arXiv:2611.28534v1 [cs.ET] 21 Nov 2026

用于 sparsemax 注意力的耦合动力系统

Lianlong Sun*
大学 ity of Roch
校lianlongsun@rochester.edu

Michael Huang
University of Roch
校michael.huang@rochester.edu

摘要

我们提出了一系列用于计算 sparsemax 注意力的耦合动力系统。每个系统都通过一次物理弛豫，将查询-键相似度与归一化融合起来。具体而言，我们将 sparsemax 注意力表述为一个约束能量最小化问题，并推导出一对耦合微分方程：一个驱动能量最小化，另一个则施加约束。在平衡状态下，系统状态恰好给出 sparsemax 注意力概率。我们采用连续伊辛自旋和 Kuramoto 振荡器实现了这一表述。这些弛豫过程可原生运行于新兴的非冯·诺依曼平台，例如伊辛机；与数字处理器相比，伊辛机已展现出高出数个数量级的能效，但迄今主要用于无约束二次组合优化。由于模拟硬件的精度有限且存在噪声，我们在这些非理想因素下对两种系统进行了模拟，结果显示推理质量会平稳退化。我们的表述将这些新兴计算基底的应用范围从组合优化拓展至注意力，为面向 Transformer 推理的新一类物理加速器开辟了道路。

1 引言

Transformer 是现代深度学习中的主流架构 [1]。它最初是为机器翻译而提出的，如今已成为大语言模型 [2, 3] 和视觉模型 [4, 5]的基础。其核心组件是注意力机制，它使每个词元都能直接从序列中的其他词元获取信息。对于每个词元，注意力机制会将其查询与其他词元的键进行比较，再将所得得分转换为注意力概率，并根据这些概率对值求加权平均。Transformer 会在多个注意力头和层中重复这一操作，其计算成本会随序列长度呈二次增长。将得分转换为概率的归一化方式会影响注意力的行为。标准选择是 softmax，它会为每个键分配非零概率。sparsemax [6] 和 entmax [7] 等稀疏归一化器则会为得分较低的键分配恰好为零的概率。它们既能确保概率之和为一，又能让每个查询专注于

键更少，因此注意力更容易解释 [6, 8]。

伊辛机是通过自然弛豫进行计算的物理系统 [9, 10]。其自旋通过可编程耦合器相互作用，并最终达到伊辛能量较低的状态，该状态编码了问题的解 [11]。伊辛机已基于多种计算基底构建，包括量子退火机 [12],、相干光学机 [13–16], 、耦合振荡器 [17–22],和 CMOS 电路 [23–26]。与数字处理器相比，它们已展现出高出数个数量级的能效 [10, 21, 24]。其应用包括最大割 [14, 17, 27, 28], 、布尔可满足性 [22, 29–31]和 MIMO 检测 [32, 33]。其他应用包括投资组合优化、车辆路径规划、聚类、化学设计和电力系统 [34–39]。近期研究还训练玻尔兹曼网络 [40],、学习图和函数 [41–44], , 并生成图像 [45]。然而，几乎所有这些应用都遵循一个共同原理。在模型层面，机器运行着相同的简单动力学：每个自旋都受到其与其他自旋之间的耦合以及偏置的驱动。二次无约束二元优化

*通讯作者。

预印本。

（QUBO）问题 [11] 可以自然地映射到该模型中。注意力不符合这种形式。少数研究使用伊辛模型对离散注意力模式进行采样或搜索 [46–48]。这些研究都没有让物理系统弛豫至注意力概率本身。

我们重新审视伊辛机的效率来源。它通过在少数简单的模拟基元上进行物理弛豫来计算，例如自旋、耦合器和偏置电流。我们在保留这些基元的同时改变动力学。新的动力学必须满足两个目标。首先，它们必须超越简单的伊辛动力学，并具备计算 Transformer 注意力的表达能力。其次，它们必须能够自然地映射到伊辛机或类似的模拟计算基底上。因此，现有的硬件设计可以大体复用，并保留效率增益。其结果不再是简单的伊辛动力学。我们将其称为耦合动力系统，而不是伊辛机。它将自旋动力学与用于施加约束的第二种动力学耦合起来。

Sparsemax 注意力是这一扩展的自然目标。Sparsemax 是一个二次规划的解 [6]。其目标是二次能量，与伊辛机中的情况相同。新的挑战在于要求概率之和为一这一约束。通过该规划的拉格朗日函数，我们推导出一对耦合常微分方程（ODEs）。第一条方程主要降低能量。每个键都在自身的查询-键得分和一个共享反馈信号作用下弛豫。第二条方程主要施加约束。它将概率之和与一进行比较，并调整反馈信号。在平衡态，系统状态恰好给出 sparsemax 注意力概率。因此，查询-键相似度与归一化被融合到单一的物理弛豫过程中，相较于传统数字计算，这可能带来效率增益。

我们用连续伊辛自旋和 Kuramoto 振荡器 [49] 实现这一表述。每个键对应一个自旋或振荡器，它与查询之间的耦合构成得分。每一项都映射到伊辛机硬件已有的一种原语 [17, 24, 50]。一个约束节点运行第二个常微分方程。在硬件层面，伊辛机通常通过模拟电路计算，在弛豫过程中无需进行模数转换。模拟基底会受到精度有限和噪声的影响 [51]。因此，我们在量化和读出噪声条件下模拟这两个系统，以衡量其推理质量。这也是对耦合动力系统的评估。

我们的贡献如下。

- 我们提出了一系列耦合动力系统，将简单的伊辛动力学加以扩展。我们推导出一对耦合常微分方程，其平衡态恰好对应 sparsemax 注意力。
- 我们将动力学映射到伊辛机原语，采用连续伊辛自旋和 Kuramoto 振荡器。这种映射为在伊辛机及类似模拟基底上构建注意力提供了一套方案，相较于传统数字计算，可能带来效率提升。
- 我们在 Transformer 内部模拟这两个系统，并评估它们在量化和读出噪声下的推理质量。结果表明，这两个系统的质量都会平稳下降。这为未来的硬件实现奠定了基础。

2 方法

Sparsemax 可以表述为一个二次规划的解 [6]。我们首先将该规划转化为一对耦合常微分方程。然后，我们将每一项映射到伊辛机的基本元件上，分别采用连续伊辛自旋和 Kuramoto 振荡器。附录 A 给出了两者的最简实现。

2.1 从 sparsemax 到耦合常微分方程

设 $q \in [-1,1]^d$ 为一个查询，n 键构成 $K \in \mathbb{R}^{n\times d}$的各行。键 i 的得分为 $s_i = \sum_j K_{ij} q_j/\sqrt{d}$。Sparsemax 将得分映射到概率单纯形 [6]上距离最近的点：

$$\mathrm{sparsemax}(s) = \arg\min_{a_i \geq 0,\ \sum_i a_i = 1} \tfrac{1}{2}\|a - s\|^2 = [s - \tau]_+. \tag{1}$$

这里，a 是注意力概率向量，每个键对应一个元素 a_i。括号中的内容为 $[\cdot]_+$ =max(·, 0)，阈值 τ 使输出之和为 1。$s_i \leq \tau$ 对应的键的概率为 0。我们为和为 1 这一约束引入拉格朗日乘子 μ，并将每个概率 a_i 保持在 $[0,1]$内。

由此得到拉格朗日函数

$$\mathcal{L}(a,\mu)=\tfrac{1}{2}\|a-s\|^2+\mu\Big(\sum_i a_i-1\Big),\qquad 0\le a_i\le 1. \tag{2}$$

a 中的下降和 μ 中的上升构成了这组耦合常微分方程：

$$\frac{da_i}{dt}=s_i-a_i-\mu\quad\text{with } a_i\in[0,1],\qquad\qquad \frac{d\mu}{dt}=\sum_i a_i-1. \tag{3}$$

这里， t 表示时间。第一个方程主要用于降低能量 $\frac{1}{2}\|a-s\|^2$。每个键只需要自己的得分和共享反馈信号 μ。第二个方程主要用于施加约束，使概率之和为一。一个节点对概率求和，并移动 μ，直到总和为一。我们称其为约束节点。在平衡态，每个概率都被 $s_i-\mu$ 截断到 $[0,1],$，且概率之和为一。这正是 sparsemax，其中 μ 充当 τ。

2.2 连续 Ising 自旋

We 从线性 Ising 机开始。在模型层面，每个自旋 i 都有一个连续状态 x_i，并 d foll遵循线性动力学：

$$\frac{dx_i}{dt}=\sum_j J_{ij}x_j+h_i,\qquad x_i\in[-1,1]. \tag{4}$$

这里， J_{ij} 是从自旋 j 到自旋 i的耦合， h_i 是自旋 i的偏置。上限 $[-1,1]$ 是模型的一部分。在硬件层面，x_i 是自旋电容器上的电压，耦合器和偏置电流会对其充电 [24, 26]。此上限即为电路的电压上限。

映射。 我们使用两组自旋。 d 查询自旋保存输入，查询自旋 j 被钳制在 q_j。 n 键自旋保存输出，键自旋 i 给出概率 $a_i=(x_i+1)/2$。此读出将自旋范围 $[-1,1]$ 映射到概率范围 $[0,1]$。我们通过以下方式对机器进行编程：

$$J_{ij}=\frac{2K_{ij}}{\sqrt{d}},\qquad J_{ii}=-1,\qquad h_i=-1-2\mu. \tag{5}$$

从查询自旋 j 到键自旋 i 的耦合 J_{ij} 存储键条目 K_{ij}。自耦合 J_{ii} 将每个键自旋拉回到其输入值。偏置 h_i 对所有键自旋都相同，并由约束节点通过 μ设置。键自旋之间没有耦合。

动力学。 在这些设置和约束节点的作用下，机器运行

$$\frac{dx_i}{dt}=\underbrace{\sum_j J_{ij}q_j}_{2s_i}-x_i-1-2\mu,\qquad \frac{d\mu}{dt}=\sum_i\frac{x_i+1}{2}-1. \tag{6}$$

当 $a_i=(x_i+1)/2$ 时，这些正是公式（3）中的耦合常微分方程。键自旋通过共享偏置相互竞争。当它们的概率之和大于 1 时， μ 会上升并降低每个键自旋的值。取值为 $s_i\le\mu$ 的键自旋会稳定在边界 $x_i=-1$，其概率恰好为零。其他键自旋则稳定在边界内，并对应非零概率。

2.3 Kuramoto 振荡器

基于振荡器的伊辛机（OIM）将每个自旋存储在振荡器的相位 θ_i 中 [17]。在模型层面，相位遵循带有二次谐波注入的 Kuramoto 动力学 [49]：

$$\frac{d\theta_i}{dt}=-\sum_j J_{ij}\sin(\theta_i-\theta_j)-K_s\sin 2\theta_i. \tag{7}$$

这里 J_{ij} 是振荡器 i 与 j之间的耦合， K_s 是二次谐波注入的强度。在硬件层面，耦合器将振荡器连接起来，并向每个振荡器注入二次谐波信号。我们复用这些基本元件。

映射。 我们使用三种振荡器。每个键对应一个相位为 θ_i 的振荡器。每个查询分量由一对具有固定相位 $\pm\varphi_j$ 的振荡器承载。参考振荡器的固定相位为 0。一对振荡器作用于一个键振荡器时，相当于一个强度为 2 cos φ_j 的输入，因为 sin($\theta-\varphi$)+sin($\theta+\varphi$) =2 cos φ sin θ。我们对机器进行编程，使其

$$\cos\varphi_j = q_j, \qquad J_{ij} = \frac{K_{ij}}{4\sqrt{d}}, \qquad J_{i0} = -\mu, \qquad K_s = -\frac{1}{2}. \tag{8}$$

耦合 J_{ij} 将键振荡器 i 连接到成对振荡器 j 中的两个振荡器，并存储键条目 K_{ij}。耦合 J_{i0} 将键振荡器 i 连接到参考振荡器，而约束节点通过 μ设置它。

动力学。 在这些设置下，键振荡器运行

$$\frac{d\theta_i}{dt} = -\sum_j J_{ij}\Big[\sin(\theta_i-\varphi_j)+\sin(\theta_i+\varphi_j)\Big] + \mu\sin\theta_i + \frac{1}{2}\sin 2\theta_i. \tag{9}$$

这是公式（7）的 OIM 动力学，采用公式（8）中的设置。与查询对的耦合向键振荡器 i传递 $-\frac{s_i}{2}$sin $\theta_{\overline{i}}$。读出值是相位余弦的整流结果。约束节点将读出值之和与二分之一进行比较:

$$r_i = [\cos\theta_i]_+, \qquad a_i = 2r_i, \qquad \frac{d\mu}{dt} = \sum_i r_i - \frac{1}{2}. \tag{10}$$

与自旋一样，键振荡器通过共享的参考耦合相互竞争。在平衡态，$a_i = [s_i - 2\mu]_+$ 与概率之和为一。这正是 sparsemax，其中 $\tau = 2\mu$。

2.4 更多实现方式

连续伊辛自旋和 Kuramoto 振荡器是同一原理的两种实现方式。该原理并不局限于这两种设计。公式（3）中的一对耦合常微分方程规定了计算基底如何计算注意力。每台伊辛机或模拟平台都可以通过自身的基本构件、读出方式和缩放方式来实现它。也可以调整动力学，使其适应自身的物理特性。

3 评估

在理想硬件上，两个系统给出的预测结果与 sparsemax 的数字实现相同。然而，模拟基底会受到精度有限、噪声和器件失配等非理想因素的影响 [51]。因此，我们研究量化和读出噪声对推理质量的影响。

3.1 设置

语言模型。 我们从头开始，在 FineWeb-Edu [52] 上训练一个包含 51.5M 个参数的 GPT [3] 。它有 8 层，每层 8 个注意力头，宽度为 512，使用 GPT-2 词表，上下文长度为 1024 个词元。使用 AdamW [53] 在 0.8B 个词元上训练。我们报告模型在 512 个留出窗口（每个窗口含 1024 个词元）上的困惑度，以及通过 LM Evaluation Harness [55] 测得的零样本 LAMBADA 准确率 [54] 。

视觉 Transformer。 我们从在 ImageNet-1k [56, 57] 上预训练的 DeiT-S/16 [4, 5]，开始，并在 ImageNet-100 [58] 的 100 类子集上对其进行 15 个 epoch 的微调。在前 3 个 epoch 中，α-entmax [7] 从 $\alpha = 1$ 逐渐增至 2，随后使用 sparsemax。我们报告模型在 5,000 张验证图像上的 Top-1 准确率。

有界查询。为方便起见，查询自旋被限制在 $[-1, 1]$ 范围内，而查询振荡器对的 cos φ_j 也处于相同范围（第 2.2 节和第 2.3 节）。因此，我们对每个注意力头的查询投影应用 tanh，并将此模型称为有界稀疏最大化。所有评估的模型都在每个注意力头中使用有界稀疏最大化。

非理想硬件。 在推理过程中，耦合动力系统计算每个层中每个注意力头的注意力概率。随后，使用这些概率以数字方式对值向量进行平均。我们对每次弛豫进行 $T = 20t_0$，其中 t_0 是自旋或振荡器的时间常数。我们

表 1：每个注意力头中采用耦合动力系统时的推理质量。第一行表示理想硬件，其结果等同于精确的 s parsemax。其他各行对 q 和 K 进行量化、添加读出噪声，或同时进行这两种处理。在存在读出噪声时，各项数值为 20 次噪声采样的均值 ± 标准差。

		ImageNet-100 前 1 名（%）↑		FineWeb-Edu 困惑度↓		LAMBADA 准确率（%）↑	
精度	读出噪声	伊辛	Kuramoto	伊辛	Kuramoto	伊辛	Kuramoto
Full	None	91.02	91.02	35.60	35.60	13.02	13.02
Full	2%	90.77 ±0.12	90.77 ±0.12	35.90 ±0.01	35.90 ±0.01	12.94 ±0.13	12.94 ±0.13
Full	5%	90.31 ±0.23	90.31 ±0.23	37.78 ±0.02	37.78 ±0.02	11.97 ±0.21	11.97 ±0.21
5 位	None	90.74	90.86	35.87	35.87	12.89	12.87
5 位	2%	90.63 ±0.16	90.65 ±0.18	36.19 ±0.01	36.19 ±0.01	12.65 ±0.12	12.64 ±0.11
5 位	5%	90.22 ±0.19	90.24 ±0.20	38.14 ±0.02	38.14 ±0.02	11.77 ±0.22	11.78 ±0.22
4 位	None	90.54	90.50	37.03	37.03	12.25	12.25
4 位	2%	90.50 ±0.19	90.46 ±0.18	37.41 ±0.01	37.41 ±0.01	12.10 ±0.16	12.10 ±0.16
4 位	5%	89.95 ±0.19	89.97 ±0.21	39.64 ±0.03	39.64 ±0.03	11.35 ±0.22	11.34 ±0.22

使用自适应 Heun–Euler 方法。为模拟非理想硬件，我们对编程到系统中的查询 q 和键 K 施加扰动。二者均量化为固定比特数，并根据其完整范围添加相应水平的高斯读出噪声。对于不同的非理想情况，我们不会重新训练模型。在存在读出噪声时，我们报告 20 次噪声抽样的均值和标准差。

3.2 非理想条件下的推理质量

表 1 展示了结果。在使用理想硬件时，两个系统在每张图像和每个 LAMBADA 示例上都给出了精确 sparsemax 的预测结果。我们未采用量化感知训练或噪声感知训练。随着精度降低、噪声增加，质量会平稳下降。在 5 位精度、噪声为 2% 时，平均 Top-1 准确率下降 0.4 个百分点，困惑度上升 1.7%。最差的情况是 4 位精度、噪声为 5%。此时 Top-1 准确率下降 1.1 个百分点，LAMBADA 得分下降 1.7 个百分点，困惑度上升 11%。在所有情况下，伊辛自旋和 Kuramoto 振荡器的性能相近。

4 相关工作

伊辛机和振荡器网络。 伊辛机将问题编码为耦合和偏置，并让物理系统弛豫至低能态 [9, 10]。它们已通过超导量子比特 [12], 、光学参量振荡器 [13–16, 59, 60], 、概率比特 [61, 62],、CMOS 电路 [24–26, 63, 64], 、约瑟夫森参量振荡器 [65, 66] 以及耦合电子振荡器 [17–21, 50, 67] 构建而成。基于振荡器的机器遵循 Kuramoto 动力学，其计算特性已得到深入研究 [49, 68, 69]。伊辛机可求解最大割、可满足性、MIMO 检测及其他优化问题 [14, 17, 22, 25, 27–39, 66, 70–73]。一些研究通过引入高阶相互作用 [22, 30, 31, 74] 或为每个自旋设置一个误差变量 [75, 76]，扩展了简单的伊辛动力学。除了优化之外，伊辛机及相关动力系统还能够学习、求解微分方程并生成图像 [40–45, 77–79]。我们通过改变伊辛动力学本身，同时保持其对硬件友好的特性，拓展了这一研究方向。这使伊辛机能够应用于 Transformer 注意力等现代机器学习任务。

注意力。注意力最初被引入翻译模型，使每个输出词都能与相关的输入词对齐 [80]。Transformer 基于缩放点积注意力，并使用 softmax 对每个查询的得分进行归一化 [1]。由于其开销会随序列长度呈二次增长，许多研究致力于降低注意力的计算成本。固定稀疏模式会预先限制哪些查询-键对需要计算得分 [81–83]。线性注意力以特征映射替代 softmax 核函数，实现线性计算开销 [84, 85]。FlashAttention 保留精确的 softmax 注意力，并重新组织其在 GPU 上的内存访问 [86]。逐元素的 ReLU 或 sigmoid 注意力则舍弃了和为一的约束 [87–89]。Sparsemax 和 entmax 保留该约束，并将得分较低的键的概率精确置为零 [6–8]。有理论证明，随着输入规模增大，softmax 注意力会变得分散，而稀疏归一化器可以避免这一问题 [90, 91]。我们的设计将查询-键分数与归一化融合为一次物理弛豫，其平衡态恰好就是 sparsemax 注意力。

将注意力视为动力系统或能量最小化。Hopfield网络通过降低能量来检索模式 [92]。在现代霍普菲尔德网络中，对该能量执行一步更新就等同于 softmax 注意力 [93], , 而稀疏变体则对应 s parsemax 和 entmax 注意力 [94, 95]。能量 Transformer 用一个循环块取代多个块的堆叠，使其沿词元对一种设计能量进行下降 [96]。优化视角将自注意力视为沿词元能量或编码率能量执行的一步下降 [97, 98]。Sinkformers 将 softmax 替换为 Sinkhorn 迭代，使注意力矩阵成为双随机矩阵 [99]。其他研究将跨层的词元建模为相互作用粒子，它们会随时间聚类 [100–102]。优化层输出二次规划的解 [103], , 经典神经网络则通过拉格朗日乘子动力学精确施加约束 [104, 105]。AKOrN 用 Kuramoto 振荡器取代阈值单元 [106]。Kuramoto Attention 和 KoPE 为每个词元赋予相位，这些相位通过经 softmax 归一化的耦合实现同步 [107, 108]。我们也将注意力计算为动力系统的平衡态。我们的动力学直接作用于注意力概率，并映射到伊辛机原语。

5 结论

我们提出了一类耦合动力系统，将伊辛动力学扩展到了 Transformer 注意力。基于 sparsemax 二次规划，我们推导出一对耦合常微分方程，其平衡态恰好是 sparsemax 注意力。这两个常微分方程都可以逐项映射到伊辛机原语，分别对应连续伊辛自旋和 Kuramoto 振荡器。这一映射为在伊辛机及类似模拟基底上构建注意力提供了一套方法。我们在语言模型和视觉 Transformer 中对这两个系统进行了评估。随着精度降低和读出噪声增大，它们的推理质量会平稳下降，这为更大规模的研究铺平了道路。

将这些系统扩展到完整的 Transformer，会在架构、电路和算法层面带来问题。在架构层面，注意力必须与投影层和前馈层集成，系统也必须扩展到长上下文。在电路层面，约束节点需要一种新电路，用于感测所有自旋，并将一个共享偏置反馈给每个自旋。在算法层面，键值缓存、注意力头以及模型训练都可以与硬件协同设计。解决这些问题，便能从动力系统迈向用于高效 Transformer 推理的物理加速器。

参考文献

[1] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is all you need。发表于神经信息处理系统进展（*NeurIPS*），第 30 卷，第 5998–6008 页，2017 年。[2] Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。通过生成式预训练改进语言理解。技术报告，OpenAI，2018 年。[3] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督多任务学习器。技术报告，OpenAI，2019 年。[4] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像抵得上 16×16 个词：面向大规模图像识别的 Transformer。发表于国际学习表征会议（*ICLR*），2021 年。[5] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Hervé Jégou。通过注意力实现训练数据高效的图像Transformer与蒸馏。发表于第 *38* 届国际机器学习会议（*ICML*）论文集，《机器学习研究论文集》第 139 卷，第 10347–10357 页。PMLR，2021 年。[6] André F. T. Martins 和 Ramón Fernandez Astudillo。从 softmax 到 sparsemax：一种稀疏注意力模型与多标签分类方法。发表于 第 *33* 届国际机器学习会议（*ICML*）论文集，《机器学习研究论文集》第 48 卷，第 1614–1623 页。PMLR，2016 年。[7] Ben Peters、Vlad Niculae 和 André F. T. Martins。稀疏序列到序列模型。发表于第 *57* 届计算语言学协会年会（*ACL*）论文集，第 1504–1519 页，2019 年。doi：10.18653/v1/P19-1146。[8] Gonçalo M. Correia、Vlad Niculae 和 André F. T. Martins。自适应稀疏 Transformer。发表于 *2019* 年自然语言处理实证方法会议暨第 *9* 届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 2174–2184 页，2019 年。doi：10.18653/v1/D19-1223。[9] Naeimeh Mohseni、Peter L. McMahon 和 Tim Byrnes。作为组合优化问题硬件求解器的伊辛机。*Nature Reviews Physics*，4(6)：363–379，2022 年。doi：10.1038/s42254-022-00440-8。[10] Tingting Zhang、Qichao Tao、Bailiang Liu、Andrea Grimaldi、Eleonora Raimondo、Manuel Jiménez、María José Avedillo、Juan Nuñez、Bernabé Linares-Barranco、Teresa Serrano-Gotarredona、Giovanni Finocchio 和 Jie Han。对采用传统技术和新兴技术实现的伊辛机的综述。*IEEE* 纳米技术汇刊，23：704–717，2024 年。doi：10.1109/TNANO.2024.3457533。[11] Andrew Lucas。许多 NP问题的伊辛表述。*Frontiers in Physics*，2：5，2014 年。[12] M. W. Johnson、M. H. S. Amin、S. Gildert、T. Lanting、F. Hamze、N. Dickson、R. Harris、A. J. Berkley、J. Johansson、P. Bunyk、E. M. Chapple、C. Enderud、J. P. Hilton、K. Karimi、E. Ladizinsky、N. Ladizinsky、T. Oh、I. Perminov、C. Rich、M. C. Thom、E. Tolkacheva、C. J. S. Truncik、S. Uchaikin、J. Wang、B. Wilson 和 G. Rose。利用人工自旋进行量子退火。*Nature*，473(7346)：194–198，2011 年。doi：10.1038/nature10012。[13] Takahiro Inagaki、Yoshitaka Haribara、Koji Igarashi、Tomohiro Sonobe、Shuhei Tamate、Toshi- mori Honjo、Alireza Marandi、Peter L. McMahon、Takeshi Umeki、Koji Enbutsu、Osamu Tadanaga、Hirokazu Takenouchi、Kazuyuki Aihara、Ken-ichi Kawarabayashi、Kyo Inoue、Shoko Utsunomiya 和 Hiroki Takesue。一种用于 2000 节点优化问题的相干伊辛机。*Science*，354(6312)：603–606，2016 年。[14] Peter L. McMahon、Alireza Marandi、Yoshitaka Haribara、Ryan Hamerly、Carsten Langrock、Shuhei Tamate、Takahiro Inagaki、Hiroki Takesue、Shoko Utsunomiya、Kazuyuki Aihara、Robert L. Byer、M. M. Fejer、Hideo Mabuchi 和 Yoshihisa Yamamoto。一种具备全连接的完全可编程 100 自旋相干伊辛机。*Science*，354(6312)：614–617，2016 年。

[15] Yoshihisa Yamamoto、Kazuyuki Aihara、Timothée Leleu、Ken-ichi Kawarabayashi、Satoshi Kako、Martin Fejer、Kyo Inoue 和 Hiroki Takesue。相干伊辛机：在量子极限下运行的光神经网络。*npj QuantumInformation*，3(1):49，2017。doi: 10.1038/s41534-017-0048-9。[16] Toshimori Honjo、Tomohiro Sonobe、Kensuke Inaba、Takahiro Inagaki、Takuya Ikuta、Ya- suhiro Yamada、风间拓志、Koji Enbutsu、Takeshi Umeki、Ryoichi Kasahara、Ken-ichi Kawarabayashi 和 Hiroki Takesue。十万自旋相干伊辛机。*Science Advances*，7(40):eabh0952，2021。[17] Tianshi Wang 和 Jaijeet Roychowdhury。OIM：用于求解组合优化问题的振荡器伊辛机。载于非常规计算与自然计算（*UCNC*），计算机科学讲义第11493卷，第232–256页。施普林格，2019。doi: 10.1007/978-3-030-19311-9 19。_[18] Jeffrey Chou、Suraj Bramhavar、Siddhartha Ghosh 和 William Herzog。模拟耦合振荡器加权伊辛机。*ScientificReports*，9(1):14786，2019。[19] William Moy、Ibrahim Ahmed、Po-wei Chiu、John Moy、Sachin S. Sapatnekar 和 Chris H. Kim。用于求解组合优化问题的1,968节点耦合环形振荡器电路。*Nature Electronics*，5(5):310–317，2022。doi: 10.1038/s41928-022-00749-3。[20] Jaykumar Vaidya、R. S. Surya Kanthi 和 Nikhil Shukla。无需外部注入锁定的基于电子振荡器的伊辛机。*ScientificReports*，12(1):981，2022。[21] Hüsrev Cılasun、William Moy、Ziqing Zeng、Tahmida Islam、Hao Lo、Alex Vanasse、Megan Tan、Mohammad Anees、Ramprasath S、Abhimanyu Kumar、Sachin S. Sapatnekar、Chris H. Kim 和 Ulya R. Karpuzcu。用于组合优化的基于耦合振荡器的伊辛芯片。*NatureElectronics*，8(6):537–546，2025。doi: 10.1038/s41928-025-01393-3。[22] Lianlong Sun、Matthew X. Burns 和 Michael C. Huang。具有相位-振幅动力学和多项式相互作用的通用基于振荡器的伊辛机模型。*PhysicalReviewApplied*，24(4):044035，2025。doi: 10.1103/7s4v-4hs4。[23] 山冈正直、吉村千寻、林正人、奥山拓也、青木英孝和水野博之。采用 CMOS 退火求解组合优化问题的两万自旋伊辛芯片。*IEEE* 固态电路期刊，51(1):303–309，2016。[24] Richard Afoakwa、张一桥、Uday Kumar Reddy Vengalam、Zeljko Ignjatovic 和 Michael Huang。BRIM：双稳态电阻耦合伊辛机。载于*IEEE* 高性能计算机体系结构国际研讨会（*HPCA*），第749–760页，2021。doi: 10.1109/HPCA51647.2021.00068。[25] 苏宇奇、Tony Tae-Hyoung Kim 和 Bongjin Kim。用于通过直接映射求解布尔可满足性的三体自旋相互作用可重构 CMOS 伊辛机。*IEEE* 固态电路快报，6:221–224，2023。[26] Yongchao Liu、Lianlong Sun、Michael Huang 和 Hui Wu。基于朗之万动力学的伊辛机集成硬件退火。载于欧洲设计、自动化与测试会议（*DATE*），第1–6页，2025。doi: 10.23919/DATE64628.2025.10993070。[27] Yoshitaka Haribara、Shoko Utsunomiya 和 Yoshihisa Yamamoto。用于最大割问题的相干伊辛机：与半定规划和模拟退火的性能评估。载于Yoshihisa Yamamoto和千波耕一（编），量子信息技术原理与方法，《物理学讲义》，第251–262页。日本施普林格，东京，2016。doi: 10.1007/978-4-431-55756-2_12。[28] Mohammad Khairul Bashar、Antik Mallick、Daniel S. Truesdell、Benton H. Calhoun、Siddhartha Joshi 和 Nikhil Shukla。通过实验演示可重构耦合振荡器平台以求解最大割问题。*IEEE* 探索性固态计算器件与电路期刊，6(2):116–121，2020。

[29] Mohammad Khairul Bashar 和 Nikhil Shukla。设计具有高阶自旋相互作用的伊辛机及其在求解组合优化问题中的应用。*Scientific Reports*，13(1):9558，2023。[30] Connor Bybee、Denis Kleyko、Dmitri E. Nikonov、Amir Khosrowshahi、Bruno A. Olshausen 和 Friedrich T. Sommer。利用高阶伊辛机实现高效优化。*NatureCommunications*，14(1):6033，2023。[31] Anshujit Sharma、Matthew Burns、Andrew Hahn 和 Michael Huang。增强电子伊辛机，以有效求解布尔可满足性问题。*Scientific Reports*，13(1): 22858，2023。[32] Abhishek Kumar Singh、Kyle Jamieson、Peter L. McMahon 和 Davide Venturelli。伊辛机的动力学与正则化：实现近最优 MIMO 检测。*IEEETransactionsonWirelessCommunications*，21(12):11080–11094，2022。[33] Shreesha Sreedhara、Jaijeet Roychowdhury、Joachim Wabnig 和 Pavan Koteshwar Srinath。使用振荡器伊辛机进行 MU-MIMO 检测。载于*IEEE/ACM*国际计算机辅助设计会议（*ICCAD*），第1–9页，2023。doi: 10.1109/ICCAD57390.2023. 10323680。[34] Kotaro Tanahashi、Shinichi Takayanagi、Tomomitsu Motohashi 和 Shu Tanaka。伊辛机的应用及其软件开发。*Journal of the PhysicalSociety of Japan*，88(6):061010，2019。[35] Matthieu Parizy、Przemyslaw Sadowski 和 Nozomu Togawa。伊辛机上的基数约束投资组合优化。载于*IEEE*第*35*届片上系统国际会议（*SOCC*），第1–6页，2022。doi: 10.1109/SOCC56010.2022.9908082。[36] Siya Bao、Masashi Tawada、Shu Tanaka 和 Nozomu Togawa。基于伊辛机的平衡取货车辆路径问题求解器。*IEEE Transactions on ConsumerElectronics*，70(1):445–459，2024。doi: 10.1109/TCE.2023.3335392。[37] Nasa Matsumoto、Yohei Hamakawa、Kosuke Tatsumura 和 Kazue Kudo。使用QUBO表述进行基于距离的聚类。*Scientific Reports*，12(1):2669，2022。[38] Zetian Mao、Yoshiki Matsuda、Ryo Tamura 和 Koji Tsuda。利用基于GPU的伊辛机进行化学设计。*Digital Discovery*，2(4):1098–1103，2023。[39] Kenta Kirihara、Hidetaka Imai、Eisuke Kuroda、Jun Yamazaki 和 Akiko Masaki-Kato。探索伊辛机在电力系统运行中的潜在应用。*IEEE Access*，11:68004–68017，2023。[40] Shaila Niazi、Shuvro Chowdhury、Navid Anjum Aadit、Masoud Mohseni、Yao Qin 和 Kerem Y. Camsari。使用稀疏伊辛机训练深度玻尔兹曼网络。*NatureElectronics*，7(7):610–619，2024。doi: 10.1038/s41928-024-01182-4。[41] Ruibing Song、Chunshu Wu、Chuan Liu、Ang Li、Michael Huang 和 Tony Tong Geng。DS-GL：借助可扩展动力系统中蕴含的自然力量，推进图学习。载于*ACM/IEEE*第*51*届年度国际计算机体系结构研讨会（*ISCA*），第45–57页，2024。doi: 10.1109/ISCA59077.2024.00014。[42] Chunshu Wu、Ruibing Song、Chuan Liu、Yunan Yang、Ang Li、Michael Huang 和 Tong Geng。将自然的力量从二值图学习拓展至现实世界中的实值图学习。载于国际学习表征会议（*ICLR*），第57041–57053页，2024。[43] Chuan Liu、Ruibing Song、Chunshu Wu、Pouya Haghi 和 Tong Geng。InstaTrain：在动力系统中通过超快速自然退火实现自适应训练。载于国际学习表征会议（*ICLR*），第42729–42741页，2025。[44] Chuan Liu、Chunshu Wu、Ruibing Song、Ang Li、Ying Nian Wu 和 Tong Geng。用于高效函数学习的高表达力、自适应动力系统。载于第*42*届国际机器学习会议（*ICML*）论文集，计算机科学研究论文集第267卷，机器学习研究论文集。PMLR，2025。

[45] Lianlong Sun、Chuan Liu、Yongchao Liu、Hui Wu、Tony Geng 和 Michael Huang。用于图像生成的条件动力系统。*arXiv* 预印本 *arXiv:2608.14961*，2026年。[46] Yu Liu、Bolin Zhang、Jialiang Yin、Wentao Huang、Zhenyu Guan、Weisheng Zhao、Deming Zhang 和 Lang Zeng。基于异步 SOT-P-Bits 的 Transformer 注意力节能型并行伊辛 SoftMax 引擎。《*IEEE* 电子器件快报》，47(7):1458–1461，2026年。doi: 10.1109/LED.2026.3691358。[47] Peng Du、Jinjing Shi、Wenxuan Wang、Yin Ma、Kai Wen 和 Xuelong Li。QAMA：用于深度学习的可扩展量子退火多头注意力算子。*arXiv* 预印本 *arXiv:2504.11083*，2025年。[48] Gilhan Kim 和 Daniel K. Park。玻尔兹曼注意力：用于协作式注意力的可学习伊辛耦合。*arXiv* 预印本 *arXiv:2606.12478*，2026年。[49] Yoshiki Kuramoto。耦合非线性振荡器群体的自同步。在 Huzihiro Araki 编，理论物理数学问题国际研讨会，第39卷，《物理学讲义》，第420–422页。施普林格，1975年。doi: 10.1007/BFb0013365。[50] Hao Lo、William Moy、Hanzhao Yu、Sachin Sapatnekar 和 Chris H. Kim。基于耦合环形振荡器、采用48节点全连接阵列架构的伊辛求解器芯片。*Nature Electronics*，6(10):771–778，2023年。doi: 10.1038/s41928-023-01021-y。[51] T. Patrick Xiao、Ben Feinberg、Christopher H. Bennett、Venkatraman Prabhakar、Prashant Saxena、Vineet Agrawal、Sapan Agarwal 和 Matthew J. Marinella。论模拟神经网络推理加速器的精度。《*IEEE* 电路与系统杂志》，22(4):26–48，2022年。doi: 10.1109/MCAS.2022.3214409。[52] Guilherme Penedo、Hynek Kydlíˇcek、Loubna Ben Allal、Anton Lozhkov、Margaret Mitchell、Colin Raffel、Leandro von Werra 和 Thomas Wolf。FineWeb 数据集：大规模提炼网络中的高质量文本数据。在 神经信息处理系统进展（*NeurIPS*），数据集与基准赛道，第37卷，2024年。[53] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。在 国际学习表征会议（*ICLR*），2019年。[54] Denis Paperno、Germán Kruszewski、Angeliki Lazaridou、Ngoc Quan Pham、Raffaella Bernardi、Sandro Pezzelle、Marco Baroni、Gemma Boleda 和 Raquel Fernández。LAMBADA 数据集：需要广泛语篇上下文的词语预测。在 第*54*届计算语言学协会年会（*ACL*）论文集，第1525–1534页，2016年。doi: 10.18653/v1/P16-1144。[55] Leo Gao、Jonathan Tow、Baber Abbasi、Stella Biderman、Sid Black、Anthony DiPofi、Charles Foster、Laurence Golding、Jeffrey Hsu、Alain Le Noac'h、Haonan Li、Kyle McDonell、Niklas Muennighoff、Chris Ociepa、Jason Phang、Laria Reynolds、Hailey Schoelkopf、Aviya Skowron、Lintang Sutawika、Eric Tang、Anish Thite、Ben Wang、Kevin Wang 和 Andy Zou。语言模型评测工具套件。Zenodo，2024年。[56] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大规模层次化图像数据库。在 *IEEE* 计算机视觉与模式识别会议（*CVPR*），第248–255页，2009年。doi: 10.1109/CVPR.2009.5206848。[57] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein、Alexander C. Berg 和 Li Fei-Fei。ImageNet 大规模视觉识别挑战赛。《国际计算机视觉杂志》，115(3):211–252，2015年。doi: 10.1007/s11263-015-0816-y。[58] Yonglong Tian、Dilip Krishnan 和 Phillip Isola。对比式多视图编码。在 欧洲计算机视觉会议（*ECCV*），第12356卷，计算机科学讲义，第776–794页，2020年。doi: 10.1007/978-3-030-58621-8_45。

[59] Takahiro Inagaki、Kensuke Inaba、Ryan Hamerly、Kyo Inoue、Yoshihisa Yamamoto 和 Hiroki Takesue。基于简并光学参量振荡器的大规模伊辛自旋网络。*Nature Photonics*，10(6):415–419，2016年。doi: 10.1038/nphoton.2016.68。[60] Ryan Hamerly、Takahiro Inagaki、Peter L. McMahon、Davide Venturelli、Alireza Marandi、Tatsuhiro Onodera、Edwin Ng、Carsten Langrock、Kensuke Inaba、Toshimori Honjo、Koji Enbutsu、Takeshi Umeki、Ryoichi Kasahara、Shoko Utsunomiya、Satoshi Kako、Ken-ichi Kawarabayashi、Robert L. Byer、Martin M. Fejer、Hideo Mabuchi、Dirk Englund、Eleanor Rieffel、Hiroki Takesue 和 Yoshihisa Yamamoto。相干伊辛机与量子退火机性能差异的实验研究。*Science Advances*，5(5):eaau0823，2019年。doi: 10.1126/sciadv.aau0823。[61] William A. Borders、Ahmed Z. Pervaiz、Shunsuke Fukami、Kerem Y. Camsari、Hideo Ohno 和 Supriyo Datta。利用随机磁性隧道结进行整数因式分解。*Nature*，573(7774):390–393，2019年。doi: 10.1038/s41586-019-1557-9。[62] Navid Anjum Aadit、Andrea Grimaldi、Mario Carpentieri、Luke Theogarajan、John M. Martinis、Giovanni Finocchio 和 Kerem Y. Camsari。利用稀疏伊辛机实现大规模并行概率计算。*NatureElectronics*，5(7):460–468，2022年。doi: 10.1038/s41928-022-00774-2。[63] Yimin Wang 和 Xuanyao Fong。利用电容耦合锁存器实现节能型伊辛机以求解最大割。在 *IEEE*国际电路与系统研讨会（*ISCAS*），第1–5页，2024年。doi: 10.1109/ISCAS58744.2024.10558647。[64] Yongchao Liu、Lianlong Sun、Matthew Burns、Michael Huang 和 Hui Wu。基于电荷再分布的伊辛机。在 *IEEE*国际电路与系统研讨会（*ISCAS*），第1–5页，2025年。doi: 10.1109/ISCAS56072.2025.11044217。[65] Sasan Razmkhah、Mehdi Kamal、Nobuyuki Yoshikawa 和 Massoud Pedram。基于约瑟夫森参量振荡器的伊辛机。*Physical ReviewB*，109(1):014511，2024年。doi: 10.1103/PhysRevB.109.014511。[66] Beyza Zeynep Ucpinar、Sasan Razmkhah、Mehdi Kamal 和 Massoud Pedram。用于组合优化问题的可扩展超导伊辛机。在 *IEEE*计算机学会*VLSI*年度研讨会（*ISVLSI*），第565–570页，2024年。doi: 10.1109/ISVLSI61997.2024.00107。[67] Dagur I. Albertsson 和 Ana Rusu。通过耦合强度的准周期调制实现高度可重构的基于振荡器的伊辛机。*Scientific Reports*，13(1):4005，2023年。doi: 10.1038/s41598-023-31155-0。[68] Mikhail Erementchouk、Aditya Shukla 和 Pinaki Mazumder。论基于非线性振荡器的伊辛机的计算能力。*Physica D: Nonlinear Phenomena*，437:133334，2022年。doi: 10.1016/j.physd.2022.133334。[69] E. M. H. E. B. Ekanayake 和 Nikhil Shukla。路径不同，目的地相同：设计具有设计稳定性的受物理启发的动力系统，以最小化伊辛哈密顿量。*PhysicalReviewApplied*，24(2):024008，2025年。doi: 10.1103/cdc9-y234。[70] Salwa Shaglel、Markus Kirsch、Marten Winkler、Christian Münch、Stefan Walter、Fritz Schinkel 和 Martin Kliesch。伊辛机求解最大割问题的全面基准测试。*NewJournalof Physics*，2026年。[71] Venkata Pavan Sumanth Sikhakollu、Shreesha Sreedhara、Rajit Manohar、Alan Mishchenko 和 Jaijeet Roychowdhury。面向振荡器伊辛机的高质量基于电路的3-SAT映射。在 非常规计算与自然计算（*UCNC*），《计算机科学讲义》，第269–285页。施普林格，2024年。[72] Abhishek Kumar Singh、Ari Kapelyan、Minsung Kim、Davide Venturelli、Peter L. McMahon 和 Kyle Jamieson。使用伊辛机进行上行链路MIMO检测：一种多阶段伊辛方法。*IEEE Transactions onWireless Communications*，23(11):17037–17053，2024年。

[73] Yui Tsuyumine、Kenichi Masuda、Takeshi Hachikawa、Tsuyoshi Haga、Yuta Yachi、Tatsuhiko Shirai、Masashi Tawada 和 Nozomu Togawa。利用伊辛机优化实际的时变车辆路径问题。载于*IEEE*国际消费电子会议 *(ICCE)*，第 1–5 页，2024。doi: 10.1109/ICCE59016.2024.10444436。[74] Robbe De Prins、Guy Van der Sande、Peter Bienstman 和 Thomas Van Vaerenbergh。如何在模拟伊辛机中纳入高阶相互作用。*arXiv*预印本 *arXiv:2507.23621*，2025年。[75] Timothée Leleu、Yoshihisa Yamamoto、Peter L. McMahon 和 Kazuyuki Aihara。通过校正振幅异质性，使模拟自旋系统中的局部极小值失稳。*PhysicalReviewLetters*，122(4):040607，2019。doi: 10.1103/PhysRevLett.122.040607。[76] Timothée Leleu、Farad Khoyratee、Timothée Levi、Ryan Hamerly、Takashi Kohno 和 Kazuyuki Aihara。混沌振幅控制在高性能组合优化中的规模优势。*CommunicationsPhysics*，4(1):266，2021。doi: 10.1038/ s42005-021-00768-0。[77] Jérémie Laydevant、Danijela Markovi´c 和 Julie Grollier。利用平衡传播训练伊辛机。*Nature Communications*，15(1):3671，2024。doi: 10.1038/ s41467-024-46879-4。[78] Qingshan Wang、Clara C. Wanjura 和 Florian Marquardt。利用平衡传播训练耦合相位振荡器，将其用作神经形态平台。*Neuromorphic Computing andEngineering*，4(3):034014，2024。doi: 10.1088/2634-4386/ad752b。[79] Chuan Liu、Chunshu Wu、Ruibing Song、Guangyan Sun、Ying Nian Wu、Yousu Chen、Ang Li 和 Tong Geng。DS-TIDE：利用动力系统高效求解时不变微分方程。载于第*58*届*IEEE/ACM*国际微体系结构研讨会论文集 *(MICRO)*，第 1690–1703 页，2025年。[80] Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。通过联合学习对齐与翻译实现神经机器翻译。载于国际学习表征会议 *(ICLR)*，2015。[81] Rewon Child 、 Scott Gray 、 Alec Radford 和 Ilya Sutskever。利用稀疏Transformer生成长序列。*arXiv*预印本 *arXiv:1904.10509*，2019。[82] Iz Beltagy、Matthew E. Peters 和 Arman Cohan。Longformer：面向长文档的 Transformer。*arXiv*预印本 *arXiv:2004.05150*，2020。[83] Manzil Zaheer、Guru Guruganesh、Kumar Avinava Dubey、Joshua Ainslie、Chris Alberti、Santiago Ontanon、Philip Pham、Anirudh Ravula、Qifan Wang、Li Yang 和 Amr Ahmed。Big Bird：面向更长序列的Transformer。载于神经信息处理系统进展 *(NeurIPS)*，第 33 卷，第 17283–17297 页，2020。[84] Angelos Katharopoulos、Apoorv Vyas、Nikolaos Pappas 和 François Fleuret。Transformer就是循环神经网络：具有线性注意力的快速自回归Transformer。载于第*37*届国际机器学习会议 *(ICML)*论文集，机器学习研究论文集第 119 卷，第 5156–5165 页。PMLR，2020。[85] Krzysztof Choromanski、Valerii Likhosherstov、David Dohan、Xingyou Song、Andreea Gane、Tamas Sarlos、Peter Hawkins、Jared Davis、Afroz Mohiuddin、Lukasz Kaiser、David Belanger、Lucy Colwell 和 Adrian Weller。借助 Performers 重新思考注意力机制。载于国际学习表征会议 *(ICLR)*，2021。[86] Tri Dao、Daniel Y. Fu、Stefano Ermon、Atri Rudra 和 Christopher Ré。FlashAttention：具备 I/O 感知能力的快速、内存高效精确注意力。载于神经信息处理系统进展 *(NeurIPS)*，第 35 卷，第 16344–16359 页，2022。[87] Biao Zhang、Ivan Titov 和 Rico Sennrich。利用线性单元实现稀疏注意力。载于*2021*年自然语言处理实证方法会议 *(EMNLP)*论文集，第 6507–6520 页。计算语言学协会，2021。doi: 10.18653/v1/2021. emnl

p-main.523.

[88] Mitchell Wortsman、Jaehoon Lee、Justin Gilmer 和 Simon Kornblith。用 ReLU 替代视觉 Transformer 中的 softmax。*arXiv* 预印本 *arXiv:2309.08586*，2023年。[89] Jason Ramapuram、Federico Danieli、Eeshan Dhekane、Floris Weers、Dan Busbridge、Pierre Ablin、Tatiana Likhomanenko、Jagrit Digani、Zijin Gu、Amitis Shidani 和 Russ Webb。sigmoid自注意力的理论、分析与最佳实践。载于国际学习表征会议（*ICLR*），2025年。[90] Petar Veli˘ckovi´c、Christos Perivolaropoulos、Federico Barbero 和 Razvan Pascanu。仅有 softmax 还不够（对于尖锐规模泛化）。载于第*42*届国际机器学习会议（*ICML*）论文集，第267卷，机器学习研究论文集，第61190–61211页。PMLR，2025年。[91] Pavlo Vasylenko、Hugo Pitorro、André F. T. Martins 和 Marcos Treviso。利用稀疏注意力实现长上下文泛化。载于国际学习表征会议（*ICLR*），2026年。[92] J. J. Hopfield。具有分级响应的神经元拥有类似双态神经元的集体计算特性。美国国家科学院院刊，81(10)：3088– 3092，1984年。doi：10.1073/pnas.81.10.3088。[93] Hubert Ramsauer、Bernhard Schäfl、Johannes Lehner、Philipp Seidl、Michael Widrich、Lukas Gruber、Markus Holzleitner、Thomas Adler、David Kreil、Michael K. Kopp、Günter Klambauer、Johannes Brandstetter 和 Sepp Hochreiter。你所需要的就是 Hopfield网络。载于国际学习表征会议（*ICLR*），2021年。[94] Jerry Yao-Chieh Hu、Donglin Yang、Dennis Wu、Chenwei Xu、Bo-Yu Chen 和 Han Liu。关于稀疏现代Hopfield模型。载于神经信息处理系统进展（*NeurIPS*），第36卷，2023年。[95] Saul Santos、Vlad Niculae、Daniel McNamee 和 André F. T. Martins。稀疏结构化Hopfield网络。载于第*41*届国际机器学习会议（*ICML*）论文集，第235卷，机器学习研究论文集，第43368–43388页。PMLR，2024年。[96] Benjamin Hoover、Yuchen Liang、Bao Pham、Rameswar Panda、Hendrik Strobelt、Duen Horng Chau、Mohammed Zaki 和 Dmitry Krotov。能量 Transformer。载于神经信息处理系统进展（*NeurIPS*），第36卷，2023年。[97] Yongyi Yang、Zengfeng Huang 和 David P. Wipf。从优化视角看 Transformer。载于神经信息处理系统进展（*NeurIPS*），第35卷，第36958–36971页，2022年。doi：10.52202/068431-2678。[98] Yaodong Yu、Sam Buchanan、Druv Pai、Tianzhe Chu、Ziyang Wu、Shengbang Tong、Ben- jamin D. Haeffele 和 Yi Ma。通过稀疏率缩减实现白盒Transformer。载于神经信息处理系统进展（*NeurIPS*），第36卷，第9422–9457页，2023年。doi：10.52202/075280-0413。[99] Michael E. Sander、Pierre Ablin、Mathieu Blondel 和 Gabriel Peyré。Sinkformers：具有双随机注意力的 Transformer。载于第*25*届国际人工智能与统计会议（*AISTATS*）论文集，第151卷，机器学习研究论文集，第3515–3530页。PMLR，2022年。

[100] Yiping Lu、Zhuohan Li、Di He、Zhiqing Sun、Bin Dong、Tao Qin、Liwei Wang 和 Tie-Yan Liu。从多粒子动力系统的视角理解并改进 Transformer。*arXiv* 预印本 *arXiv:1906.02762*，2019。

[101] Borjan Geshkovski、Cyril Letrouit、Yury Polyanskiy 和 Philippe Rigollet。自注意力动力学中簇的涌现。载于 神经信息处理系统进展（*NeurIPS*），第 36 卷，第 57026–57037 页，2023年。doi：10.52202/075280-2493。

[102] Borjan Geshkovski、Cyril Letrouit、Yury Polyanskiy 和 Philippe Rigollet。《Transformer 的数学视角》。美国数学会通报，62(3):427–479，2025年。doi: 10.1090/bull/1863.

[103] Brandon Amos 和 J. Zico Kolter。OptNet：作为神经网络一层的可微优化。载于第 *34* 届国际机器学习会议（*ICML*）论文集，第 70 卷机器学习研究论文集，第 136–145 页。PMLR，2017。[104] John C. Platt 和 Alan H. Barr。约束微分优化。载于神经信息处理系统进展（*NIPS*），第 612–621 页，1987年。[105] S. Zhang 和 A. G. Constantinides。拉格朗日规划神经网络。*IEEE* 电路与系统 *II*：模拟与数字信号处理汇刊，39(7):441–452，1992年。doi: 10.1109/82.160169。[106] Takeru Miyato、Sindy Löwe、Andreas Geiger 和 Max Welling。人工库拉莫托振荡神经元。载于国际学习表征会议（*ICLR*），2025年。[107] Joshua Nunley。Kuramoto Attention：在环面上同步自注意力。*arXiv* 预印本 *arXiv:2606.11585*，2026年。[108] Mingqing Xiao、Yansen Wang、Dongqi Han、Caihua Shan 和 Dongsheng Li。库拉莫托振荡相位编码：通过神经启发式同步提高学习效率。载于第 *43* 届国际机器学习会议（*ICML*）论文集，第 306 卷机器学习研究论文集，第 139020–139043 页。PMLR，2026年。

A 实现

代码清单 1 在一个示例中复现了两个系统。函数 `ising` 是公式 (6)。函数`kuramoto` 是采用公式 (8) 的设置和公式 (10) 的读出的公式 (7)。

代码清单 1：两个耦合动力系统的最小复现。

```
import numpy as np
from scipy.integrate import solve_ivp

q = np.array([0.6, 0.8])
K = np.array([[1., 1.], [1., 0.5], [-1., 1.]])
V = np.array([[1., 0.], [0., 1.], [-1., -1.]])
n, d = K.shape

s = K @ q / np.sqrt(d)
z = np.sort(s)[::-1]
cutoffs = (np.cumsum(z) - 1) / np.arange(1, n + 1)
a = np.maximum(s - cutoffs[z > cutoffs][-1], 0)

def ising(t, state):
    x, mu = np.clip(state[:-1], -1, 1), state[-1]
    J = 2 * K / np.sqrt(d)
    dx = J @ q - x - 1 - 2 * mu
    dx[((x <= -1) & (dx < 0)) | ((x >= 1) & (dx > 0))] = 0
    return np.r_[dx, np.sum((x + 1) / 2) - 1]

def kuramoto(t, state):
    theta, mu = state[:-1], state[-1]
    phi = np.arccos(q)
    J = K / (4 * np.sqrt(d))
    th = theta[:, None]
    pairs = (J * (np.sin(th - phi) + np.sin(th + phi))).sum(1)
    dtheta = -pairs + mu * np.sin(theta) + 0.5 * np.sin(2 * theta)
    return np.r_[dtheta, np.sum(np.maximum(np.cos(theta), 0)) - 0.5]

print(f"{'Sparsemax':9}: {a} -> {a @ V}")
for name, ode, x0, readout in [
    ("Ising", ising,  np.full(n, 2 / n - 1),
     lambda x: (np.clip(x, -1, 1) + 1) / 2),
    ("Kuramoto", kuramoto, np.full(n, np.pi / 2),
     lambda theta: 2 * np.maximum(np.cos(theta), 0)),
]:
    sol = solve_ivp(ode, (0, 200),  np.r_[x0, 0.],  rtol=1e-9,  atol=1e-11)
    b = readout(sol.y[:-1, -1])
    print(f"{name:9}: {b} -> {b @ V}")
    assert sol.success and np.allclose(a, b, atol=1e-6, rtol=0)
```