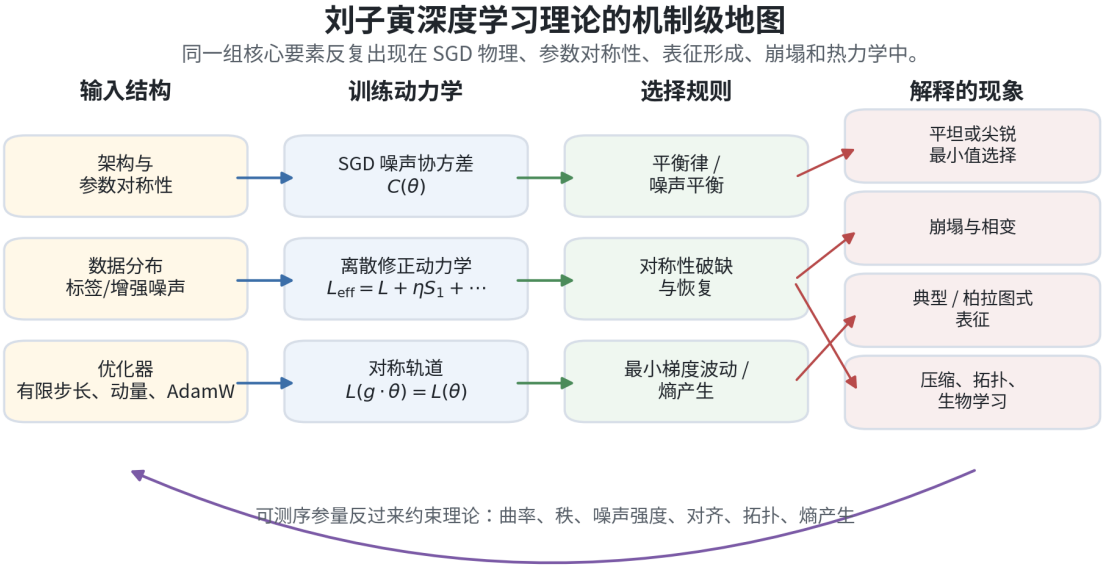
刘子寅的深度学习理论

对称性、随机梯度下降、表征与神经热力学

中文技术报告：机制化组织、主要结果与开放问题



对象： 用户

撰写： OpenAI GPT-5.5 Pro

版本： 中文版，基于改进版报告，2026-07-08

图表： 本报告所有图均为重新绘制的原创解释性示意图

摘要

刘子寅关于深度学习理论的一系列工作，可以被组织成一个统一的研究纲领：现代神经网络不是普通的有限维优化问题，而是具有大量参数冗余、对称性和退化流形的随机动力系统；真实训练也不是无穷小步长的梯度流，而是有限步长、带有小批量噪声、归一化、正则化和优化器偏置的非平衡过程。其主要结论包括：有限学习率会改变 SGD 的平稳涨落；小批量噪声是各向异性且依赖状态的；SGD 逃逸可由对数势垒而非普通损失势垒控制；参数对称性会诱导约束、平衡律、Noether 型漂移和拓扑保持/破坏；表征学习表现为隐藏表征、权重和神经元梯度之间的典范对齐；坍缩、特征学习和泛化相变可以在精确可解模型中出现；神经热力学进一步把训练理解为带有熵力和不可逆性的物理过程。本文按机制而非年份重组这些结果，给出核心方程、图示和主要剩余开放问题。部分 2025–2026 年工作仍为预印本或新近会议版本，题名、版本和细节可能继续变化。

目录

1 执行摘要 **1**

1.1 主要结果一页版 2

2 文献范围、来源状态与技术分类 **3**

2.1 范围 3

2.2 按技术路线的时间轴 3

2.3 论文地图 3

3 符号与技术背景 **5**

4 核心结果 I：有限步长 SGD 是离散随机物理过程 **6**

4.1 离散协方差替代扩散协方差 6

4.2 小批量噪声是结构化的 7

4.3 逃逸由对数势垒控制 8

4.4 大常数学习率可违背梯度流直觉 8

4.5 平坦性与梯度涨落 8

5 核心结果 II：参数对称性是隐藏结构 **9**

5.1 对称诱导约束 9

5.2 噪声平衡与 Noether 流 9

5.3 平衡律 10
5.4 等变性、拓扑和压缩 10

6 核心结果 III：表征通过对齐形成 11
6.1 典范表征假说 11
6.2 柏拉图式表征与多模态对齐 11
6.3 组合泛化需要的不只是瓶颈解缠 11

7 核心结果 IV：可解模型揭示坍缩、逐步学习与相变 12
7.1 周期函数与归纳偏置 12
7.2 深线性网络与原点的特殊性 12
7.3 后验坍缩 12
7.4 相变与逐步学习 13
7.5 自监督损失景观 13

8 核心结果 V：神经热力学与不可逆性 14
8.1 有效损失与熵力 14
8.2 训练算法的热力学不可逆性 14
8.3 热力学综合 15

9 核心结果 VI：理论驱动的算法干预 15
9.1 Snake 激活 15
9.2 LaProp 与自适应优化 15
9.3 spred 15
9.4 syre 15

10 核心结果 VII：生物学习可被视为梯度机器 15

11 综合：一个紧凑的数学图景 16

12 主要剩余开放问题 17

13 推荐阅读路径 **19**

13.1 最快概念路线 19

13.2 SGD 物理路线 20

13.3 表征路线 20

13.4 可解模型路线 20

14 批判性评估 **21**

14.1 最强之处 21

14.2 仍然具有推测性的部分 21

14.3 底线 21

来源与图表说明 **22**

参考文献 **23**

插图

1 机制级地图。输入结构包括架构、数据和优化器；训练动力学包括小批量噪声、有限步长修正和对称轨道；选择规则决定最终可观测的平坦/尖锐选择、坍缩、表征、压缩和生物学习现象。 1

2 阅读路线图。该图把相关工作组织成四条可读路径：SGD 物理、对称性、表征/自监督学习、热力学/压缩/生物学习。 3

3 有限学习率 SGD 的离散统计。左图显示同一二次盆地中，连续近似与离散精确协方差给出的稳态云团不同；右图显示当归一化步长靠近稳定边界 $\eta\lambda_{\max} = 2$ 时，离散方差相对连续近似急剧放大。 7

4 平坦性与梯度涨落的概念分离。普通损失在退化解集上几乎相同，曲率最平坦点与梯度涨落最小点可以不同；随机训练的有效偏置可沿对称坐标选择后者。 8

5 对称类型与结构偏置的对应。反射、重缩放、旋转和置换对称分别提示固定子空间、平衡/稀疏、低秩和同质集成/坍缩等诊断方向。 9

6 噪声平衡的示意图。普通损失在轨道上退化；有限步长随机训练诱导有效势，从而把参数推向涨落较小或熵产生较低的代表元。 10

7 表征形成的对齐观点。左图把隐藏表征、权重和神经元梯度视为三个需要同时对齐的几何对象；右图显示典型训练过程中对齐指标可以逐步上升。 11

8 相变与逐步学习示意。左图展示参数控制下从核/懒惰区到特征学习区再到坍缩或拓扑破坏区的相图；右图展示自监督学习中嵌入维度或特征模态可按阶段逐一获得。 13

9 不可逆性视角。正向训练路径通常不能简单反演；后向误差、时间反演不对称和熵产生等度量在小步长展开中具有共同首阶结构。 14

10 开放问题地图。机制、测量、规模化和控制/验证四类问题互相依赖：没有测量就难以验证机制，没有规模化就难以服务真实模型，没有控制算法就难以证明理论的实用性。 . . 17

1 执行摘要

中心论点

最简洁的总论点是：深度学习应被建模为退化参数流形上的随机离散动力学。经验风险只给出一组函数等价或近似等价的好解；有限步长 SGD、梯度噪声、权重衰减、归一化和参数对称性共同决定训练最终在该等价类中选择哪一个代表元、哪一个相或哪一条商空间轨迹。

刘子寅公开发表和列出的深度学习理论工作覆盖 SGD 噪声、有限步长效应、精确可解模型、对称性、表征形成、自监督学习坍缩、神经热力学、压缩和生物可实现学习等方向 [1, 2]。这些论文看似分散，但可以被压缩为如下公式：

$$\boxed{\text{参数对称性} + \text{有限步长随机动力学} + \text{正则化/归一化} \quad \Longrightarrow \quad \text{隐式结构选择.}} \tag{1}$$

所谓“结构选择”可以是平衡参数化、低秩性、稀疏性、表征对齐、表示坍缩、彩票票据式压缩、拓扑简化，或者生物电路中可实现的局部学习规则。

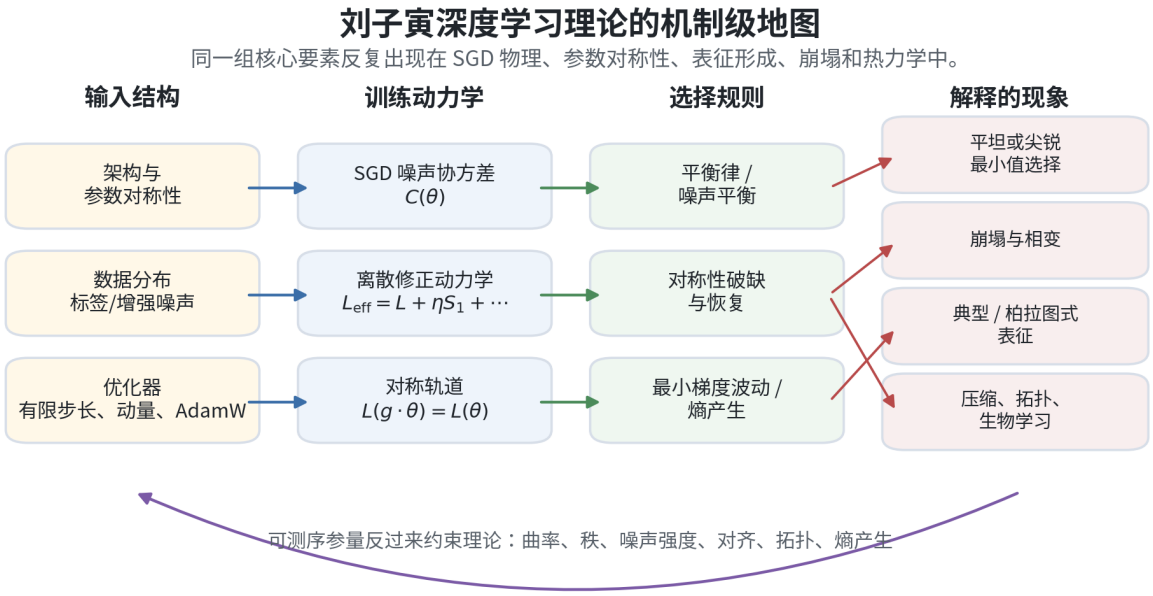


图 1: 机制级地图。输入结构包括架构、数据和优化器；训练动力学包括小批量噪声、有限步长修正和对称轨道；选择规则决定最终可观测的平坦/尖锐选择、坍缩、表征、压缩和生物学习现象。

1.1 主要结果一页版

结果族 1：有限步长 SGD 具有自己的平稳统计物理

在二次盆地中，离散 SGD 的精确平稳协方差满足

$$\Sigma = (I - \eta H)\Sigma(I - \eta H)^\top + \eta^2 C, \tag{2}$$

而不仅是连续扩散近似 $H\Sigma + \Sigma H \approx \eta C$。只有当所有相关方向都满足 $\eta\lambda_i(H) \ll 1$ 时，二者才接近。刘子寅关于有限学习率和小批量噪声的工作说明，现代训练的许多现象正发生在这个小步长近似失效的区域 [6, 8]。

结果族 2：平坦性不是更基本的隐式偏置

2026 年精确可解模型的核心观点是：SGD 没有普适的“偏好平坦最小值”的内置目标。更基本的选择量是梯度涨落。当噪声几何与 Hessian 曲率对齐时，看起来像是偏好平坦解；当标签噪声或数据噪声各向异性时，SGD 也可能选择尖锐但低涨落的解 [36]。

结果族 3：参数对称性把优化问题变成代表元选择问题

若 $L(g\cdot\theta) = L(\theta)$，则普通损失不能区分同一对称轨道 $G\cdot\theta$ 上的点。但权重衰减、噪声协方差、有限步长修正和熵项可以区分这些点，从而诱导稀疏、低秩、平衡、同质集成或坍缩等结构 [19, 20, 18]。

结果族 4：表征形成是对齐与对称破缺的动力学结果

典范表征假说认为，训练过程中隐藏表征 H、权重 W 与神经元梯度 G 会趋向某种互相对齐；当精确对齐不成立时，多项式对齐假说预测这些量之间存在幂律关系。这把表征形成从“静态架构性质”转化为可测的动力学几何问题 [23]。

结果族 5：热力学修正把训练视为非平衡不可逆过程

神经热力学把有限步长随机训练解释为带有有效熵力的过程：某些连续重参数化对称性被打破，离散或正交对称性可被保留；训练算法的不可逆性可用后向误差、时间反演不对称和熵产生等量刻画 [25, 37]。

结果族 6：坍缩、特征学习和相变不是例外，而是可解析的相

在自监督学习、线性 VAE、深线性网络和有限宽度可解模型中，坍缩可以是稳定相，特征学习可以通过连续或不连续相变出现，表征维度也可能逐步获得 [14, 12, 13, 15, 22]。

2 文献范围、来源状态与技术分类

2.1 范围

本文讨论的核心语料是 Liu Ziyin / 刘子寅大约 2019–2026 年间公开列出的深度学习理论相关工作，主要依据其 MIT 主页和论文列表 [1, 2]。纳入标准是：论文直接涉及深度学习理论、优化动力学、SGD 噪声、对称性、表征、自监督坍缩、相变、神经热力学、压缩或生物可实现学习。像 *LaProp*、*Snake*、*spred* 和 *syre* 这样的算法论文也被纳入，因为它们是由理论机制直接驱动的干预方法。金融类论文和纯应用多模态论文一般不纳入，除非它们直接服务于表征对齐问题。

文献状态说明

2025–2026 年若干项目仍是预印本或新近会议论文，正式题名、版本、证明细节或发表信息可能继续更新。本文的目标是解释主要思想和机制，而不是给出最终历史书目。

2.2 按技术路线的时间轴

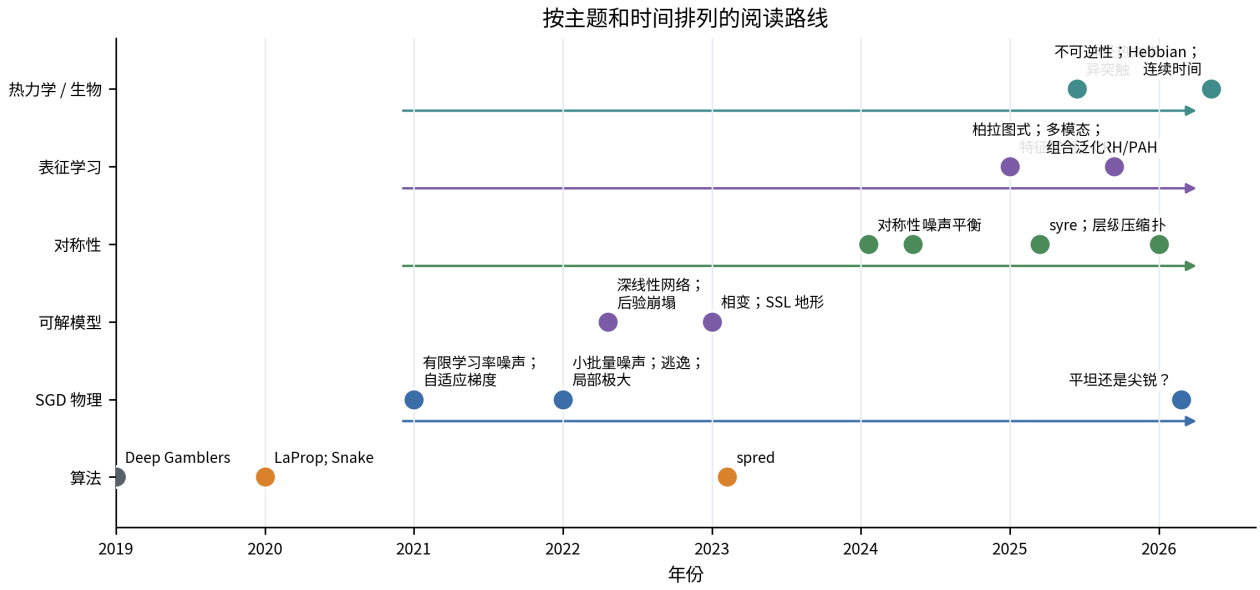


图 2: 阅读路线图。该图把相关工作组织成四条可读路径：SGD 物理、对称性、表征/自监督学习、热力学/压缩/生物学习。

2.3 论文地图

表 1: 刘子寅主要深度学习理论论文的机制化地图。

年份	论文或论文组	数学对象	主要贡献
2019	*Deep Gamblers* [3]	选择性分类损失；投资组合类比	设计允许模型在不确定时弃权的损失函数。它不是后续理论主线的中心，但展示了从其他数学领域引入结构的研究风格。
2020	*LaProp* [4]	自适应优化器更新规则	将动量与自适应缩放分离，避免 Adam 类优化器中二者耦合导致的不稳定。
2020	*Neural Networks Fail to Learn Periodic Functions* [5]	激活函数的归纳偏置	说明 ReLU、tanh、sigmoid 等标准激活难以外推周期函数；提出 Snake 激活 $x + \sin^2 x$。
2021	*Noise and Fluctuation of Finite Learning Rate SGD* [6]	极小值附近的离散随机递推	证明有限学习率改变平稳分布、动量行为和逃逸规律，不能由连续扩散近似完全替代。
2021	*Distributional Properties of Adaptive Gradients* [7]	自适应梯度更新统计量	分析自适应梯度分布性质，并质疑对 Adam warmup 的一种常见解释。
2022	*Strength of Minibatch Noise in SGD* [8]	单样本梯度协方差	推导小批量噪声强度公式，并联系稳定性、学习率/批量大小缩放和隐式正则化。
2022	*SGD with a Constant Large Learning Rate Can Converge to Local Maxima* [9]	离散随机动力学	构造例子说明大常数学习率 SGD 可收敛到局部最大值、缓慢逃离鞍点或偏好尖锐极小值。
2022	*Logarithmic Landscape and Power-Law Escape Rate of SGD* [10]	逃逸率理论	用对数景观势垒和幂律逃逸率替代普通损失势垒直觉。
2022–2023	*Exact Solutions of a Deep Linear Network* [11]	带正则/随机性的深线性网络	求解全局极小值，揭示深参数化如何在原点产生坏极小值和异常退化。
2022	*Posterior Collapse of a Linear Latent Variable Model* [12]	线性 VAE 型潜变量模型	给出后验坍缩的精确机制：似然项与先验诱导正则化之间的竞争。
2023	*Zeroth, First, and Second-Order Phase Transitions* [13]	统计力学序参量	展示深度学习训练状态可发生零阶、一阶或二阶相变。
2023	*What Shapes the Loss Landscape of SSL?* [14]	联合嵌入自监督损失	刻画完全坍缩、维度坍缩以及归一化和偏置的作用。
2023	*On the Stepwise Nature of SSL* [15]	对比核的特征模态	说明自监督学习可逐维、逐模态地获得嵌入维度。
2023	*spred* [16]	冗余重参数化	用可微重参数化和 L_2 型衰减求解 L_1 稀疏学习问题。
2023	*Type-II Saddles* [17]	鞍点附近的随机矩阵乘积	说明某些鞍点由于噪声消失或与方向对齐，可对 SGD 表现出概率稳定性。

年份	论文或论文组	数学对象	主要贡献
2023–2025	*Law of Balance and Stationary Distribution of SGD* [18]	重缩放对称乘积结构	研究平衡律、平稳测度、遍历破缺和涨落反转等对称丰富模型中的现象。
2024	*Symmetry Induces Structure and Constraint of Learning* [19]	群作用与固定子空间	证明对称性结合权重衰减或噪声可诱导稀疏、低秩或同质集成。
2024	*Parameter Symmetry and Noise Equilibrium of SGD* [20]	对称轨道上的 Noether 型漂移	说明梯度噪声可沿对称方向产生流，直到达到噪声平衡。
2025	*Remove Symmetries to Control Model Expressivity* [21]	对称诱导低容量状态	提出 syre，用模型无关方式破坏限制表达能力的对称性。
2025	*When Does Feature Learning Happen?* [22]	有限宽度可解析模型	指出 alignment、disalignment 和 rescaling 是核极限中缺失的特征学习机制。
2025	*Formation of Representations in Neural Networks* [23]	R、W 和 G 的对齐	提出典范表征假说和多项式对齐假说。
2025	对称破缺/恢复统一框架 [24]	对称层级结构	将层级学习和表征形成解释为参数对称性的破缺与恢复。
2025–2026	*Neural Thermodynamics* 与不可逆性 [25, 37]	熵力与熵产生	将有限步长随机训练重述为具有有效力的非平衡不可逆过程。
2025–2026	柏拉图表征、多模态对齐和组合泛化 [26, 27, 28]	表征等价与因子化	研究不同架构、模态和任务中的表征何时对齐，以及仅有解缠表征为何不足。
2025–2026	拓扑、等变性、压缩 [33, 34, 35]	等变动力学；置换不变压缩	将对称分析推广到二阶几何、拓扑保持/破坏和彩票票据式压缩。
2024–2026	生物学习论文 [29, 30, 31, 32]	异突触回路与连续时间学习	论证梯度型学习可能由生物上可行的回路和时间信用分配机制实现。

3 符号与技术背景

令 $\theta \in \mathbb{R}^p$ 表示网络参数，$L(\theta)$ 表示经验损失。小批量 SGD 的基本形式为

$$\theta_{t+1} = \theta_t - \eta g_t, \qquad g_t = \nabla L(\theta_t) + \xi_t, \qquad \mathbb{E}[\xi_t \mid \theta_t] = 0. \tag{3}$$

条件噪声协方差定义为

$$C(\theta) = \mathbb{E}[\xi_t \xi_t^\top \mid \theta_t = \theta]. \tag{4}$$

在无放回或近似独立抽样的小批量情形中，常见有限总体近似为

$$C_B(\theta) \approx \left(\frac{1}{B} - \frac{1}{N}\right) C_{\text{single}}(\theta), \tag{5}$$

其中 B 为 batch size，N 为样本数，C_{single} 是单样本梯度协方差。这只是起点；关键在于 $C(\theta)$ 一般既非各向同性，也非与 θ 无关。

参数对称性是满足

$$L(g \cdot \theta) = L(\theta), \qquad \text{或更强地} \qquad f_{g\cdot\theta} = f_\theta \tag{6}$$

的变换 g。例子包括隐藏单元置换、齐次层的重缩放、潜变量旋转、符号翻转以及深线性网络中的规范自由度。对称性产生轨道

$$\mathcal{O}_\theta = \{g \cdot \theta : g \in G\}, \tag{7}$$

普通损失在该轨道上无法选择唯一代表元。刘子寅理论的核心问题可表述为：有限步长随机训练会在这些轨道上选择什么？

4 核心结果 I：有限步长 SGD 是离散随机物理过程

4.1 离散协方差替代扩散协方差

在非退化局部极小值附近，可令

$$L(\theta) = \frac{1}{2}\theta^\top H \theta. \tag{8}$$

若局部噪声协方差近似为常数 C，离散 SGD 的精确平稳协方差就是式 (2)。连续扩散近似给出

$$H\Sigma + \Sigma H \approx \eta C. \tag{9}$$

两者只在 ηH 的所有相关特征值都很小时一致。现代训练经常接近 edge of stability，因此离散修正不是高阶装饰，而是可改变平稳测度的主效应。

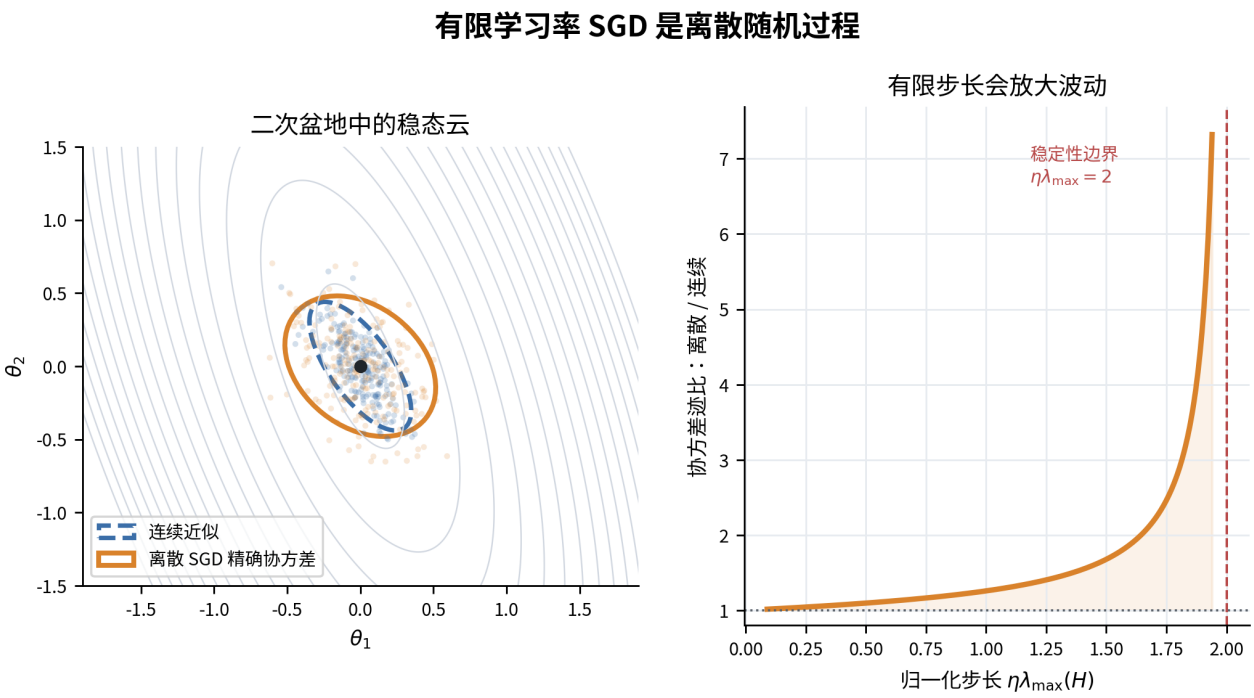


图 3: 有限学习率 SGD 的离散统计。左图显示同一二次盆地中，连续近似与离散精确协方差给出的稳态云团不同；右图显示当归一化步长靠近稳定边界 $\eta\lambda_{\max} = 2$ 时，离散方差相对连续近似急剧放大。

Noise and Fluctuation of Finite Learning Rate SGD 研究了上述差异及其在动量、小批量噪声、Adam 类方法和逃逸问题中的推广 [6]。其概念性结论是：算法本身改变平稳测度；仅知道损失和 Hessian 并不足以预测训练状态。

4.2 小批量噪声是结构化的

Strength of Minibatch Noise in SGD 推导了极小值附近小批量噪声强度公式，并把它连接到训练稳定性、学习率/批量大小缩放以及隐式正则化 [8]。关键点在于：梯度噪声不是简单的各向同性温度。它来自单样本梯度，因此反映数据几何和模型表征。任何把 SGD 简化为“loss 加白噪声”的理论都会丢掉重要信息。

一个合理诊断量是噪声与曲率的相对几何，例如

$$\frac{v^\top C(\theta) v}{v^\top H(\theta) v} \tag{10}$$

沿 Hessian 特征方向或数据相关方向的变化。这种量比单独的最大特征值或 trace 更接近刘子寅理论中“SGD 选择什么”的问题。

4.3 逃逸由对数势垒控制

Logarithmic Landscape and Power-Law Escape Rate of SGD 的思想是：SGD 从局部极小值逃逸时，支配概率的不一定是普通损失差 ΔL，而可能是对数景观中的势垒 [10]。这会导致幂律逃逸率，并使“平坦极小值偏好”与有效维度、噪声方向和 Hessian 结构相关，而不是一个简单的标量平坦性指标。

4.4 大常数学习率可违背梯度流直觉

SGD with a Constant Large Learning Rate Can Converge to Local Maxima 构造了反例，说明在标准梯度流直觉之外，SGD 可以收敛到局部最大值、极慢地逃离鞍点，甚至偏好尖锐极小值 [9]。这不是说常规训练总会如此，而是说明理论上不能把 SGD 简化为“带噪声的下降”。有限步长、采样机制与局部几何必须同时分析。

4.5 平坦性与梯度涨落

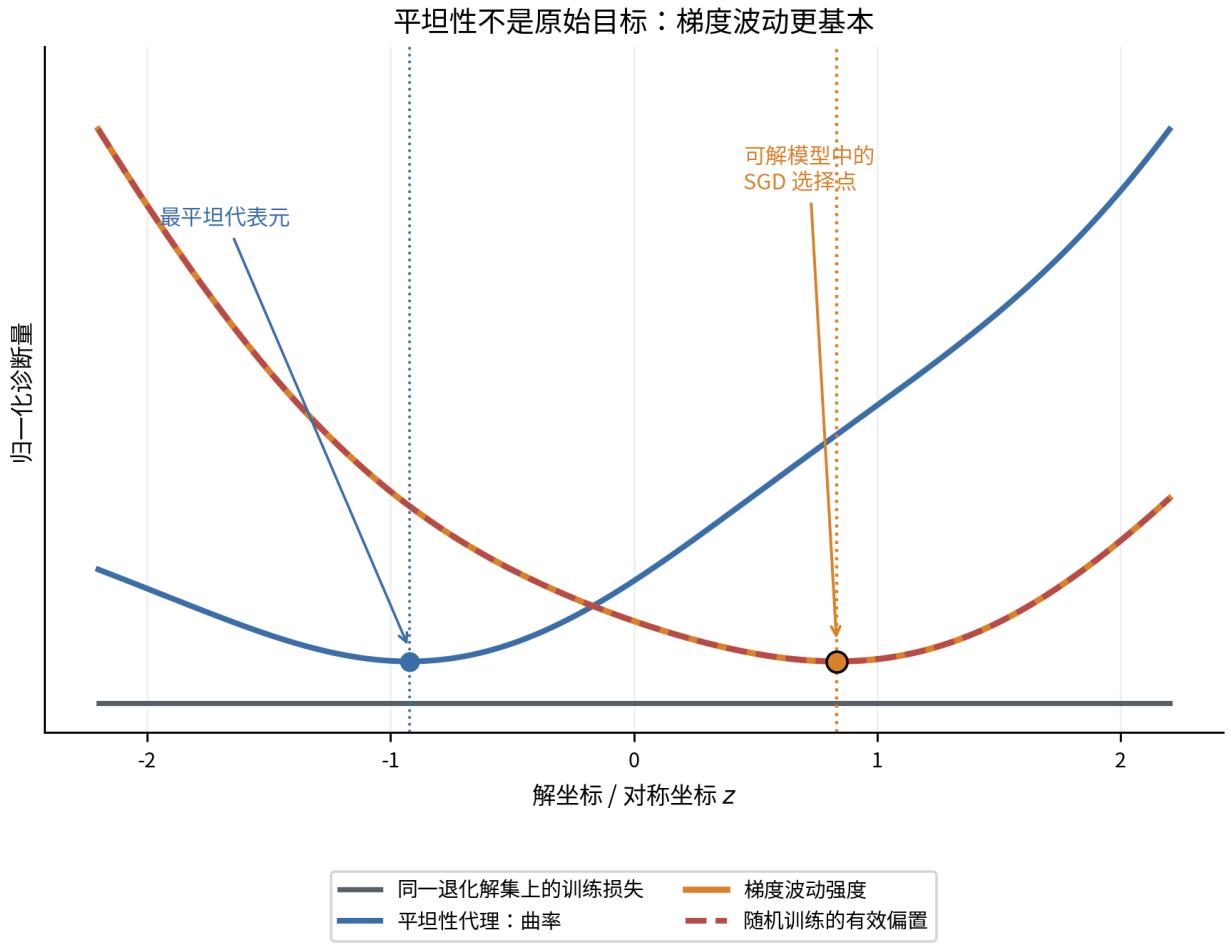


图 4: 平坦性与梯度涨落的概念分离。普通损失在退化解集上几乎相同，曲率最平坦点与梯度涨落最小点可以不同；随机训练的有效偏置可沿对称坐标选择后者。

Does SGD Seek Flatness or Sharpness? 给出一个精确可解模型来分离曲率平坦性与梯度涨落 [36]。其主张可以概括为：SGD 不是直接最小化 sharpness，而是倾向于最小化由数据和优化器共同定义的梯度涨落。若噪声近似各向同性，低涨落往往与平坦性重合；若标签噪声或数据几何各向异性，SGD 也可能选择尖锐解。

5 核心结果 II：参数对称性是隐藏结构

5.1 对称诱导约束

Symmetry Induces Structure and Constraint of Learning 的核心是：对称性不只是“多出一些等价参数”，而是会通过优化器和正则化产生可观测约束 [19]。若损失在某群作用下不变，则沿群轨道的普通梯度为零或退化；但权重衰减、梯度噪声和离散更新可在这些方向上产生选择力。

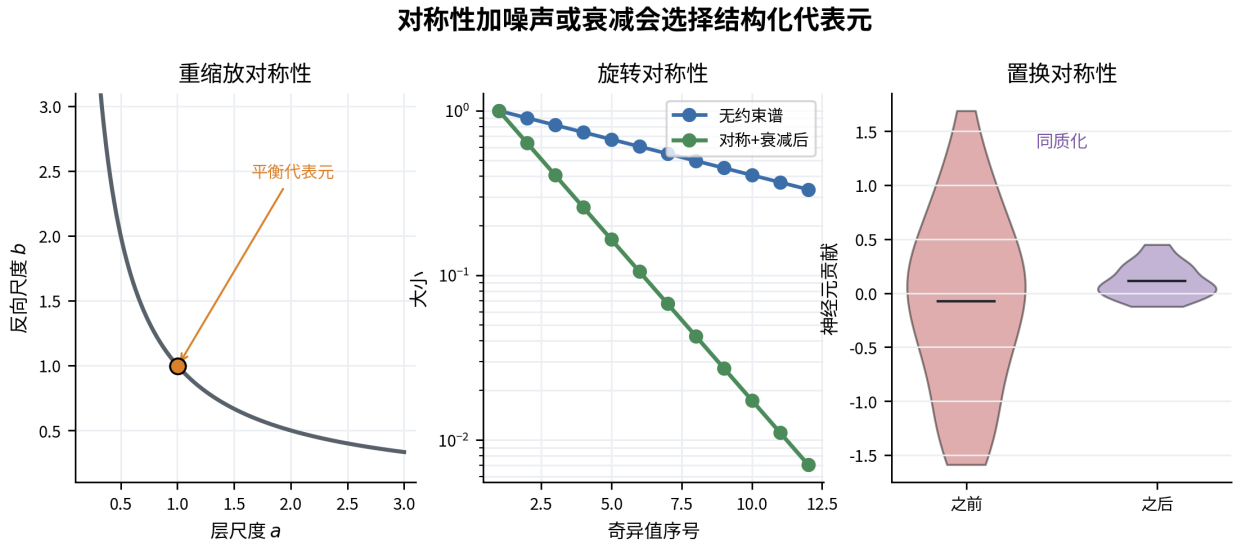


图 5: 对称类型与结构偏置的对应。反射、重缩放、旋转和置换对称分别提示固定子空间、平衡/稀疏、低秩和同质集成/坍缩等诊断方向。

可以把该结果理解为一条实践原则：先找出模型参数化中的群作用，再问训练算法在群轨道上是否有额外势函数。若有，则它将把函数等价类中的某些代表元选出来。反射对称可诱导固定子空间约束；齐次重缩放可诱导层间平衡；旋转自由度可诱导低秩结构；置换对称可诱导同质化或集成型解。

5.2 噪声平衡与 Noether 流

Parameter Symmetry and Noise Equilibrium of SGD 将上述思想推向动力学层面：沿对称轨道，普通损失不提供力，但噪声协方差和有限步长修正可产生 Noether 型漂移，使参数沿轨道移动直到达到噪声平衡 [20]。

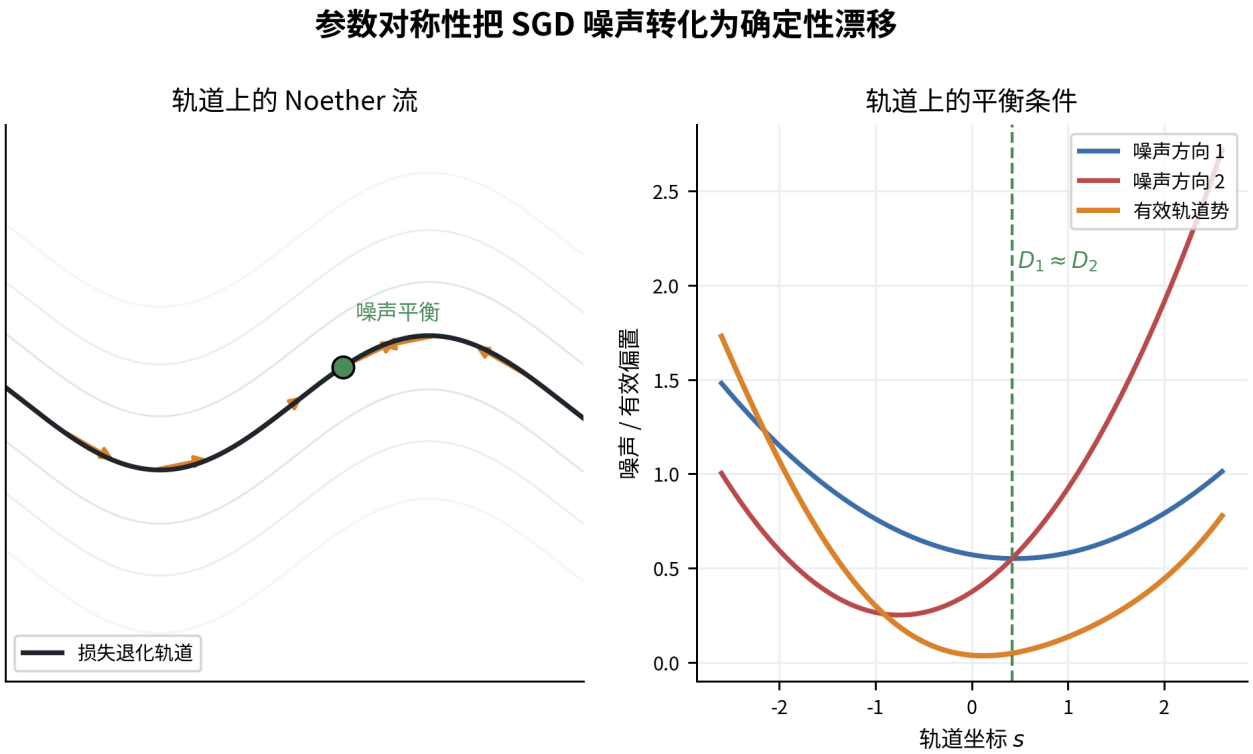


图 6: 噪声平衡的示意图。普通损失在轨道上退化；有限步长随机训练诱导有效势，从而把参数推向涨落较小或熵产生较低的代表元。

这给出了一种解释 progressive sharpening、flattening、warmup、normalization 以及表示形成的方法：这些现象不一定来自损失显式项，而可能来自对称轨道上的噪声诱导有效动力学。

5.3 平衡律

在具有重缩放对称性的模型中，例如两层乘积参数化 $w = uv$，所有满足 $uv = w$ 的 (u, v) 给出同一函数。梯度下降或 SGD 常常趋向所谓 balanced state，例如 $|u| \approx |v|$。*Law of Balance and Stationary Distribution of SGD* 研究了这种平衡律和平稳分布，并指出可出现遍历破缺、相变和涨落反转 [18]。这说明“平衡”不是简单的美学选择，而是对称性、噪声和正则化共同决定的统计态。

5.4 等变性、拓扑和压缩

等变学习规则满足“先变换再学习”和“先学习再变换”相容。*An Equivariance Toolbox for Learning Dynamics* 把这种相容性翻译成梯度、Hessian 和二阶动力学的约束 [34]。*Topological Invariance and Breakdown in Learning* 进一步说明，在置换等变学习规则下，小学习率可保持神经元分布拓扑，而超过临界学习率后拓扑可简化 [33]。压缩理论则把置换不变性与彩票票据式压缩相连，主张在一定光滑性与对称性条件下，许多对象级函数可被超多项式地压缩 [35]。

6 核心结果 III：表征通过对齐形成

6.1 典范表征假说

Formation of Representations in Neural Networks 提出两个相连假说 [23]。第一，典范表征假说：训练过程中，隐藏表征 R 或 H、权重 W 和神经元梯度 G 倾向于互相对齐，形成共享的低维坐标系。第二，多项式对齐假说：当严格对齐失败时，这些对象之间仍可能存在幂律或多项式关系。

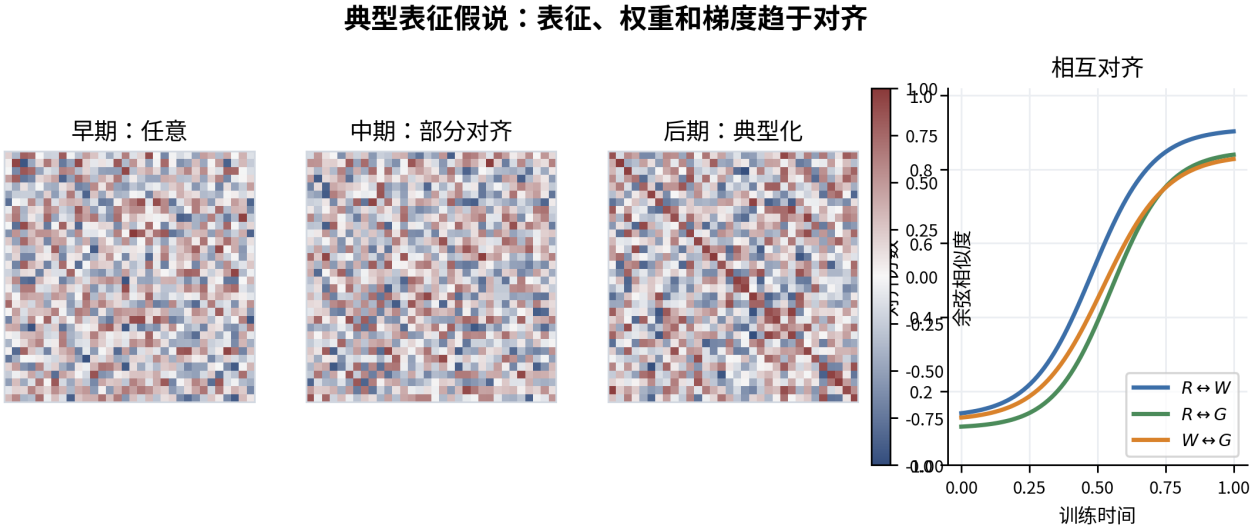


图 7: 表征形成的对齐观点。左图把隐藏表征、权重和神经元梯度视为三个需要同时对齐的几何对象；右图显示典型训练过程中对齐指标可以逐步上升。

这一理论的重要性在于，它把“表示学习发生了”变成一组可测量的量：表示矩阵的主方向、权重子空间、神经元梯度方向以及它们之间的夹角或谱相似性。它也把神经坍缩、neural feature ansatz 和 self-supervised representation alignment 放入同一几何语言中。

6.2 柏拉图式表征与多模态对齐

Proof of a Perfect Platonic Representation Hypothesis 在特定嵌入深线性网络中证明，不同宽度、深度和数据设置下训练出的表征可在旋转意义下匹配 [26]。多模态对齐论文则讨论不同模态独立训练后何时会隐式对齐，结论依赖于模态相似性以及冗余信息与独有信息的比例 [27]。这些工作共同提出一个问题：表征对齐是架构偶然性，还是训练动力学中更普遍的吸引子？

6.3 组合泛化需要的不只是瓶颈解缠

Compositional Generalization Requires More Than Disentangled Representations 说明，仅在瓶颈层拥有解缠坐标并不足以保证 OOD 组合泛化，因为后续层可以重新纠缠因素并通过 superposition 记忆训

练组合 [28]。稳健组合泛化需要在完整表示链或像素空间中保持因子化结构，而不只是局部瓶颈约束。

7 核心结果 IV：可解模型揭示坍缩、逐步学习与相变

7.1 周期函数与归纳偏置

Neural Networks Fail to Learn Periodic Functions and How to Fix It 证明和展示标准激活函数难以外推简单周期函数，并提出 Snake 激活

$$\phi(x) = x + \sin^2 x \tag{11}$$

以加入周期归纳偏置，同时保留较好的优化性质 [5]。这篇论文表明：看似简单的外推失败可以用激活函数的函数空间偏置来解释。

7.2 深线性网络与原点的特殊性

Exact Solutions of a Deep Linear Network 求解带权重衰减和随机神经元的深线性网络全局极小值 [11]。重要结论是：深参数化使原点成为特殊点；在多隐藏层情形中，权重衰减可在原点产生坏极小值。这与“线性模型总是简单”的直觉相反，因为深线性网络的参数空间有非平凡退化结构。

7.3 后验坍缩

Posterior Collapse of a Linear Latent Variable Model 在线性潜变量模型中刻画 VAE 后验坍缩的机制 [12]。坍缩来自似然项与先验诱导正则之间的竞争：当先验正则过强或数据方向不足以支撑潜变量使用时，编码器趋向忽略潜变量。该模型把“坍缩”从经验现象转化为可求解的相。

7.4 相变与逐步学习

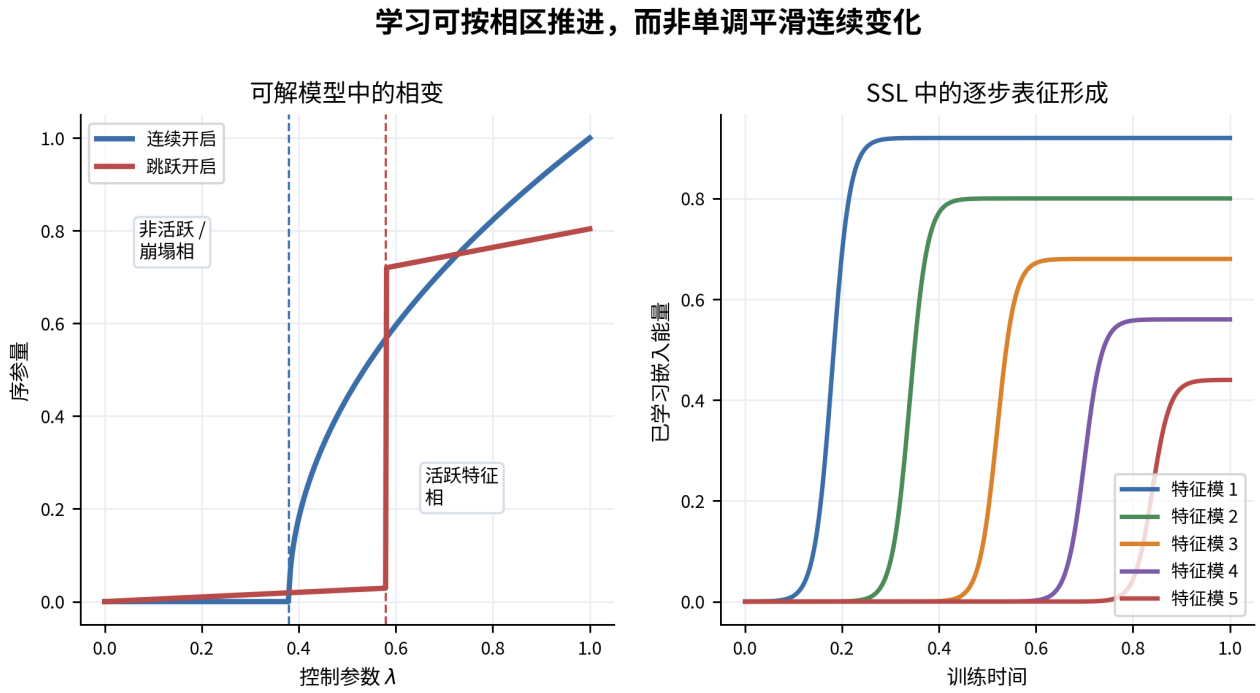


图 8: 相变与逐步学习示意。左图展示参数控制下从核/懒惰区到特征学习区再到坍缩或拓扑破坏区的相图；右图展示自监督学习中嵌入维度或特征模态可按阶段逐一获得。

Zeroth, First, and Second-Order Phase Transitions in Deep Neural Networks 使用统计物理序参量说明学习状态可发生不同阶数的相变 [13]。*On the Stepwise Nature of SSL* 则显示联合嵌入自监督方法可逐个学习对比核的特征模态 [15]。*When Does Feature Learning Happen?* 在有限宽度可解模型中识别 alignment、disalignment 和 rescaling 三种特征学习机制 [22]。这些结果共同说明：特征学习不是“连续变好”的单一路径，而可能是分段、阈值化和相变式的。

7.5 自监督损失景观

What Shapes the Loss Landscape of Self-Supervised Learning? 研究联合嵌入 SSL 目标，刻画完全坍缩、维度坍缩、归一化和偏置项的影响 [14]。它解释了为何某些 SSL 方法需要 stop-gradient、batch normalization、variance/covariance regularization 或显式防坍缩项：这些机制改变了坍缩相的稳定性。

8 核心结果 V：神经热力学与不可逆性

8.1 有效损失与熵力

神经热力学把有限步长随机训练近似写成普通损失加有效修正的形式，例如

$$L_{\text{eff}}(\theta) = L(\theta) + \eta S_1(\theta) + \eta^2 S_2(\theta) + \cdots , \tag{12}$$

其中 S_i 由噪声协方差、曲率、更新离散性和参数化决定。若普通损失在某对称轨道上常数，S_i 仍可沿轨道变化，从而产生熵力。这正是“损失不变但训练仍然选择某个代表元”的热力学解释 [25]。

8.2 训练算法的热力学不可逆性

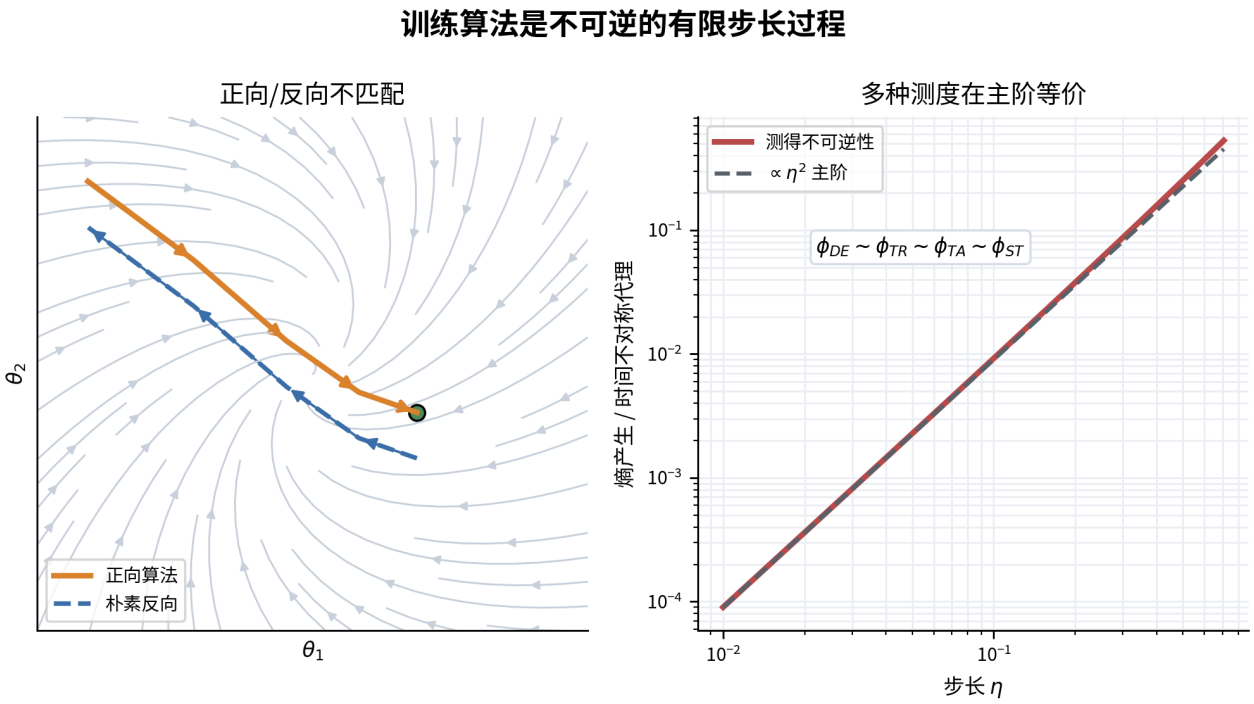


图 9: 不可逆性视角。正向训练路径通常不能简单反演；后向误差、时间反演不对称和熵产生等度量在小步长展开中具有共同首阶结构。

Thermodynamic Irreversibility of Training Algorithms 建立了刻画训练不可逆性的框架，说明多种定义在步长展开的首阶意义下等价：数值后向误差、时间重整化修正、微观时间反演不对称以及随机热力学熵产生 [37]。其含义是：训练不只是寻找静态损失极小值，而是一条带有方向性的非平衡轨迹。

8.3 热力学综合

把对称性和热力学结合起来，可得到一个统一解释：普通损失定义等价类，有限步长随机训练给出等价类上的有效势，熵产生或梯度涨落决定训练更倾向的代表元。该框架同时解释平衡律、表征对齐、平坦/尖锐选择、坍缩和拓扑破坏，但在真实 Transformer 规模上的定量版本仍是开放问题。

9 核心结果 VI：理论驱动的算法干预

9.1 Snake 激活

Snake 激活把周期结构直接写入激活函数，使网络在周期函数外推上具有更合适的归纳偏置 [5]。它体现了刘子寅工作中的一个算法原则：不是只扩大模型，而是把目标函数或函数类的几何结构注入参数化。

9.2 LaProp 与自适应优化

LaProp 发现 Adam 类优化器中动量与自适应缩放的耦合会产生不稳定，并通过解耦二者改进稳定性 [4]。这与后续 SGD 物理工作相连：优化器的更新规则本身定义了训练动力学，不能被视为无关实现细节。

9.3 spred

spred 使用冗余重参数化把 L_1 稀疏惩罚转化为可由 SGD 和 L_2 型衰减处理的问题 [16]。该方法说明，改变参数化可以改变优化器的隐式偏置，从而让标准训练过程解决原本非光滑的稀疏学习问题。

9.4 syre

Remove Symmetries to Control Model Expressivity 提出 syre，用以破坏导致低容量状态的对称性 [21]。它的理论动机是：若对称性把模型困在表达能力不足的折叠态，训练算法应主动打破这些对称性，而不是被动依赖随机初始化。

10 核心结果 VII：生物学习可被视为梯度机器

生物可实现学习论文把前述理论带入神经科学：是否存在局部电路能近似实现梯度更新？*Self-Assembly of a Biologically Plausible Learning Circuit* 提出可自组装的学习回路，并预测上行/下行皮层流之间的四突触 motif [29]。*Heterosynaptic Circuits Are Universal Gradient Machines* 进一步主张异突触可塑性可实现广泛的梯度型元学习 [30]。*Emergence of Hebbian Dynamics in Regularized Non-Local Learners* 把

SGD 加权重衰减或噪声注入与 Hebbian/anti-Hebbian 动力学联系起来 [31]。*On Biologically Plausible Learning in Continuous Time* 则研究连续时间模型，并强调输入信号与误差信号的时间重叠窗口 [32]。

这条线的意义在于：如果梯度计算可以由局部异突触机制间接实现，那么“深度学习中的梯度动力学”和“生物学习中的局部可塑性”之间的鸿沟可能比通常想象的小。

11 综合：一个紧凑的数学图景

可以用以下步骤概括刘子寅理论风格。

第一步，识别参数化中的对称群 G，写出等价轨道 $\mathcal{O}_\theta = G \cdot \theta$。第二步，写出训练更新和噪声协方差 $C(\theta)$。第三步，在小步长或可解模型中推导有效动力学：

$$\theta_{t+1} - \theta_t = -\eta \nabla L(\theta_t) + \eta \xi_t + \eta^2 b(\theta_t) + O(\eta^3), \tag{13}$$

其中 $b(\theta)$ 代表离散性、噪声诱导漂移或优化器修正。第四步，把 $b(\theta)$ 投影到对称轨道的切空间，得到代表元选择规则：

$$\Pi_{T\mathcal{O}_\theta} b(\theta) eq 0 \quad \Longrightarrow \quad \text{沿轨道发生 Noether 型流或熵力选择。} \tag{14}$$

第五步，用该代表元解释可观测现象：平衡、稀疏、低秩、坍缩、表征对齐、拓扑破坏或压缩。

该图景的优点是统一；缺点是目前许多定量版本仍主要在可解模型、小网络或受控实验中成立。真正困难的问题是把这些可测量量带到大规模 Transformer、AdamW、LayerNorm、数据增强和预训练管线中。

12 主要剩余开放问题

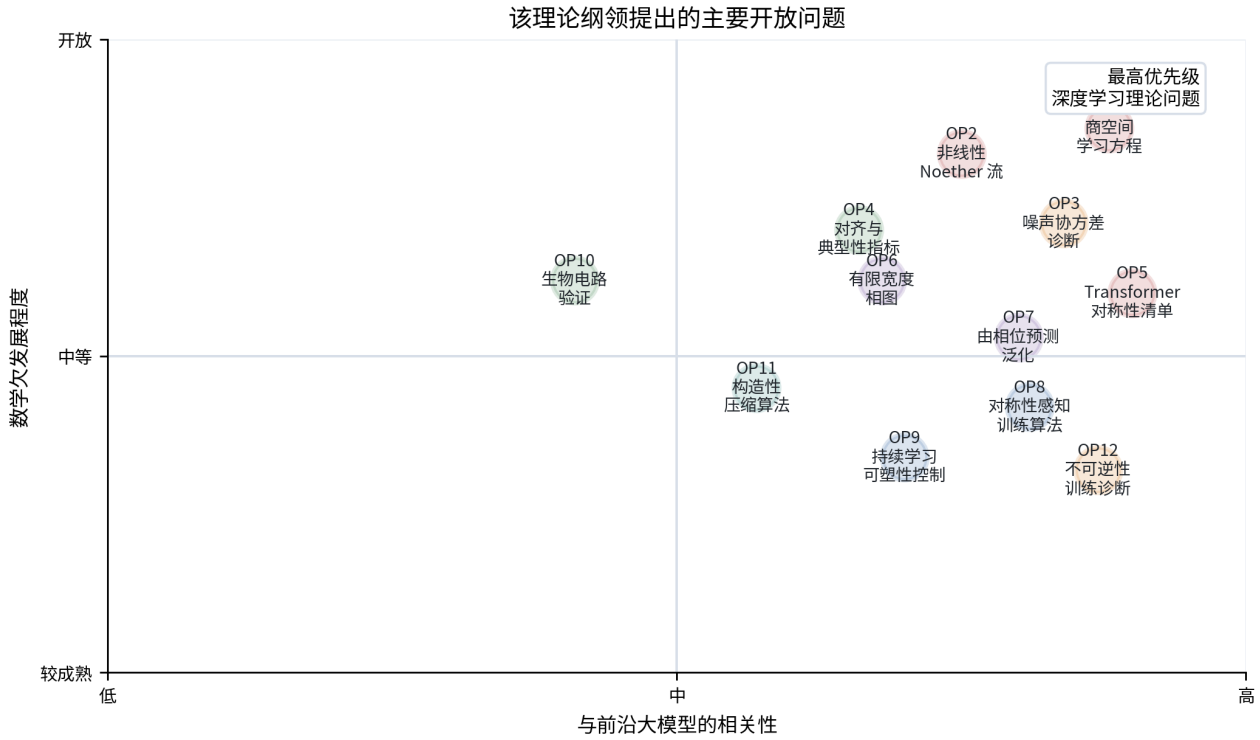


图 10: 开放问题地图。机制、测量、规模化和控制/验证四类问题互相依赖：没有测量就难以验证机制，没有规模化就难以服务真实模型，没有控制算法就难以证明理论的实用性。

OP1：商空间学习方程

问题。对具有群作用 G 的神经网络，推导训练在商空间 Θ/G 上的有效学习方程，而不是只在原始参数空间中写 SGD。

重要性。普通参数空间包含大量规范自由度；如果理论不能消去这些自由度，就会把不可观测代表元误认为函数学习本身。

进展标准。给出适用于非线性网络、归一化层和 AdamW 的商空间动力学，并能预测轨道上的代表元选择。

OP2：非线性 Noether 流定理

问题。把噪声诱导 Noether 流从可解或局部模型推广到一般非线性神经网络。

重要性。当前理论解释了若干对称轨道上的漂移，但还缺少覆盖真实架构中多重对称性、归一化和残差连接的通用定理。

进展标准。给出可检验的漂移公式，并在 MLP、CNN、Transformer 中准确预测平衡、sharpening 或 flattening 方向。

OP3：大模型噪声协方差诊断

问题。在大模型训练中高效估计 $C(\theta)$ 的关键谱、方向性和与 Hessian 的相对几何。
重要性。如果 SGD 的基本选择量是梯度涨落而非平坦性，理论必须提供可实际测量的噪声几何指标。
进展标准。开发低成本估计器，能在不显著增加训练成本的情况下预测稳定边界、逃逸方向、相变点和泛化差异。

OP4：对齐与典范性的可操作度量

问题。为 H、W、G 的典范对齐和多项式对齐建立标准化指标。
重要性。表征理论需要从定性描述变成可复现实验协议。没有统一指标，很难比较不同架构、数据集和优化器。
进展标准。指标应能区分神经坍缩、典范表征、柏拉图式跨模型对齐和失败的组合泛化，并能预测下游迁移。

OP5：Transformer 级对称性清单

问题。系统列出 Transformer 中的所有重要参数对称性和近似对称性，包括注意力头置换、MLP 宽度置换、LayerNorm 尺度自由度、残差流规范和 embedding 旋转。
重要性。刘子寅理论若要服务当前主模型，必须对 Transformer 的实际对称结构给出完整地图。
进展标准。一个能解释 head specialization、activation sparsity、rank collapse、attention redundancy 和 scaling behavior 的对称性分类。

OP6：有限宽度相图

问题。在有限宽度而非无限宽度或核极限中，绘制训练相图：哪些参数区域是核区域、特征学习区域、坍缩区域或拓扑破坏区域?
重要性。大模型既不是经典小模型，也不是严格无限宽度模型。有限宽度修正很可能决定真实表征学习。
进展标准。理论相图能预测 learning-rate/batch-size/width/depth 改变时的阶段转变，并与实验相变点匹配。

OP7：从被选择的相推导泛化

问题。不只是解释训练会选择哪个代表元，还要推导该代表元为什么泛化更好或更差。
重要性。对称性和热力学理论解释了选择机制，但泛化仍需要把代表元几何连接到数据分布、任务结构和测试误差。
进展标准。给出从噪声平衡、低秩、稀疏、对齐或坍缩指标到泛化误差的定量预测。

OP8：对称感知训练算法

问题。设计显式使用对称性和噪声几何的训练算法，而不仅是事后解释 SGD。
重要性。理论的实用价值取决于能否改进训练稳定性、样本效率、压缩或持续学习。
进展标准。算法能自动检测并调节对称轨道上的代表元选择，优于 AdamW、SAM、warmup、weight decay 和标准归一化基线。

OP9：持续学习中的可塑性控制

问题。用对称性和代表元选择解释 loss of plasticity，并设计可恢复可塑性的训练干预。
重要性。持续学习失败可能不是单纯的过拟合，而是模型被噪声、衰减和对称性推入低容量代表元。
进展标准。能预测何时发生可塑性丧失，并通过对称破缺、再平衡或噪声控制恢复新任务学习能力。

OP10：生物回路验证

问题。对异突触梯度机器和自组装学习电路给出可实验验证的神经科学预测。
重要性。生物学习理论目前很有启发性，但必须连接到真实皮层回路、突触可塑性时间窗和可观测 motif。
进展标准。实验证据支持四突触 motif、异突触梯度计算或秒级塑性窗口约束。

OP11：压缩定理到构造算法

问题。把通用压缩理论和动态彩票票据理论转化为真实网络和数据集上的实际压缩算法。
重要性。压缩定理有力，但依赖光滑性和置换不变性等假设；真正挑战是高效找到被压缩对象。
进展标准。方法不仅保持最终输出，还能保持训练动力学，并显著优于标准剪枝、蒸馏或数据选择基线。

OP12：不可逆性作为训练诊断

问题。在真实训练中测量熵产生或等价不可逆性，并用它预测学习质量。
重要性。热力学不可逆性提供统一概念，但只有变成可测信号后才具有工程价值。
进展标准。低成本不可逆性估计器能预测相变、表征形成、优化器不稳定或泛化跨任务变化。

13 推荐阅读路径

13.1 最快概念路线

按如下顺序阅读：

1. *Symmetry Induces Structure and Constraint of Learning* [19]。
2. *Parameter Symmetry and Noise Equilibrium of SGD* [20]。

3. *Formation of Representations in Neural Networks* [23]。
4. *Neural Thermodynamics* [25]。
5. *Thermodynamic Irreversibility of Training Algorithms* [37]。

这条路线最快地从对称性到热力学统一框架。

13.2 SGD 物理路线

阅读:

1. *Noise and Fluctuation of Finite Learning Rate SGD* [6]。
2. *Strength of Minibatch Noise in SGD* [8]。
3. *Logarithmic Landscape and Power-Law Escape Rate of SGD* [10]。
4. *Type-II Saddles and Probabilistic Stability of SGD* [17]。
5. *Does SGD Seek Flatness or Sharpness?* [36]。

这条路线解释为什么 SGD 必须作为有限步长随机过程分析。

13.3 表征路线

阅读:

1. *What Shapes the Loss Landscape of SSL?* [14]。
2. *On the Stepwise Nature of SSL* [15]。
3. *When Does Feature Learning Happen?* [22]。
4. *Formation of Representations in Neural Networks* [23]。
5. *Proof of a Perfect Platonic Representation Hypothesis* [26]。

这条路线聚焦隐藏层几何、特征形成和跨模型表征对齐。

13.4 可解模型路线

阅读:

1. *Neural Networks Fail to Learn Periodic Functions* [5]。
2. *Exact Solutions of a Deep Linear Network* [11]。
3. *Posterior Collapse of a Linear Latent Variable Model* [12]。

4. *Zeroth, First, and Second-Order Phase Transitions* [13]。
5. *When Does Feature Learning Happen?* [22]。

这条路线适合先通过玩具模型掌握机制，再读综合性论文。

14 批判性评估

14.1 最强之处

刘子寅理论最强的地方是把模糊经验命题替换为可分析的数学对象：

- “SGD 喜欢平坦极小值”变为梯度涨落几何问题。
- “过参数化有帮助”变为对称轨道和冗余参数化问题。
- “表征涌现”变为 H、W、G 之间的对齐问题。
- “坍缩发生了”变为目标函数和动力学中的稳定相。
- “训练是优化”变为非平衡热力学过程。

这些替换产生可证伪的诊断量和干预目标，因此具有科学价值。

14.2 仍然具有推测性的部分

最宽泛的统一主张很有吸引力，但尚未完全闭合。精确推导主要在简化模型中最强；一般 Transformer、AdamW、LayerNorm、数据增强、课程学习和大规模预训练仍需要更直接的定量理论。上文开放问题正是该研究纲领需要变得更可预测、更工程化的地方。

14.3 底线

刘子寅的贡献最好描述为一种理论风格：先识别参数对称性，再测量有限步长随机修正，随后推导被选择的代表元或相，最后检验该代表元能否解释坍缩、平衡、表征对齐、压缩或生物实现。该纲领的价值不在于已经解决所有深度学习理论问题，而在于为许多原本互不相干的现象提供了共同机制。

来源与图表说明

本报告基于截至 2026-07-08 可见的公开网页和公开 arXiv/OpenReview 风格记录整理。所有图都是为本报告重新绘制的原创解释性示意图，并非复制自刘子寅论文。除非图中明确由玩具方程生成，图形均旨在澄清机制，不代表复现实验结果。

参考文献

[1] Liu Ziyin. Personal website. https://www.mit.edu/~ziyinl/. Accessed July 8, 2026.

[2] Liu Ziyin. Publications page. https://www.mit.edu/~ziyinl/publication.html. Accessed July 8, 2026.

[3] Ziyin Liu, Zhikang Wang, Paul P. Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. *Deep Gamblers: Learning to Abstain with Portfolio Theory*. NeurIPS, 2019. https://arxiv.org/abs/1907.00208.

[4] Ziyin Liu, Zhikang Wang, and Masahito Ueda. *LaProp: Separating Momentum and Adaptivity in Adam*. 2020. https://arxiv.org/abs/2002.04839.

[5] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. *Neural Networks Fail to Learn Periodic Functions and How to Fix It*. NeurIPS, 2020. https://arxiv.org/abs/2006.08195.

[6] Kangqiao Liu, Liu Ziyin, and Masahito Ueda. *Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent*. ICML, 2021. https://arxiv.org/abs/2012.03636.

[7] Zhang Zhiyi and Liu Ziyin. *On the Distributional Properties of Adaptive Gradients*. UAI, 2021. https://arxiv.org/abs/2105.07222.

[8] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. *Strength of Minibatch Noise in SGD*. ICLR, 2022. https://arxiv.org/abs/2102.05375.

[9] Liu Ziyin, Botao Li, James B. Simon, and Masahito Ueda. *SGD with a Constant Large Learning Rate Can Converge to Local Maxima*. ICLR, 2022. https://arxiv.org/abs/2107.11774.

[10] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. *Logarithmic Landscape and Power-Law Escape Rate of SGD*. ICML, 2022. https://arxiv.org/abs/2105.09557.

[11] Liu Ziyin, Botao Li, and Xiangming Meng. *Exact Solutions of a Deep Linear Network*. NeurIPS, 2022; Journal of Statistical Mechanics: Theory and Experiment, 2023. https://arxiv.org/abs/2202.04777.

[12] Zihao Wang and Liu Ziyin. *Posterior Collapse of a Linear Latent Variable Model*. NeurIPS, 2022. https://arxiv.org/abs/2205.04009.

[13] Liu Ziyin and Masahito Ueda. *Zeroth, First, and Second-Order Phase Transitions in Deep Neural Networks*. Physical Review Research, 2023. https://arxiv.org/abs/2205.12510.

[14] Liu Ziyin, Ekdeep Singh Lubana, Masahito Ueda, and Hidenori Tanaka. *What Shapes the Loss Landscape of Self-Supervised Learning?* ICLR, 2023. https://arxiv.org/abs/2210.00638.

[15] James B. Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J. Fetterman, and Joshua Albrecht. *On the Stepwise Nature of Self-Supervised Learning*. ICML, 2023. https://arxiv.org/abs/2303.15438.

[16] Liu Ziyin and Zihao Wang. *Sparsity by Redundancy: Solving L_1 with SGD*. ICML, 2023. https://arxiv.org/abs/2210.01212.

[17] Liu Ziyin, Botao Li, and collaborators. *Type-II Saddles and Probabilistic Stability of Stochastic Gradient Descent*. 2023. https://arxiv.org/abs/2303.13093.

[18] Liu Ziyin, Hongchao Li, and Masahito Ueda. *Law of Balance and Stationary Distribution of Stochastic Gradient Descent*. Physical Review E. https://arxiv.org/abs/2308.06671.

[19] Liu Ziyin. *Symmetry Induces Structure and Constraint of Learning*. ICML, 2024. https://arxiv.org/abs/2309.16932.

[20] Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. *Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent*. NeurIPS, 2024. https://arxiv.org/abs/2402.07193.

[21] Liu Ziyin, Yizhou Xu, and Isaac Chuang. *Remove Symmetries to Control Model Expressivity*. ICLR, 2025. https://arxiv.org/abs/2408.15495.

[22] Yizhou Xu and Liu Ziyin. *When Does Feature Learning Happen? Perspective from an Analytically Solvable Model*. ICLR, 2025. https://arxiv.org/abs/2401.07085.

[23] Liu Ziyin, Isaac Chuang, Tomer Galanti, and Tomaso Poggio. *Formation of Representations in Neural Networks*. ICLR, 2025. https://arxiv.org/abs/2410.03006.

[24] Liu Ziyin, Yizhou Xu, Tomaso Poggio, and Isaac Chuang. *Parameter Symmetry Breaking and Restoration Determines the Hierarchical Learning in AI Systems*. 2025. https://arxiv.org/abs/2502.05300.

[25] Liu Ziyin, Yizhou Xu, and Isaac Chuang. *Neural Thermodynamics: Entropic Forces in Deep and Universal Representation Learning*. NeurIPS, 2025. https://arxiv.org/abs/2505.12387.

[26] Liu Ziyin and Isaac Chuang. *Proof of a Perfect Platonic Representation Hypothesis*. 2025. https://arxiv.org/abs/2507.01098.

[27] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. *Understanding the Emergence of Multimodal Representation Alignment*. ICML, 2025. https://arxiv.org/abs/2502.16282.

[28] Qiyao Liang, Daoyuan Qian, Liu Ziyin, and Ila Fiete. *Compositional Generalization Requires More Than Disentangled Representations*. ICML, 2025. https://arxiv.org/abs/2501.18797.

[29] Qianli Liao, Liu Ziyin, Yulu Gan, Brian Cheung, Mark Harnett, and Tomaso Poggio. *Self-Assembly of a Biologically Plausible Learning Circuit*. 2024. https://arxiv.org/abs/2412.20018.

[30] Liu Ziyin, Isaac Chuang, and Tomaso Poggio. *Heterosynaptic Circuits Are Universal Gradient Machines*. 2025. https://arxiv.org/abs/2505.02248.

[31] David Koplow, Tomaso Poggio, and Liu Ziyin. *Emergence of Hebbian Dynamics in Regularized Non-Local Learners*. ICML, 2026. https://arxiv.org/abs/2505.18069.

[32] Marc Gong Bacvanski, Liu Ziyin, and Tomaso Poggio. *On Biologically Plausible Learning in Continuous Time*. 2026. https://arxiv.org/abs/2510.18808.

[33] Yongyi Yang, Tomaso Poggio, Isaac Chuang, and Liu Ziyin. *Topological Invariance and Breakdown in Learning*. 2025. https://arxiv.org/abs/2510.02670.

[34] Yongyi Yang and Liu Ziyin. *An Equivariance Toolbox for Learning Dynamics*. 2025. https://arxiv.org/abs/2512.21447.

[35] Hong-Yi Wang, Di Luo, Tomaso Poggio, Isaac L. Chuang, and Liu Ziyin. *A Universal Compression Theory: Lottery Ticket Hypothesis and Superpolynomial Scaling Laws*. ICLR, 2026. https://arxiv.org/abs/2510.00504.

[36] Yizhou Xu, Pierfrancesco Beneventano, Isaac Chuang, and Liu Ziyin. *Does SGD Seek Flatness or Sharpness? An Exactly Solvable Model*. 2026. https://arxiv.org/abs/2602.05065.

[37] Liu Ziyin, Yuanjie Ren, Adam Levine, and Isaac Chuang. *Thermodynamic Irreversibility of Training Algorithms*. 2026. https://arxiv.org/abs/2605.21933.