_{L-1,t})) & \text{for } \ell = L-1 \end{array} \right\}$$

可改写为

$$x_{\ell,t+1} = x_{\ell,t} - \gamma' \left\{ \begin{array}{ll} (k \cdot \Sigma_\ell^{-1})(x_{\ell,t} - f_{\ell-1}(x_{\ell-1,t})) & \\ \quad - \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,t}}^{T} (k \cdot \Sigma_{\ell+1}^{-1})(x_{\ell+1} - f(x_{\ell,t})) & \text{for } \ell = 1, ..., L-2 \\ (k \cdot \Sigma_{L-1}^{-1})(x_{L-1,t} - f_{L-2}(x_{L-2,t})) & \\ \quad - \left.\frac{\partial f_{L-1}}{\partial x_{L-1}}\right|_{x_{L-1,t}}^{T} (1 \cdot \Sigma_L^{-1})(x_L - f(x_{L-1,t})) & \text{for } \ell = L-1 \end{array} \right\}$$

其中，对于推理，$\frac{1}{k}$ 因子已被吸收到新的步长 γ' 中，因此，只要欧拉积分不发散，推理结束时就会得到相同的收敛误差。随后，我们对参数 θ_ℓ 的更新为

$$\Delta\theta_\ell = \frac{1}{k}\alpha \left\{ \begin{array}{ll} \frac{\partial f_\ell}{\partial \theta_\ell}^T \Sigma_{\ell+1}^{-1}(x_{\ell+1} - f(x_{\ell,t})) & \text{for } \ell = 1, ..., L-2 \\ \frac{\partial f_\ell}{\partial \theta_\ell}^T (\frac{1}{k} \cdot \Sigma_L^{-1})(x_L - f(x_{L-1,t})) & \text{for } \ell = L-1 \end{array} \right\}$$

可改写为

$$\Delta\theta_\ell = \alpha \left\{ \begin{array}{ll} \frac{\partial f_\ell}{\partial \theta_\ell}^T (\frac{1}{k} \cdot \Sigma_{\ell+1}^{-1})(x_{\ell+1} - f(x_{\ell,t})) & \text{for } \ell = 1, ..., L-2 \\ \frac{\partial f_\ell}{\partial \theta_\ell}^T (1 \cdot \Sigma_L^{-1})(x_L - f(x_{L-1,t})) & \text{for } \ell = L-1 \end{array} \right\}$$

因此，**情形 2** 中的修改相当于将中间潜在状态的方差缩小至 0，同时保持输出似然方差不变。要与这个非潜在图对应的反向传播学习更新完全等价，就需要令 k 趋于无穷大；若要在实践中获得近似等价，则需要取一个相当大的值；（Whittington & Bogacz, 2017）在一个小型网络中采用了 k=100 。但这种缩放显然有数值不稳定的风险，在生物学上也不现实。

3.4 固定预测假设

在（Millidge、Tschantz 和 Buckley，2020）中提出了一种替代上述缩放过程的方法，称为“固定预测假设”（FPA）。在该方法中，整个推理过程中，与每个潜变量相对应的自上而下预测 $f(x_\ell)$ 和雅可比矩阵 $\left(\frac{\partial f_\ell}{\partial x_\ell}\right)$ 均固定为其前馈预测 $\mu_{\ell+1}$ 和 $\left(\frac{\partial f_\ell}{\partial x_\ell}|_{\mu_\ell}\right)$。

PC 的 FPA 修改将离散化梯度流中的以下项固定为与其前馈初始化相对应的特定值。也就是说

$$f_{\ell-1}(x_{\ell-1,t}) = f_{\ell-1}(x_{\ell,0}) = f_{\ell-1}(\mu_{\ell-1}) \tag{25}$$

$$f_\ell(x_\ell, t) = f_\ell(x_{\ell,0}) \quad = f_\ell(\mu_\ell) \tag{26}$$

$$\left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,t}} = \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{x_{\ell,0}} = \left.\frac{\partial f_\ell}{\partial x_\ell}\right|_{\mu_\ell} \tag{27}$$

其中我们假设了 **要求 1**，即潜在状态被初始化为相应计算链的前馈值。对于特定的潜在状态 x_ℓ，我们将与这些 FPA 修改相对应的推理更新方程称为 $u(x_{\ell,\,t},\ x_{\ell+1,\,t})$，以表明它仅取决于 $x_{\ell,\,t}$ 的瞬时值及其子变量 $x_{\ell+1,\,t}$，而不取决于父级潜在状态 $x_{\ell-1}$。每个潜在状态的动态与其父级状态值的这种解耦，意味着潜在状态的轨迹不再保证遵循联合对数的梯度（或标准预测编码中定义的自由能（方程 11））， 因此，对于标准 VPC中定义的模型，这些动态也不保证能够找到潜在状态的 MAP 估计，

$$x_{\ell,t+1} = x_{\ell,t} - \gamma\, u(x_{\ell,t}, x_{\ell+1,t}) \tag{28}$$

with

$$u(x_\ell, x_{\ell+1}) = \begin{Bmatrix} (x_\ell - \mu_\ell) - \frac{\partial f_\ell}{\partial x_\ell}^T (x_{\ell+1} - \mu_{\ell+1}) & \text{for } \ell = 1, ..., L-2 \\ (x_{L-1} - \mu_{L-1}) - \frac{\partial f_{L-1}}{\partial x_{L-1}}\Big|_{\mu_{L-1}}^T \frac{\partial f_L}{\partial \mu_L}^T & \text{for } \ell = L-1 \end{Bmatrix} \tag{29}$$

Millidge、Tschantz 和 Buckley，2020年提出，上述固定预测修改（方程 25 和27）等价于（Whittington& Bogacz，2017年）**案例2** 的无限方差极限。这是因为在输出方差无限大的极限情况下，到推理结束时，潜在状态的预测相较于其前馈初始化值几乎没有变化。然而，我们认为这种等价性并不成立，因为在**案例2**的无限方差或高方差极限情况下，潜在状态 x_ℓ 也会保持固定，或与其前馈值任意接近，而不仅仅是预测 $f_\ell(x_\ell)$。这种行为对于使推理准确对应于对潜在中间状态 x_ℓ的推理是必要的。在等价的反向传播模型所假设的中间潜在状态零方差（确定性）极限下，这些状态会变得等于前馈值μ_ℓ。

对于预测编码的FPA修改版本，情况并非如此，其中潜在状态不受约束，并且预测误差必然不为零，除非权重更新的幅度为零这一平凡情况。
这表明，FPA-PC推理既不对应于标准VPC所假设的概率图模型，也不对应于上述**案例2** 中缩放方差（即确定性潜在状态）修改后的概率图模型。这就引出了一个问题：如果不将FPA-PC推理视为在概率图模型上的推断（如标准VPC中那样），那么应如何解释FPA-PC推理？对此，一个尤为简单而优雅的答案是：可以将FPA-PC修改后的推断方程解释为通过一组常微分方程的稳态，直接实现递归反向传播方程。于是，FPA-PC推理可以被解释为执行欧拉积分方案，直至达到这一稳态。我们将在下一节中推导这一视角。

3.4.1 将 FPA-PC 推导为反向传播的稳态实现

通过展示 FPA-PC方程可以被解释为通过微分方程直接实现反向传播，我们可以进一步阐明 FPA-PC与反向传播之间的关系。为说明这一点，我们从反向传播出发反向推导，展示如何直接实现其递归关系，从而得到 FPA-PC 推理方程。

反向传播需要计算的关键量是通过递归表达式（公式 2）得到的误差 e_i。我们可以轻松地将这些表达式转换为一组简单微分方程的稳态形式。

$$\dot{e}_\ell = -\begin{Bmatrix} e_\ell - \frac{\partial \mu_{\ell+1}}{\partial \mu_\ell}^T e_{\ell+1} & \text{for } \ell = 1, \ldots, L-1 \\ e_\ell - \frac{\partial f_L}{\partial \mu_L}^T & \text{for } \ell = L \end{Bmatrix} \tag{30}$$

使得当 $\dot{e}_\ell = 0$时，我们获得所需的递归关系（方程 2）。注意，$\frac{\partial \mu_{\ell+1}}{\partial \mu_\ell}$ 和 $\frac{\partial E}{\partial \mu_L}$ 是固定的雅可比矩阵，需要在前馈值 μ_ℓ 处进行评估。还要注意，添加外部负号是为了更清楚地表明其与 FPA-PC 的关系，这不会影响 e_ℓ 在稳态下的值。

然后，我们可以将 e_ℓ 改写为某个变量 x_ℓ 减去常数前馈值 μ_ℓ 的形式，从而得到：$e_\ell = x_\ell - \mu_\ell$，据此可将上述动态方程表示为变量 x_ℓ 的形式：

$$\dot{x}_\ell = -\begin{Bmatrix} (x_\ell - \mu_\ell) - \frac{\partial \mu_{\ell+1}}{\partial \mu_\ell}^T (x_{\ell+1} - \mu_{\ell+1}) & \text{for } \ell = 1, \ldots, L-1 \\ (x_\ell - \mu_\ell) - \frac{\partial f_L}{\partial \mu_L}^T & \text{for } \ell = L \end{Bmatrix} \tag{31}$$

如果我们直接将 e_L 赋值（如 FPA-PC 所做的那样），并让其余误差保持动态，则可得到

$$\dot{x}_\ell = -\begin{Bmatrix} (x_\ell - \mu_\ell) - \frac{\partial \mu_{\ell+1}}{\partial \mu_\ell}^T (x_{\ell+1} - \mu_{\ell+1}) & \text{for } \ell = 1, \ldots, L-2 \\ (x_{L-1} - \mu_{L-1}) - \frac{\partial \mu_L}{\partial \mu_{L-1}}^T \frac{\partial f_L}{\partial \mu_L}^T & \text{for } \ell = L-1 \end{Bmatrix} \tag{32}$$

这恰好等同于 FPA-PC 的更新方程（28-29）。

这种修改版预测编码的实际后果首先在 (Rosenbaum, 2022) 中受到质疑，该文表明，当步长等于 1 时，FPA-PC 不仅在功能上而且在算法上等价于反向传播，

这表明至少在推理步骤大小等于 1 的特定情况下，FPA-PC 可能并不比反向传播有优势。我们扩展了这项工作，并进一步研究了 (Millidge、Tschantz 和 Buckley，2020) 以及其他近期修改 (Salvatori 等，2021；Song 等，2020) 的特性，表明这些修改对于任何步长都具有可证明更差的计算与时间复杂度。

3.4.2 计算与时间复杂度

为了研究 FPA-PC 和其他修改后的预测编码算法的计算与时间复杂度，我们首先证明，在上述动态机制下，误差从计算链中的输出节点传播到中间节点所经历的推理步骤数至少为该中间节点与输出节点之间的距离。参见定理 1 和推论 1.1。

定理1。对于计算链中相对于输出损失 $\mathcal{L}$具有非零梯度的特定节点（x_ℓ），在初始化为网络的前馈值后，在标准预测编码或 *FPA-PC*动力学下演化，该节点的误差（e_ℓ）将在时间 $t = L - \ell$首次变为非零。其中 L 是链的长度，ℓ 是该节点在链中的索引。

证明。令变分模态 x 初始化为网络的前馈值（即 **要求 1**）。也就是说，令 $x_\ell = \mu_\ell$ 对于 $\ell \in [1, ..., L-1]$。

让方程 28 和 29 描述我们潜在状态在 FPA-PC 动力学下的动态，方程 15 和16 描述标准预测编码下的动态。

> **1** 由于我们忽略了关于输出损失的梯度为零这一平凡情况，因此可以保证雅可比矩阵 $\frac{\partial f_\ell}{\partial x_\ell}$均非零，如 $\frac{\partial \mathcal{L}}{\partial x_\ell} = \frac{\partial f_\ell}{\partial x_\ell} \cdots \frac{\partial f_{L-1}}{\partial x_{L-1}} \frac{\partial \mathcal{L}}{\partial f_{\ell-1}} \neq 0$所示。**2** 根据更新方程的定义和**要求 1**，在时间 $t = 0$，除 x_{L-1}之外，所有节点的动力学 $u(x_\ell,\ x_{\ell+1})$ 均为 0。**3** 因此，在时间 $t = 1$，节点 x_{L-1} 按比例更新为 $u(x_L)$**4** 在任意时间步 $t = t + 1$，如果与节点 x_ℓ 相关的动力学 $u(x_\ell,\ x_{\ell+1})$ 在时间步 t 为 0，那么当且仅当 $x_{\ell+1}$ 在时间步 t 到 $t+1$ 之间发生变化时，它们才会在 $t+1$ 变为非零。（根据 FPA-PC（28 和 29）或标准预测编码（15 和 16）更新方程的定义，以及陈述 1）**5** 因此，通过归纳法，对于计算链中任意位置 ℓ 的任意变分模式 x_ℓ，其首次出现的非零变化将在时间 $t = L - \ell$ 发生，即其与输出之间的距离，按中间节点的数量计算。

□

推论 1.1. 该 *co* 计算链中与任意节点相关的误差的收敛时间
其下界为 *d* 到输出节点的节点距离。

证明。

> **1** FPA-PC 动力学方程的固定收敛点保证会收敛到一个等于输出损失梯度的误差。**2** 对于一个相对于输出损失具有非零梯度的节点而言，要实现收敛，该节点的收敛误差因此必须为非零。**3** 根据定理 1，对于长度为 L 的计算链，误差首次变为非零所需的时间等于其到输出节点 x_L 的距离，即对于 L 的计算链为 $t = L - \ell$。

□

换言之，该证明描述了：尽管 FPA-PC 中表面上进行的是并行计算，但要实现收敛，仍然需要足够的步数，让误差信息从输出节点向后传播。这是必需的，因为信息归根结底仍只能通过相邻节点传递，就像反向传播一样。因此，在收敛发生之前，推理必须至少执行与这一距离相等的推理步骤。

现在，我们可以考察在 FPA-PC 动力学下，单个推理步骤的时间复杂度是多少，进而考察 FPA-PC 整体推理的时间复杂度，以及将其与反向传播进行比较。需要注意的是，在这项比较中，对于 FPA-PC 和反向传播，我们都不计入与特定 θ 相关的梯度的最终向量-雅可比积（VJP）计算，因为该计算对两种算法具有相同的计算成本，并且只在推理和误差传播分别结束时执行一次。

在考量 FPA-PC推理的时间复杂度时，我们假设误差计算（$e_\ell = x_\ell - \mu_\ell$）对每个推理步骤的计算延迟影响可以忽略不计。对于中间误差而言，这通常是合理的假设，因为这些误差可以完全并行计算，因此其延迟相当于一次减法运算；而减法通常也是现代硬件上延迟最低的算术指令之一（参见 Fog & others，2011；Wong, Papadopoulou, Sadooghi-Alvandi, & Moshovos，2010）。

$$\text{TIME}\{\textit{FPA-PC}(F)\} = \text{TIME}\{\textit{forward}(F)\} + \text{TIME}\{\textit{FPA-Inf}(F)\} \tag{33}$$

我们假设每个推理步骤都能最大限度地并行执行，因此所有节点上的向量-雅可比积计算会同时进行。不过，由于每个推理步骤都必须等待最慢的（向量-雅可比积）计算完成，因此每个推理步骤的最小时间复杂度可视为等于最慢的向量-雅可比积计算的时间复杂度：（注意，若假设每个函数 f_i 的向量-雅可比积计算所需时间大致相同，也会得出类似结论）

$$\begin{aligned}
\text{TIME}\{\textit{FPA-PC}(F)\} &= \sum_{i=0}^{L} \text{TIME}\{\textit{forward}(f_i)\} + \sum_{t=0}^{t_c} \text{TIME}\{\textit{vjp}(f_{\text{slow}})\} && (34)\\
&\geq \sum_{i=0}^{L} \text{TIME}\{\textit{forward}(f_i)\} + \sum_{t=0}^{L} \text{TIME}\{\textit{vjp}(f_{\text{slow}})\} && (35)\\
&\geq w \sum_{i=0}^{L} \mathcal{C}_i + w \sum_{i=0}^{L} 2\mathcal{C}_{\max} && (36)\\
&\geq w \sum_{i=0}^{L} \mathcal{C}_i + w \sum_{i=0}^{L} 2\mathcal{C}_i && (37)\\
&\geq \text{TIME}\{\textit{backprop}(F)\} && (38)
\end{aligned}$$

其中 $C_{\max}$ = 最大值$\{C_i:\ i = 1, \ldots\ldots, L\}$ = 工作{向量-雅可比积(f_{slow})}。因此，对于等效的计算链，FPA-PC推理的时间复杂度经证明大于或等于反向传播的时间复杂度。出于以下一些原因，我们在此给出的界限接近 FPA-PC 的最佳情况：

1.我们假设误差计算（$e_\ell = x_\ell - \mu_\ell$）不会为每个推理步骤带来显著延迟。 2.除非在特定情况下（例如推理步骤大小等于 1），否则通常不会恰好在 $t_c = L$ 步数内收敛，因此在实践中，FPA-PC的时间复杂度可能显著高于反向传播（随推理长度扩展）。3. 瓶颈（缓慢的向量-雅可比积计算）将对FPA推理产生更显著的影响，因为每个推理步骤的速度都将与最慢的向量-雅可比积计算一样慢（产生 $t_c \cdot w \cdot C_{\max}$ 延迟成本），而单次缓慢的向量-雅可比积计算对于反向传播只会产生一次 $C_{\max}$ 延迟成本。

注意，这是与(Rosenbaum, 2022)的结果不同的独立证明，该结果显示对于特定步长为1，FPA-PC在 $t = L - \ell$ 步数内计算梯度（即与反向传播等效的步数）。然而尚不清楚(1) FPA-PC是否能在任意步长下以更少的时间步数收敛，以及(2) FPA-PC推理是否具有更差的时间复杂度（或更慢的运行时间），这取决于每个推理步骤的时间复杂度。这里我们证明FPA-PC的收敛可以证明永远不会比标准反向传播更快，对于任何 步长。

我们通过在不同规模的多层感知机网络上执行 FPA-PC推理，并改变推理步骤数与网络规模的相对关系，来检验这些结果。随后，我们计算梯度更新的余弦相似度，并将其与通过反向传播获得的梯度更新进行比较。正如预期，随着推理步骤数低于网络层数，余弦相似度会迅速下降，表明算法未能收敛（图 3）。当推理学习率低于 1 时，即使推理步骤数高于网络层数，算法仍无法收敛。我们还报告了验证集和测试集的准确率；在出现上述情况时，准确率也会显著下降（图 4 和图 5）。

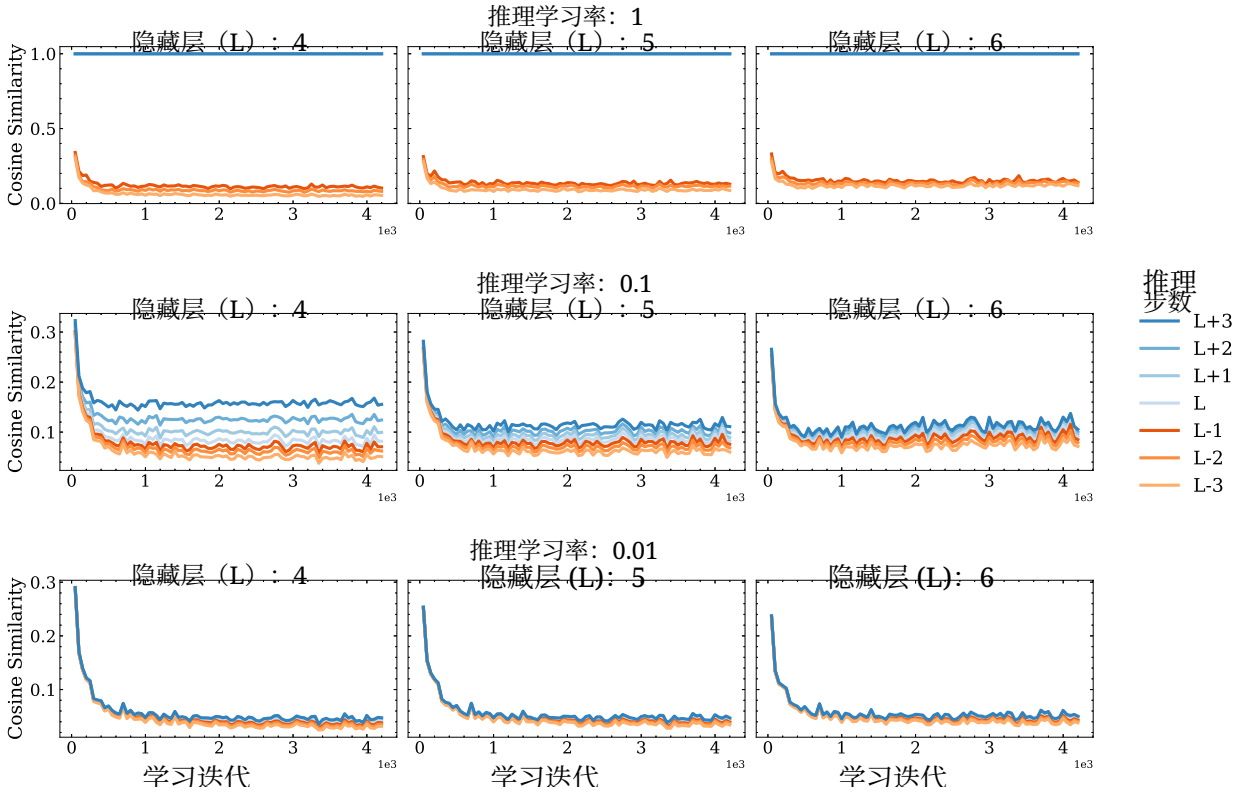

图 3：在 MNIST 上测试时，FPA-PC 学习更新与反向传播之间的余弦相似度随推理步数变化的情况，并以网络规模为参照。

3.5 零发散推理学习（Z-IL）

在对预测编码进行 FPA 修改后的表述之后，文献（Salvatori 等，2021；Song 等，2020）又提出了进一步修改的预测编码表述。由此得到的算法称为零发散推理学习 （Z-IL），无需显式固定前馈值和雅可比矩阵，便实现了与 FPA-PC 类似的效果，但需要进行一组非常特定的修改。具体而言，需要对标准预测编码算法进行以下修改：

- **要求 1：**（与前文相同）：变分模态或潜在状态（x_1，…，x_{L-1}）初始化为对应的基于反向传播的计算链（μ_1，…，μ_{L-1}）的前馈值
- **要求 2：**推理学习率 γ 设为 1
- **要求 3：**特定层仅在与其到计算链中输出节点的距离相对应的特定推理时间步（$t = L - \ell$）进行更新。

结果表明，去除这三项要求中的任何一项，都会导致 Z-IL 与反向传- 播失去等价性[1]。

该算法依赖于这样一个事实：在某个特定的时间步，与某个特定层 x_ℓ 对应的节点正在经历一种更新，而这种更新等价于在FPA-PC动力学下获得的更新。这是因为，任何特定节点（x_ℓ）的当前值在经过某个特定的（即 $t = L - \ell$）推理步骤之前，都不会偏离其初始化时的前馈值（参见（Song等，2020）补充材料中的引理 A.3）。对于与正在更新的权重相对应的节点而言，这起到了模拟 FPA 修正形式（公式25-27）下固定前馈值的效果，因为该节点的预测及其父节点（$\ell - 1$）的预测都保持固定为其前馈值。

[1]注意，出于本证明的目的，我们将忽略 FA-Z-IL（完全自主 Z-IL），它建立在 Z-IL 之上，因为其主要目的是消除在特定推理步骤手动触发权重更新这一神经生物学上不合理的要求，因此不会影响本节的结果。

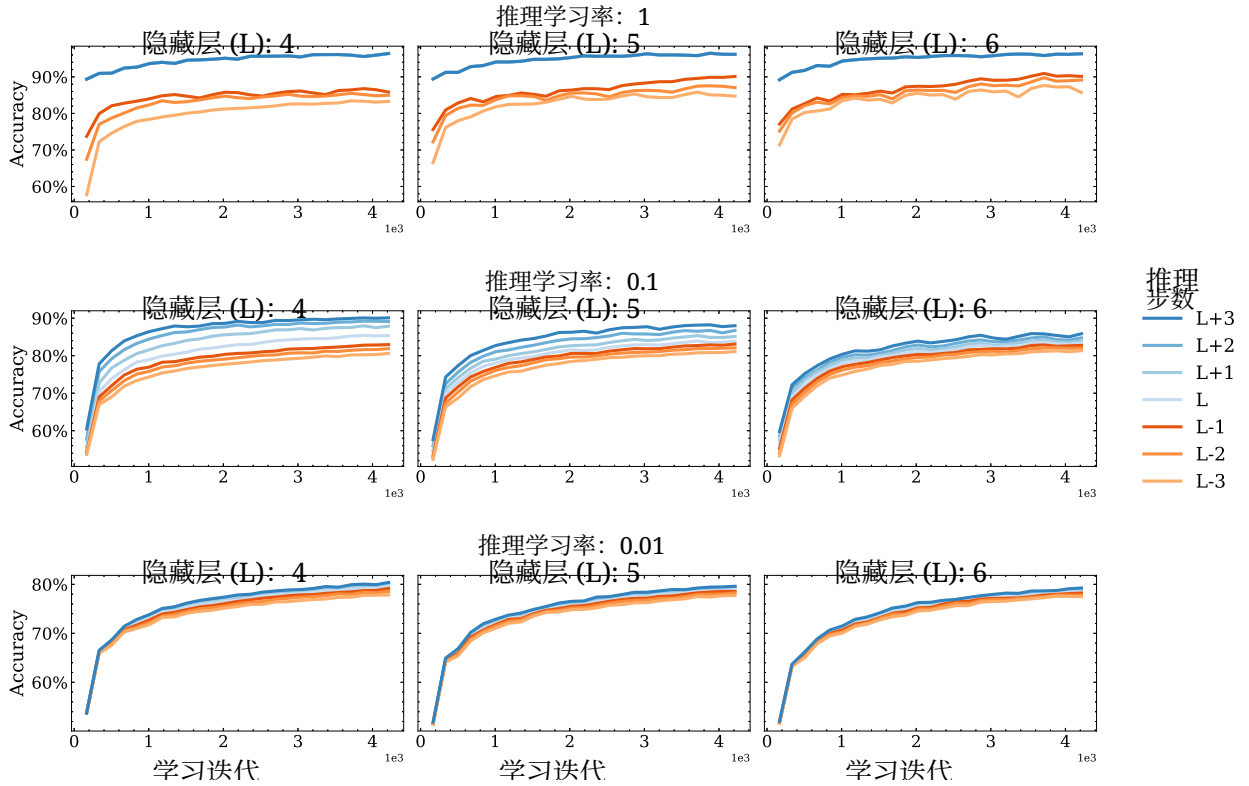

图 4：使用 FPA-PC 在 MNIST 验证集上的准确率，随网络推理步数相对于网络层数的变化而变化。

为了明确展示这一点，我们在表 3.5中描述了示例函数F在每个推理步骤中发生的相关计算。在每个推理步骤中，我们仅描述每个时间步长 ℓ中 x_ℓ、 e_ℓ 和 θ_ℓ 所发生的变化。因此，对于节点 $> \ell$上发生的任何计算，我们不会在下表中予以记录，因为它们不会对任何参数学习产生后续影响；同样，对于节点 $< \ell$，由于在该推理步骤中没有发生变化，我们也不会记录。为简化符号，我们省略时间索引；相反，表 3.5 中方程右侧的值均指前一时间步中的值。

步骤（T）	Z-IL	反向传播
$T = 0$	$e_L = \frac{\partial f_L}{\partial \mu_L}$	$e_L = \frac{\partial f_L}{\partial \mu_L}$
$T = 1$	$x_{L-1} = \underbrace{x_{L-1}}_{=\mu_{L-1}} + \frac{\partial f_{L-1}}{\partial \mu_{L-1}}^T e_L$ $e_{L-1} = x_{L-1} - \mu_{L-1}$	$e_{L-1} = \frac{\partial f_{L-1}}{\partial \mu_{L-1}}^T e_L$
$\vdots$	$\vdots$	$\vdots$
$T = L - 1$	$x_1 = \underbrace{x_1}_{=\mu_1} + \frac{\partial f_1}{\partial \mu_1}^T e_2$ $e_1 = x_1 - \mu_1$	$e_1 = \frac{\partial f_1}{\partial \mu_1}^T e_2$

我们可以看到，在每个时间步中，对于与参数梯度更新相关的计算，Z-IL 执行的计算与反向传播完全相同（在前馈值处计算的向量-雅可比积，以及随后产生的误差）；接着将这些结果加到一个常数（前馈激活 μ_ℓ）上，然后再减去该常数。除了这些计算之外，我们还注意到，Z-IL（以及由此延伸出的变体 FA-Z-IL）还进行了无效计算

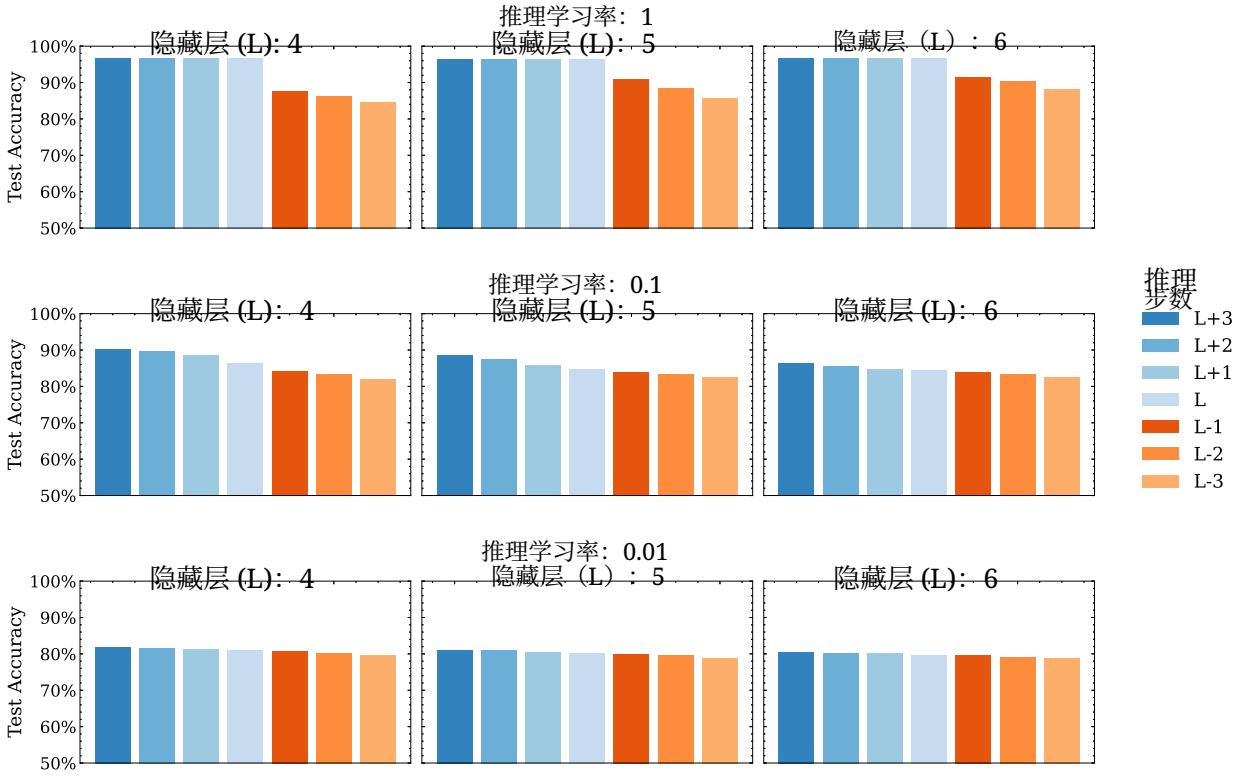

图 5：使用 FPA-PC 在 MNIST 测试集上取得的最终准确率，随推理步骤数相对于网络层数的变化情况。

节点 x_i，$i > \ell$ 的计算在上文的计算过程示意中并未展示。尽管这些计算不会产生或促成任何参数更新，却仍然会进行，因此是不必要的。

由于 Z-IL 的动态过程是标准预测编码动态过程的一个特例，我们可以再次使用定理 1、推论 1.1 以及与第 3.4 节相同的推理，再次获得 Z-IL 的复杂度界。我们发现，其复杂度下界与反向传播的复杂度下界相同。

$$\text{TIME}\{\textit{Z-IL}(F)\} \geq \sum_{i=0}^{L} \text{TIME}\{\textit{forward}(f_i)\} + \sum_{t=0}^{L} \text{TIME}\{\textit{vjp}(f_{\text{slow}})\} \tag{39}$$

$$\geq w \sum_{i=0}^{L} \mathcal{C}_i + w \sum_{i=0}^{L} 2\mathcal{C}_{\max} \tag{40}$$

$$\geq w \sum_{i=0}^{L} \mathcal{C}_i + w \sum_{i=0}^{L} 2\mathcal{C}_i \tag{41}$$

$$\geq \text{TIME}\{\textit{backprop}(F)\} \tag{42}$$

3.6 广义 IL（G-IL）、学习率稳定性与在线学习

为求完整起见，我们指出，近期有些研究表明，某些无法产生与反向传播等价更新的预测编码（PC）及相关基于能量的模型，仍可能对较高学习率具有更强的鲁棒性，并且在在线学习（批量大小为 1）设置下性能下降较少（Alonso 等，2022；Song 等，2022）。具体而言，Song 等（2022）表明，加入额外正则化项的预测编码变体——广义 IL（G-IL）——可以近似隐式随机梯度下降（SGD），且在某些极限条件下，该近似可达到精确。该研究的作者还提出了一种改进算法（IL-prox），可保证其更新与隐式 随机梯度下降相同。这些方案在学习率方面表现出更强的鲁棒性，在在线学习场景下性能下降也较少。这些特性也通过实证研究得到了验证，例如，

一般而言，基于能量的模型（EBM）也具有这些特性，这是由 Alonso 等（2022）提出的一项名为“前瞻性配置”的原理所致。
这些发现颇为引人注目，因为对学习率的稳健性符合对后向欧拉积分格式的预期——后者是随机梯度下降所源自的前向欧拉积分格式的隐式对应形式——也是一种在非随机情境下经证实更不易发散的优化方法。

然而，目前尚不清楚，在与Adam或带动量的随机梯度下降等优化器进行比较时，学习率稳定性方面的差异是否仍然存在。这两种优化器无论在理论上还是实践中都已有充分研究，能够通过改善损失景观的条件数或抑制振荡，降低高学习率下发散的风险。此外，有证据表明，小批量随机梯度下降本身对学习率具有更强的鲁棒性（Masters & Luschi, 2018；Shallueetal.,2019），这可能是因为小批量设置中的额外随机性阻止了欧拉积分误差的累积。因此，作者认为，未来研究探索在线学习设置下，当在线随机梯度下降与高学习率结合时，性能差距是否仍然存在，将是一个很有吸引力的研究方向。

尽管这些发现不会影响我们所呈现的结果，也不会影响将预测编码作为反向传播直接替代方案的用法，但它们引发了一些值得深思的问题：即便修改版和未修改版的预测编码未必会产生与反向传播完全相同的参数更新，它们是否仍可能具备神经形态学习所期望的特性。

4 结论与关键结果

在本文中，我们研究了各种预测编码变体相较于反向传播算法的计算效率、学习动力学和神经生物学合理性。现总结主要结果和要点：

> **结果 1** 来自 (Whittington & Bogacz, 2017) 的无限方差极限修正（见上文案例 2），等价于假设一个具有严格确定性（非概率性）潜在状态的模型。**结果 2** 预测编码的修正变体：FPA-PC (Millidge、Tschantz 和 Buckley，2020)，以及 Z-IL (Salva- tori et al., 2021; Song 等，2020)，不遵循自由能梯度。推论：学习并不对应于通过变分贝叶斯学习潜在概率模型。**结果 3** 预测编码的修正变体：FPA-PC 和 Z-IL，保证具有比反向传播更差的时间复杂度，即使考虑到每个推理步骤内的完全并行计算。（见定理 1、相应的推论 1.1 以及方程 33-42）**要点 1** 预测编码与反向传播的等价性或近似等价性要求采用反转方案，这与生成式预测编码的传统公式不同，对神经生物学合理性具有非平凡的影响。**要点 2** 标准预测编码的朴素实现不会遭受与朴素神经生物学反向传播提案中存在的权重 传输 问题相同意义上的问题。然而，它确实遭受类似但可能不那么不合理的权重 对称问题。

这些结果让人对预测编码变体相较于反向传播及其神经形态实现可能带来的优势产生怀疑，因为那些产生等效或近乎等效梯度更新的预测编码变体，执行的计算也与反向传播完全相同，且方式同样具有局部性或非局部性。尽管预测编码网络表面上采用并行计算，但从误差反向传播的角度看，这只会带来额外或未被利用的计算，并不会加快误差传播。

此外，通过引入修改以实现或近似等价，预测编码变体失去了标准预测编码公式的生成式变分贝叶斯解释及其多项优势，例如保留对潜在隐状态或原因的不确定性，或与我们目前对大脑皮层内通路组织和功能特性的理解保持一致。

致谢

作者谨向 Christopher L. Buckley 教授及其团队致谢，感谢他们对本稿提出宝贵意见和反馈。

参考文献

Akrout, M., Wilson, C., Humphreys, P., Lillicrap, T., & Tweed, D. B. (2019). 无权重传输的深度学习。载于《神经信息处理系统进展》（第 32 卷）。Curran Associates, Inc. 检索日期：2023-01-13，网址：https://papers.nips.cc/paper/2019/hash/f387624df552cea2f369918c5e1e12bc -Abstract.html Alonso, N., Millidge, B., Krichmar, J., & Neftci, E. (2022, May). 推理学习的理论框架。arXiv。检索日期：2023-04-03，网址：http://arxiv.org/abs/2206.00164 (arXiv:2206.00164 [cs]) doi: doi:10.48550/arXiv.2206.00164 Amit, Y. (2019). 非对称连接与赫布更新的深度学习。《计算神经科学前沿》，13。检索日期：2023-01-13，网址：https://www.frontiersin.org/articles/10.3389/ fncom.2019.00018 Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012, November). 预测编码的经典微回路。《Neuron》，76(4), 695–711。doi: doi:10.1016/j.neuron.2012.10.038 Baur, W., & Strassen, V. (1983, February). 偏导数的复杂度。《理论计算机科学》，22(3), 317–330。检索日期：2022-12-06，网址：https://www.sciencedirect.com/science/article/pii/ 030439758390110X doi: doi:10.1016/0304-3975(83)90110-X Bogacz, R. (2017, February). 用于建模感知与学习的自由能框架教程。《数学心理学杂志》，76, 198–211。检索日期：2021-07-02，网址：https://www.sciencedirect.com/science/article/pii/S0022249615000759 doi: doi:10.1016/j.jmp.2015.11.003 Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., . . . Zhang, Q. (2018). JAX：Python+NumPy 程序的可组合变换。检索自：http://github.com/google/jax Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017, December). 行动与感知的自由能原理：数学综述。《数学心理学杂志》，81, 55–79。检索日期：2022-06-13，网址：https://www.sciencedirect.com/science/article/pii/S0022249617300962 doi: doi:10.1016/j.jmp.2017.09.004 Crick, F. (1989, January). 关于神经网络的最新热潮。《Nature》，337(6203), 129–132。检索日期：2022-01-14，网址：https://www.nature.com/articles/337129a0 (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6203 Primary_atype: 评论与观点 Publisher: Nature Publishing Group) doi: doi:10.1038/337129a0 Feldman, H., & Friston, K. (2010). 注意、不确定性与自由能。《人类神经科学前沿》，4, 215。检索日期：2021-08-29，网址：https://www.frontiersin.org/article/10.3389/fnhum.2010.00215 doi: doi:10.3389/fnhum.2010.00215 Felleman, D. J., & Van Essen, D. C. (1991). 灵长类大脑皮层中的分布式层级处理。《Cerebral Cortex》（纽约，纽约州：1991），1(1), 1–47。doi: doi:10.1093/cercor/1.1.1-a Fog, A., & others. (2011). 指令表：Intel、AMD 和 VIA CPU 的指令延迟、吞吐量及微操作分解列表。哥本哈根大学工程学院，93, 110。Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., & Roseboom, W. (2022, June). 情景记忆与时间感知的预测处理模型。《神经计算》，34(7), 1501–1544。检索自：https://doi.org/10.1162/neco_a_01514 doi: doi:10.1162/neco_a_01514Friston, K. (2003, November). 大脑中的学习与推理。《神经网络》，16(9), 1325–1352。检索日期：2021-07-28，网址：https://linkinghub.elsevier.com/retrieve/pii/S0893608003002454 doi: doi:10.1016/j.neunet.2003.06.005 Friston, K. (2005, April). 大脑皮层反应理论。《英国皇家学会哲学汇刊 B：生物科学》，360(1456), 815–836。检索日期：2022-06-23，网址：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569488/ doi: doi:10.1098/rstb.2005.1622 Friston, K. (2008, November). 大脑中的层级模型。《PLOS 计算生物学》，4(11), e1000211。检索日期：2022-09-28，网址：https://journals.plos.org/ploscompbiol/article?id=10.1371/journal .pcbi.1000211 (Publisher: 公共科学图书馆) doi: doi:10.1371/journal.pcbi.1000211 Friston, K., & Kiebel, S. (2009, May). 自由能原理下的预测编码。《伦敦皇家学会哲学汇刊 B：生物科学》，364(1521), 1211–1221。doi: doi:10.1098/rstb.2008.0300 Griewank, A. (1997, February). 自动微分。Griewank, A. (2009). 梯度、雅可比矩阵与海森矩阵的复杂度。《优化百科全书》中，C. A. Floudas & P. M. Pardalos（编），425–435页。波士顿，马萨诸塞州：Springer US。检索日期：2022-12-27，网址：https://doi.org/10.1007/978-0-387-74759-0_78 doi: doi:10.1007/978-0-387-74759-0 78 Griewank, A., & Walther, A. (2008). 导数求值。Grossberg, S. (1987, January). 竞争性学习：从交互激活到自适应共振。《认知

Science, 11(1), 23–63. 检索于2022-12-07，来自 https://www.sciencedirect.com/science/article/pii/S0364021387800253 doi: doi:10.1016/S0364-0213(87)80025-3 Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., . . . Sifre, L. (2022年3月). 训练计算最优的大语言模型. arXiv. 检索于2023-01-02，来自 http://arxiv.org/abs/ 2203.15556 (arXiv:2203.15556 [cs]) doi: doi:10.48550/arXiv.2203.15556 Hohwy, J., Roepstorff, A., & Friston, K. (2008年9月). 预测编码解释双眼竞争：一项认识论综述. Cognition, 108(3), 687–701. doi: doi:10.1016/j.cognition.2008.05.010 Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015年5月). 大脑层级：预测处理、精度和丘脑枕. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1668), 20140169. 检索于2021-08-11，来自 https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0169 (出版商：Royal Society) doi: doi:10.1098/rstb.2014.0169 Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., . . . Amodei, D. (2020年1月). 神经语言模型的缩放定律. arXiv. 检索于2023-01-02，来自 http://arxiv.org/abs/2001.08361 (arXiv:2001.08361 [cs, stat]) doi: doi:10.48550/arXiv.2001.08361 Kinghorn, P. F., Millidge, B., & Buckley, C. L. (2022年9月). 防止预测编码网络中分类准确率的下降. arXiv. 检索于2023-03-07，来自 http://arxiv.org/abs/2208.07114 (arXiv:2208.07114 [cs, q-bio] version: 2) doi: doi:10.48550/arXiv.2208.07114 Kolen, J., & Pollack, J. (1994年6月). 无权重传输的反向传播. 载于 Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94) (第3卷, 第1375–1380页 vol.3). doi: doi:10.1109/ICNN.1994.374486 Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016年11月). 随机突触反馈权重支持深度学习的误差反向传播. Nature Communications, 7(1), 13276. 检索于2022-12-07，来自 https://www.nature.com/articles/ncomms13276 (编号：1 出版商：Nature Publishing Group) doi: doi:10.1038/ncomms13276 Linnainmaa, S. (1976年6月). 累积舍入误差的泰勒展开. BIT Numerical Mathematics, 16(2), 146–160. 检索于2022-12-18，来自 https://doi.org/10.1007/BF01931367 doi: doi:10.1007/BF01931367 Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., . . . Kennedy, H. (2014). 层级解剖：猕猴视觉皮层中的前馈和反馈通路. Journal of Comparative Neurology, 522(1), 225–259. 检索于2023-01-28，来自 https://onlinelibrary.wiley.com/doi/abs/10.1002/ cne.23458 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.23458) doi: doi:10.1002/cne.23458 Masters, D., & Luschi, C. (2018年4月). 重新审视深度神经网络的小批量训练. arXiv. 检索于2023-04-03，来自 http://arxiv.org/abs/1804.07612 (arXiv:1804.07612 [cs, stat]) doi: doi:10.48550/arXiv.1804.07612 Millidge, B., Salvatori, T., Song, Y., Bogacz, R., & Lukasiewicz, T. (2022年2月). 预测编码：迈向超越反向传播的深度学习未来? arXiv. 检索于2022-12-14，来自 http://arxiv.org/ abs/2202.09467 (arXiv:2202.09467 [cs]) doi: doi:10.48550/arXiv.2202.09467 Millidge, B., Tschantz, A., & Buckley, C. L. (2020年10月). 预测编码沿任意计算图近似反向传播. arXiv:2006.*04182 [*cs*]*. 检索于2021-10-06，来自 http://arxiv.org/abs/ 2006.04182 (arXiv: 2006.04182) Millidge, B., Tschantz, A., Seth, A., & Buckley, C. L. (2020b, 10月). 放宽预测编码模型的约束. arXiv. 检索于2022-12-08，来自 http://arxiv.org/abs/2010.01047 (arXiv:2010.01047 [cs, q-bio, stat]) doi: doi:10.48550/arXiv.2010.01047 Millidge, B., Tschantz, A., Seth, A. K., & Buckley, C. L. (2020a, 10月). 激活松弛：大脑中反向传播的局部动态近似. arXiv:2009.*05359 [*cs, q-bio*]*. 检索于2022-01-21，来自 http://arxiv.org/abs/2009.05359 (arXiv: 2009.05359) Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017年). PyTorch中的自动微分. Pawlak, V., Wickens, J. R., Kirkwood, A., & Kerr, J. N. D. (2010年10月). 时机并非一切：神经调制打开STDP之门. Frontiers in Synaptic Neuroscience, 2, 146. 检索于2023-01-13，来自 https:// www.ncbi.nlm.nih.gov/pmc/articles/PMC3059689/ doi: doi:10.3389/fnsyn.2010.00146 Rao, R. P. N., & Ballard, D. H. (1999年1月). 视觉皮层中的预测编码：对一些超经典感受野效应的功能解释. Nature Neuroscience, 2(1), 79–87. 检索于2022-05-17，来自 https://www.nature.com/articles/nn0199_79 (编号：1 出版商：Nature Publishing Group) doi: doi:10.1038/4580 Rosenbaum, R. (2022年3月). 论预测编码与反向传播的关系. PLOS ONE, 17(3), e0266102. 检索于2022-04-12，来自 https://journals.plos.org/plosone/article?id=10.1371/ journal.pone.0266102 (出版商：Public Library of Science) doi: doi:10.1371/journal.pone.0266102 Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986年10月). 通过反向传播误差学习表征。

Nature, *323*(6088), 533–536. 检索于2022-12-14，来自 https://www.nature.com/articles/323533a0 (编号：6088 出版商：Nature Publishing Group) doi: doi:10.1038/323533a0 Salvatori, T., Song, Y., Lukasiewicz, T., Bogacz, R., & Xu, Z. (2021年3月). 预测编码可在卷积和循环神经网络上实现精确反向传播。 arXiv. 检索于2023-01-27，来自 http://arxiv.org/abs/2103.03725 (arXiv:2103.03725 [cs]) doi: doi:10.48550/arXiv.2103.03725 Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., & Dahl, G. E. (2019). 测量数据并行对神经网络训练的影响. *Journal of Machine LearningResearch*, *20*(112), 1–49. 检索于2023-04-03，来自 http://jmlr.org/papers/v20/18-789.html Shipp, S. (2016). 预测编码的神经元件. *Frontiers in Psychology*, *7*, 1792. 检索于2021-10-10，来自 https://www.frontiersin.org/article/10.3389/fpsyg.2016.01792 doi: doi:10.3389/fpsyg.2016.01792 Sjöström, P. J., & Häusser, M. (2006年7月). 合作开关决定新皮层锥体神经元远端树突中突触可塑性的符号. *Neuron*, *51*(2), 227–238. 检索于2023-01-13，来自 https://www.sciencedirect.com/science/article/pii/S0896627306004715 doi: doi:10.1016/j.neuron.2006.06.017 Song, Y., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2020). 大脑能进行反向传播吗？——预测编码网络中反向传播的精确实现. 载于 *Advances in Neural Information Processing Systems* (第33卷, 第22566–22579页). Curran Associates, Inc. 检索于2022-04-12，来自 https://proceedings.neurips.cc/paper/2020/hash/fec87a37cdeec1c6ecf8181c0aa2d3bf-Abstract.html Song, Y., Millidge, B., Salvatori, T., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2022年5月). 可塑性之前推断神经活动：超越反向传播的学习基础。 bioRxiv. 检索于2023-04-03，来自 https://www.biorxiv.org/content/10.1101/2022.05.17.492325v1 (页码：2022.05.17.492325 栏目：New Results) doi: doi:10.1101/2022.05.17.492325 Urbanczik, R., & Senn, W. (2014年2月). 通过树突预测体细胞放电进行学习 . *Neuron*, *81*(3), 521–528. doi: doi:10.1016/j.neuron.2013.11.030 Whittington, J. C. R., & Bogacz, R. (2017年5月). 具有局部Hebbian突触可塑性的预测编码网络中误差反向传播算法的近似. *NeuralComputation*, *29*(5), 1229–1262. doi: doi:10.1162/NECO a 00949¯ ¯Wong, H., Papadopoulou, M.-M., Sadooghi-Alvandi, M., & Moshovos, A. (2010年3月). 通过微基准测试揭秘GPU微架构. 载于 *2010 IEEE InternationalSymposium on Performance Analysisof Systems & Software (ISPASS)*(第235–246页). doi: doi:10.1109/ISPASS.2010.5452013 Zahid, U., Guo, Q., Friston, K., & Fountas, Z. (2023年3月). 具有近似拉普拉斯蒙特卡洛的曲率敏感预测编码。 arXiv. 检索于2023-03-13，来自 http://arxiv.org/abs/2303.04976 (arXiv:2303.04976 [cs]version: 1) doi: doi:10.48550/arXiv.2303.04976 Zipser, D., & Rumelhart, D. E. (1993). 新学习模型的神经生物学意义。