arXiv:2601.10765v1 [cs.NE] 14 Jan 2026

作为进化的剪枝：神经网络中通过选择动力学涌现的稀疏性

Zubair Shah 科学与工程学院 哈马德·本·哈利法大学 多哈，卡塔尔zshah@hbku.edu.qa

Noaman khan 科学与工程学院 哈马德·本·哈利法大学 多哈，卡塔尔nokh88609@hbku.edu.qa

摘要

神经网络通常在高度过参数化状态下进行训练，但经验证据一致表明，许多参数在学习过程中会变得冗余。现有的大多数剪枝方法通过显式干预来引入稀疏性，例如基于重要性的阈值处理或正则化惩罚项，这实际上将剪枝视为对已训练模型做出的集中式决策。这一假设与基于梯度的训练所具有的分布式、随机且依赖路径的特性并不相符。我们提出一种关于剪枝的演化视角：将参数组（神经元、滤波器、注意力头）建模为种群，其影响力在选择压力下持续演变。从这一视角看，剪枝对应于种群灭绝：适应度持续偏低的组件会逐渐失去影响力，因此无需离散剪枝计划或均衡计算即可将其移除。我们将神经网络剪枝形式化为种群质量上的演化过程，推导出支配质量演变的选择动力学，并将适应度与局部学习信号联系起来。我们使用种群规模化多层感知机（784–512–256–10）在 MNIST 上验证该框架，其中包含 768 个可剪枝神经元种群。所有动力学的测试准确率均达到接近 98% 的稠密基线。我们对目标稀疏度为 35–50% 的训练后硬剪枝进行了基准测试：剪枝 35% 时，测试准确率为 ≈95.5%；剪枝 50% 时，根据具体动力学不同，测试准确率为 ≈88.3–88.6%。这些结果表明，演化选择能够在训练期间不采用显式剪枝计划的情况下，产生可测量的准确率–稀疏度权衡。

1 引言

神经网络通常使用多于最终所需的参数进行训练，成功的权重剪枝以及稀疏可训练子网络的存在便证明了这一点 [Han 等人，2015，Frankle 和 Carbin，2019]。剪枝方法利用这种冗余来减小模型大小和计算成本。然而，大多数现有方法都通过显式干预来施加稀疏性，例如基于重要性的阈值处理 [LeCun 等人，1990，Hassibi 和 Stork，1993], 正则化惩罚项[Louizos 等人，2018], 或离散剪枝计划 [Zhu 和 Gupta，2017]。这些方法都隐含地假设，剪枝是由集中式优化器或外部控制器在训练好的模型上做出的有意决策。

这一假设与神经网络训练所具有的去中心化、随机性和路径依赖性并不相容。单个神经元或滤波器不会推理全局模型结构、优化长期目标，也不会彼此明确协调。相反，学习是在局部梯度更新、噪声反馈以及对表征能力的隐式竞争中展开的 [Saxe 等人，2019]。从这一视角来看，稀疏性未必是理性优化决策的结果；它也可能是选择动力学随时间作用的结果。

这一观察引出了一个截然不同的问题：即使没有明确设定剪枝目标或强制实施离散剪枝机制，稀疏性是如何在训练过程中涌现的？要回答这个问题，就必须超越静态优化视角，转而动态地理解模型组件在学习过程中如何存续、衰退或消失。

在这项工作中，我们提出一种关于神经网络剪枝的演化视角。我们不将神经元或参数组建模为解决优化问题的理性行动者，而是将它们建模为承受选择压力的种群。每个参数组都在有限的表征影响力中竞争，其普遍程度会随时间演变，具体取决于反映其在训练过程中所作贡献的适应度信号。在这一视角下，剪枝对应于种群灭绝：适应度持续偏低的组件会逐渐失去影响力并被淘汰，无需显式作出剪枝决策、制定阈值规则或进行均衡计算。

这种演化框架与经典博弈论方法截然不同。我们不假设存在理性行为、最佳响应或向纳什均衡收敛。相反，我们关注局部且渐进地运作的复制子式选择动力学， [Hofbauer and Sigmund，1998]稀疏性由选择作用下的差异化增长与衰减驱动，是一种随时间演变的现象，而非全局优化问题的静态解。

演化视角具有若干优势。首先，它与基于梯度的训练所依托的随机性和去中心化机制天然契合。其次，它解释了为何剪枝往往是渐进发生的，并呈现出有效的不可逆性，而非通过外部强加的阈值处理骤然完成。第三，它为理解剪枝是学习动力学的涌现属性，而非施加于已训练模型上的辅助目标，提供了有原则的基础。

本文将神经网络剪枝形式化为参数种群上的演化过程。我们推导了支配参数组种群质量（参与度）演化的选择动力学，并在温和的适应度假设下分析其灭绝行为。重要的是，无须明确的剪枝计划，也只需承受温和的衰减压力，稀疏性便会在训练过程中通过进化选择动力学自然涌现。

1.1 贡献

本文的主要贡献如下：

1. 我们提出了一种神经网络剪枝的演化表述，无须依赖理性、最佳响应动力学或均衡假设。2. 我们将参数组建模为种群，其影响力在由局部适应度信号驱动的选择动力学下演化。3. 我们从理论上刻画了在温和的适应度异质性下，稀疏性如何通过种群灭绝涌现。4. 我们从动态的过程层面解释了剪枝如何作为学习的涌现结果。

1.2 范围与局限

本研究旨在将神经网络剪枝严谨地表述为一种演化选择过程，并刻画其灭绝行为。当适应度异质性足够显著时，稀疏性可能自然涌现；但所提出的框架并非为了在高度对称或冗余占主导的情形下强制进行激进剪枝。我们的目标不是追求最大程度的压缩，而是理解选择动力学何时以及为何会诱导稀疏性

在训练过程中涌现，从而为未来纳入更丰富适应度信号的扩展奠定基础。

2 相关工作

神经网络剪枝问题已从多个角度得到广泛研究，涵盖从经典敏感性分析到现代动态稀疏训练方法。本节综述剪枝的主要方法，并将我们的演化框架置于这一更广阔的研究背景中。

早期有理论依据的剪枝方法旨在找出哪些参数可以移除，同时尽可能降低对损失函数的影响。Optimal Brain Damage（最优脑损伤） [LeCun 等人， 1990]提出了利用二阶信息估计损失对参数移除的敏感度这一思路：通过计算 Hessian 矩阵的对角线，近似估计将单个权重置零所导致的损失变化。Optimal Brain Surgeon（最优脑外科） [Hassibi 和 Stork，1993]通过使用完整的 Hessian 矩阵，并考虑剪枝后的最优权重调整，扩展了这一框架，从而能够更准确地进行重要性估计。尽管这些方法为剪枝奠定了理论基础，但它们依赖计算开销高昂的二阶计算，因此难以应用于大型现代架构。当前研究探索了各种近似方法和可扩展的替代方案，但理论严谨性与计算效率之间的根本矛盾依然存在。

基于幅值的剪枝是一种更简单且更具可扩展性的替代方法，其依据的直觉是：幅值较小的参数对网络输出的贡献较小。Han 等人 [Han 等人，2015] 证明了，迭代式基于幅值的剪枝结合重新训练，在保持准确率的同时，可以实现较高的压缩率。由于其简单有效，这种方法已成为标准基线。基于正则化的方法则提供了互补视角：它们不是事后强制施加稀疏性，而是在训练过程中鼓励稀疏性。经典的 ℓ_1 和 ℓ_2 正则化 [Hanson 和 Pratt， 1988] 通过权重衰减和软阈值化效应间接促进稀疏性。近期关于随机 ℓ_0 正则化 [Louizos 等人， 2018]则通过连续松弛学习二值门，更直接地处理稀疏性，从而实现端到端可微稀疏训练。基于相关性的剪枝方法 [Mozer 和 Smolensky， 1989] 为单个神经元或参数组分配重要性评分，并迭代移除相关性最低的单元；通常还会加入重新训练步骤，使网络能够适应降低后的容量。

结构化剪枝已成为一个重要方向，其发展动力源于对硬件高效压缩的需求。与移除单个权重的非结构化剪枝不同，结构化剪枝会移除整个滤波器、通道或神经元，从而得到可在标准硬件上高效执行的稠密子网络 [Li 等人， 2019]。滤波器级剪枝方法 [Li 等人， 2017]会根据重要性指标对卷积滤波器进行排序，例如权重的 ℓ_1-范数或基于梯度的敏感性度量，然后移除排名靠后的滤波器，并对生成的网络进行微调。软滤波器剪枝 [He 等人， 2018] 引入了连续门控机制，使训练过程中能够进行可微通道选择，从而实现架构与权重的联合优化。Network Slimming（网络瘦身）方法 [Liu 等人， 2017] 将逐通道缩放因子与诱导稀疏性的正则化相结合，在训练过程中学习应剪除哪些通道。这些方法表明，结构化稀疏性可以通过学习有效获得，而非依靠启发式方法强行施加；不过，它们通常需要谨慎调整正则化强度和剪枝计划。

动态稀疏训练代表着一种范式转变：在整个训练过程中保持稀疏性，而不是先进行稠密训练再剪枝。Mocanu 等人 [Mocanu 等人， 2018] 提出了在训练过程中持续调整稀疏连接的概念：根据幅值定期移除连接，并依据梯度信息重新生长连接。这

种方法能够从头开始训练稀疏网络，无需经过稠密预训练阶段，挑战了剪枝必须从稠密网络开始这一传统观念。彩票假设 [Frankle 和 Carbin， 2019] 为稀疏子网络的存在提供了理论和实证支持：这些子网络单独训练时能够达到与稠密网络相当的性能。这表明，过参数化的优势或许主要体现在初始化阶段，而非贯穿整个训练过程。后续研究探讨了彩票子网络胜出的原因 [Zhang 等人，2021], 稀疏网络中的梯度流受到何种影响 [Evci 等人， 2022], 以及向图神经网络的扩展 [Chen 等人， 2021]。这些研究共同表明，稀疏性不仅仅是一种压缩技术，也是神经网络实现高效训练的一个基本方面。

大语言模型的兴起为剪枝研究带来了新的挑战与机遇。现代 Transformer 的规模动辄数十亿甚至数万亿个参数，使传统的迭代剪枝方法在计算上难以承受。SparseGPT [Frantar 和 Alistarh，2023] 通过结合近似重构的一次性剪枝来应对这一挑战：在逐层做出剪枝决策的同时，近似保留每一层的输出。这种方法无需重新训练即可剪枝超大规模模型，但需要以牺牲部分准确率为代价来换取计算效率。WANDA [Sun 等人， 2023] 提出了一种简单而有效的方法，将权重幅值与激活统计量结合起来，指出权重的重要性不仅取决于其幅值，也取决于其输入的典型幅值。LoSparse [Li 等人， 2023] 则采取了不同的思路，将低秩分解与稀疏近似结合起来，利用了大语言模型中的权重矩阵通常具有较低有效秩这一观察结果。这些方法表明，大规模剪枝要求我们重新思考重要性评估标准，以及应用剪枝决策的计算策略。

除了剪枝技术本身，我们的研究还借鉴了演化动力学和学习动力学理论中的基础概念。Taylor 和 Jonker 在进化博弈论中提出的复制子动力学 [Taylor 和 Jonker， 1978]以及 Hofbauer and Sigmund 对其进行的深入研究 [Hofbauerand Sigmund， 1998], 描述了种群频率如何在选择压力下根据相对适应度而演变。这些动力学已被用于对生物进化、文化传播以及多智能体系统中的策略学习进行建模 [Sandholm， 2010]。关键洞见在于，种群动力学可以通过基于相对表现的差异性存活来理解，而非通过显式优化或均衡寻求行为来理解。我们的研究将这些原则用于神经网络剪枝，把参数组视为种群，它们在由其对学习的贡献所决定的选择压力下演变。

近年来，从动力系统的视角理解神经网络训练逐渐受到关注。Saxe 等人 [Saxe 等人， 2019] 提出了深度网络语义发展的数学理论，指出学习动力学呈现出可预测的结构，这种结构源于网络架构、损失景观与优化算法之间的相互作用。关于梯度下降隐式偏置的研究 [Gunasekar 等人， 2018]表明，优化轨迹本身携带着损失函数所未明确指定的信息，因此即使在过参数化的情形下，也会使某些特定解更受偏好。这些研究共同表明，要理解神经网络的行为，就必须将训练过程建模为一个动力系统，而不能仅仅描述最终训练所得模型的属性。我们的演化框架将这一视角扩展到剪枝，认为稀疏性涌现应被理解为一种由选择动力学引发的时间性现象，而非静态优化问题。

我们的方法在概念基础上与现有剪枝方法有着根本区别。我们不通过显式正则化、离散剪枝计划或训练后压缩来强制实现稀疏性，而是将稀疏性建模为训练期间竞争动力学所涌现的结果。参数组并非被视为追求目标优化的理性主体，而是被视为根据其对学习的贡献而承受选择压力的种群。这一视角自然解释了剪枝中若干令人费解的现象：为何剪枝往往是渐进发生而非突然发生，为何某些神经元在被移除前会逐渐变得不活跃，以及为何

冗余参数可以在未显式检测冗余的情况下被消除。我们以演化动力学而非优化理论为剪枝奠定基础，由此提供了一个统一框架，既将现有启发式方法涵盖为特例，也为设计与基于梯度的学习的分散式、随机性本质相契合的新剪枝算法提供了有原则的指导。

3 为什么均衡假设不适用于神经网络训练

尽管许多剪枝方法并未明确引入博弈论均衡，但它们往往依赖对训练动力学的静态或类似均衡的解释。本节将阐明，为何此类假设对神经网络训练而言过于严苛。更广泛地说，许多关于学习系统的理论抽象，将参数建模为优化定义明确的目标，并通过理性调整或全局协调收敛至稳定解。尽管这些假设在数学上便于处理，却与神经网络实际训练所依赖的分散、随机且受路径影响的机制并不相符。

首先，神经网络训练是分散且局部的。单个神经元或滤波器无法获取有关模型、数据集或长期结果的全局信息。参数更新由通过反向传播计算的局部梯度驱动 [Rumelhart 等人， 1986], 这些梯度提供的是有噪声且不完整的信号，并且取决于所有其他参数的当前状态。不存在任何机制能让单个组件评估最佳响应，或从全局角度预测其他组件的策略。

其次，学习动力学具有随机性且依赖于路径。小批量训练 [Bottou，2010], 随机初始化和非凸损失景观 [Li 等人， 2018] 都会给优化过程带来显著的变异性。训练早期的微小扰动可能导致截然不同的轨迹，即使最终性能相近也是如此。在这种情况下，假设系统会收敛到唯一或稳定解的均衡概念往往过于僵化，无法捕捉观察到的行为。

第三，神经网络是在有限理性条件下训练的。参数不会明确地最大化长期效用函数，也不会针对未来状态求解优化问题。相反，学习通过渐进、短视的更新来进行，这些更新只响应即时反馈。这种有限理性使得难以为假设理性主体或单个参数层面的均衡寻求行为的模型提供合理依据。

考虑稀疏性时，这些局限尤为明显。实践中，剪枝很少表现为突然转变为稳定的稀疏配置。相反，冗余会逐渐减少：一些神经元变得越来越不活跃，其梯度逐渐减小，对网络的影响力也随时间消退。这一过程更像是选择与灭绝，而不是收敛到静态均衡。

此外，基于均衡的解释隐含地假设，稀疏配置是系统有意选择的。然而，神经网络很少以明确的剪枝目标进行训练，而且即使在稠密训练机制下，也常常会涌现出稀疏性。这表明，与其将剪枝视为静态优化问题的解，不如将其理解为竞争动力学随时间演化的结果。

这些观察促使我们从以均衡为中心的推理转向动态的进化视角。我们不再询问网络是否会收敛到稀疏均衡，而是探究稀疏性如何在训练过程中通过局部相互作用、选择压力和差异性存活涌现。这一视角不要求理性主体、全局协调或均衡计算，只要求参数组在贡献能力上存在差异，且这种差异会影响它们随时间的持续存在。

在本文余下部分，我们将采用这一进化视角。我们将参数组建模为种群，其影响力会根据由贡献和冗余驱动的选择动力学而演变。在这一表述下，剪枝自然地表现为种群灭绝，从过程层面解释稀疏性，同时与神经网络训练的去中心化和随机性相一致。

4 参数组作为种群

我们首先从进化论的角度重新审视神经网络的组成部分。我们不再将单个参数视为决策代理，而是将参数组建模为其影响力随时间演变的种群。这一转变使我们能够将剪枝描述为选择与灭绝的过程，而非优化的结果。

4.1 参数组

考虑一个经过训练以最小化损失函数 $\mathcal{L}(\theta)$ 的神经网络 $f(x;\,\theta)$。与许多剪枝方法一样，我们将参数向量 θ 划分为一组参数组：

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_N\} \tag{1}$$

其中，每个组都可能对应一个神经元、卷积滤波器、注意力头或其他具有结构意义的单元。这种抽象使该框架能够涵盖非结构化和结构化稀疏化，同时不受架构细节的限制。

4.2 种群质量与影响力

为刻画每个参数组的普遍程度，我们为每个群体 θ_i 关联一个非负标量变量

$$p_i(t) \geq 0 \tag{2}$$

我们将其解释为该群体在训练时间 t的种群质量 或影响水平 。群体 θ_i 对网络的有效贡献为

$$\tilde{\theta}_i(t) = p_i(t)\,\theta_i \tag{3}$$

当 $p_i(t)$ 较大时，该群体会对网络的计算产生显著影响；当 $p_i(t)$ 接近零时，该群体实际上会变得不活跃。重要的是， $p_i(t)$ 并非二值剪枝掩码，也不是由外部规则设定的。它会随时间连续演化，反映影响力的逐渐增长或衰减。

4.3 将其解读为进化种群

在这一表述下，参数组的行为类似于进化系统中的种群：

- 每个群体都有一个种群质量 ，用于决定其普遍程度。
- 群体通过对共享损失的影响间接竞争。
- 冗余或无效的群体会随着时间推移逐渐失去影响力。
- 贡献度高的群体会持续存在或不断壮大。

至关重要的是，这种解释并不假设参数组会优化目标、计算效用或对他人作出策略性反应。种群质量的变化反映的是选择压力，而非理性选择。

这一差异至关重要。演化动力学描述的是系统如何通过差异性存活而演化，而非个体如何解决优化问题。因此，该框架仍与基于梯度的学习所具有的局部性、短视性和随机性相容。

4.4 剪枝作为群体灭绝

从种群视角来看，剪枝自然对应于灭绝事件。当参数组的种群质量 $p_i(t)$逐渐趋近于零时，该组便不再影响网络，因此可以将其移除，而不会对性能产生实质性影响。这种解释与实践中对剪枝的实证观察十分吻合：神经元或滤波器在被明确移除之前，往往会逐渐变得不活跃。我们不把剪枝视为一个离散决策，而是将其视为连续演化过程的最终阶段。

4.5 向演化动力学的过渡

剩下的问题是，种群质量 $p_i(t)$在训练期间如何演化。在进化系统中，这种演化受选择动力学支配：适应度较高的种群会得到扩增，而适应度较低的种群则会受到抑制。在下一节中，我们将通过引入支配种群质量时间演化的演化动力学，把这一直觉形式化。这些动力学只需要一个相对适应度的概念，无须依赖均衡假设、理性或全局协调。

5 演化选择动力学

现在，我们将第4节中引入的种群质量的时间演化形式化。我们的目标是对稀疏性如何通过选择动力学涌现进行建模，而不假设存在理性决策、效用最大化或均衡计算。为清晰起见，本节始终关注连续时间动力学。实践中使用的离散时间实现可直接推导得出。

5.1 一般选择原理

令 $p_i(t) \geq 0$ 表示参数组 i 在时刻 t的种群质量。我们假设每个群体都关联着一个适应度信号

$$\phi_i(p, \theta) \tag{4}$$

该信号反映了在网络及其他参数组的当前状态下，该群体对学习的相对贡献。

我们不假设适应度经过优化，也不假设群体明确知晓其适应度。适应度只是一种调节增长或衰减的标量。

一大类演化动力学可以写成如下形式

$$\dot{p}_i(t) = p_i(t)\, \Phi_i(p(t)) \tag{5}$$

其中， Φ_i 是根据适应度比较得出的选择函数。这种乘法结构确保：

- 质量为零的种群会继续灭绝，
- 增长和衰减与当前影响力成正比，
- 灭绝是逐渐发生的，而非突然发生。

在温和的正则性条件下，适应度持续较低的参数群体，其种群质量会单调衰减。

5.2 复制子动力学

经典的演化动力学是复制子动力学，它在进化生物学和博弈论中得到了广泛研究。在这种设定下，每个种群的增长率与其相对于种群平均适应度的适应度成正比。

形式上，复制子动力学可表示为

$$\dot{p}_i = p_i(\phi_i - \bar{\phi}), \quad \bar{\phi} = \frac{\sum_j p_j \phi_j}{\sum_j p_j} \tag{6}$$

此处， p_i 表示的是未经归一化的种群质量，而非概率。因此，平均适应度被定义为按种群质量加权的平均值。直观地说，适应度高于种群平均值的参数组，其影响力往往会增加；而适应度低于平均值的参数组，其影响力往往会衰减。

复制子动力学具有以下几个特性，使其非常适合用于建模剪枝：

- 它们只需要相对适应度信息。
- 在适应度差异稳定的情况下，被支配种群会呈指数衰减。
- 在没有突变的情况下，灭绝实际上是不可逆的。

在这种动力学下，稀疏性会自然涌现：持续贡献较低的参数组会逐渐走向灭绝。

5.3 归一化增长动力学

尽管复制子动力学在理论上颇具吸引力，但它们可能对适应度估计中的噪声较为敏感。因此，我们还考虑一种归一化增长动力学，它在提升数值稳定性的同时，保留了相同的定性选择行为。

在这一表述中，种群质量按照以下方式演化：

$$\dot{p}_i = p_i \left(\phi_i - \frac{1}{Z} \sum_j p_j \phi_j \right) \tag{7}$$

其中， $Z = \sum_j p_j$ 是总种群质量，作为归一化常数，用于保持不同种群之间的相对尺度。

这种动力学与经典复制子动力学的主要区别在于归一化和缩放。关键在于：

- 定性的灭绝行为得以保留，被支配种群逐渐衰减至零。

- 在相同的归一化约定下，定点相同（或在丰度形式下仅相差一个比例因子）。

- 在带噪声的梯度估计下，增长率更加稳定。

因此，归一化增长动力学是一种实用的替代方案，既保留了相同的诱导稀疏性的选择机制，又提高了基于梯度的训练循环中的数值稳定性。

5.4 选择–突变动力学

纯选择动力学可能导致过早灭绝，尤其是在训练早期，此时适应度估计带有噪声。为了解决这一问题，我们考虑选择–突变动力学，它引入一个小的扩散项，使种群能够从濒临灭绝的状态中恢复。

该动力学的形式为

$$\dot{p}_i = p_i(\phi_i - \bar{\phi}) + \mu\left(\frac{1}{N} - p_i\right) \tag{8}$$

其中，$\bar{\phi} = \frac{\sum_j p_j \phi_j}{\sum_j p_j}$是种群加权平均适应度，$\mu > 0$ 是突变率，N 是参数组的数量。

突变项:

- 可防止立即灭绝,
- 可维持探索多样性,
- 允许暂时受到抑制的群体重新出现。

当 $\mu \to 0$ 时，动力学会恢复标准的复制子行为。对于较小但非零的 μ，稀疏性仍会出现，但灭绝会延迟发生，且对噪声的鲁棒性更强。

5.5 灭绝与稀疏性

在这三种动力学中，都呈现出一种共同的定性行为：适应度持续较低的参数组，其种群质量会单调衰减。在适应度分离条件较为温和的假设下，这些参数组的影响力会随时间推移逐渐趋近于零。这一灭绝过程为剪枝提供了自然解释。冗余参数组并非通过阈值处理或优化被移除，而是在演化压力下逐步淘汰。具体轨迹取决于所选择的动力学，但其定性结果——涌现稀疏性——得以保留。

5.6 讨论

这里介绍的演化动力学是同一选择原则的不同具体形式。复制子动力学提供了经典的理论基准，归一化增长动力学提升了数值稳定性，而选择–突变动力学则增强了抗噪性。重要的是，这些动力学的区别在于稀疏性如何随时间涌现，而不在于稀疏性是否会涌现。这一视角使我们能够将剪枝视为一个动态过程，并分析不同的演化压力如何在训练过程中塑造稀疏性轨迹。下一节将具体说明实践中使用的适应度函数，并把这些演化动力学与源自神经网络训练的具体学习信号联系起来。

6 适应度函数与学习信号

第4节中介绍的演化动力学需要一个适应度 概念，用以决定种群质量如何随时间演化。在本文中，适应度并非被解释为需要优化的效用或需要最大化的奖励。相反，它充当一种局部选择信号，反映在网络当前状态下某个参数组对学习的相对贡献。

本节的目标有两个：（i）以兼容基于梯度的训练的方式定义适应度；（ii）使这一表述足够通用，能够支持多种演化动力学，同时不改变其定性行为。

6.1 适应度的设计原则

我们对与参数组 i 相关联的适应度信号 ϕ_i 设定三项最低要求：

1. **局部性**。 适应度应当能够根据训练过程中已有的信息计算得出，例如梯度或激活值，且无需全局协调。2. **相对性**。 适应度只需在相对比较的意义上有用；绝对尺度并不重要。3.**上下文依赖性**。 适应度应取决于其他参数组的存在及其影响力，从而能够隐式地惩罚冗余。

这些要求确保适应度反映的是选择压力，而非理性优化。

6.2 基于贡献的适应度

适应度信息的一个自然来源是损失对参数组影响力的敏感度。˜令 p_i 表示群体 i的种群质量，并考虑有效参数 $\theta_i = p_i\theta_i$。我们将群体 i 的贡献定义为

$$c_i = \left| \frac{\partial \mathcal{L}}{\partial p_i} \right| \tag{9}$$

它衡量群体影响力的变化如何影响训练损失。直观而言，对损失影响较大的群体会获得较高的适应度，而影响微乎其微的群体则会受到惩罚。通过自动微分即可轻松获得这一信号，并且它与基于梯度的学习自然契合。

6.3 冗余与竞争效应

仅凭贡献无法体现冗余。多个参数组可能各自具有较高的敏感度，但合在一起却提供了重叠的功能。我们不引入显式的成对竞争项，而是依靠演化动力学的相对性 来隐式地引发竞争。

在复制器和归一化增长动力学中，种群增长取决于适应度相对于种群平均值的大小。因此：

- 冗余群体会争夺影响力，
- 只有部分贡献者会保留下来，
- 即使其他参数的绝对贡献不为零，它们也会衰减。

这种机制无需显式地对所有参数群体之间的相互作用进行建模，就能抑制冗余。

6.4 正则化适应度与稳定性

在实践中，原始贡献信号可能存在噪声，尤其是在训练初期。为提高稳定性，我们考虑如下形式的正则化适应度：

$$\phi_i = c_i - \lambda\, r(p_i) \tag{10}$$

其中，$r(\cdot)$ 是单调递增的正则化函数，$\lambda \geq 0$ 控制其强度。在我们的实验中，我们使用 $r(p_i) = p_i$，它对种群质量施加线性衰减压力。该项可抑制不受控制的增长，并加速弱种群灭绝，其作用类似于 ℓ_2 惩罚，但它应用于进化选择框架内，而非作为损失正则化。

正则化有两个作用：

- 防止种群质量不受控制地增长，
- 加速长期处于弱势的种群灭绝。

重要的是，正则化塑造的是灭绝的轨迹和稳定性，而非其是否发生。在所考察的所有演化动力学中，在温和条件下，被支配种群都会被消除，这与具体采用哪种正则化方法无关。正则化主要影响选择动力学下稀疏性涌现的速度、平滑度和稳健性。

6.5 与演化动力学的兼容性

上述适应度定义与第4节中介绍的三种演化动力学均兼容：

- 复制子动力学使用相对适应度差异，
- 归一化增长动力学可稳定尺度，
- 选择–突变动力学可在噪声下平缓灭绝过程。

尽管这些动态过程在时间行为上有所不同，但它们具有相同的定性结果：适应度持续较低的参数群体会失去影响力，并最终被剪除。

6.6 讨论

此处提出的适应度公式刻意保持简单。我们的目标并非设计高度专门化的剪枝信号，而是证明，基本且可在局部获取的学习信号足以驱动进化稀疏性。这种简洁性进一步印证了本文的核心论点：剪枝可以作为训练期间选择动力学的内在结果自然产生，无需明确的剪枝目标、重要性启发式或均衡计算。在下一节中，我们将介绍这些演化动力学如何在实践中实现，并概述用于研究不同选择机制下涌现稀疏性的实验方案。

7 进化剪枝算法

现在，我们将介绍一种实用算法，用于实现第 5 节和第 6 节中提出的进化选择动力学。该算法旨在与标准的基于梯度的训练无缝集成，并通过统一的更新规则支持多种演化动力学。

7.1 概述

核心思想是在训练过程中共同演化网络参数和种群质量。网络参数使用标准随机梯度下降进行更新，而种群质量则根据由学习产生的适应度信号，按照进化选择动力学进行演化。训练期间不会将剪枝作为单独步骤执行。相反，参数组的种群质量会在选择压力下逐渐衰减，使相应参数组逐步受到抑制。只有在训练结束且种群质量已经收敛后，才会执行显式移除。

7.2 联合训练与演化循环

令 θ 表示网络参数，令 $p = (p_1, \ldots\ldots, p_N)$ 表示与参数组相关联的种群质量。训练过程在每次迭代或小批量中交替执行两次更新。首先，在参数更新（学习步骤）中，给定当前种群质量 $p(t)$，我们使用梯度下降更新网络参数：

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}(\theta, p) \tag{11}$$

其中，有效参数由 $\tilde{\theta}_i = p_i \theta_i$ 给出。其次，在种群更新（选择步骤）中，我们计算适应度信号 ϕ_i ，并根据所选的演化动力学更新种群质量。这种分离体现了学习（更新表征）与选择（调节存活）之间的概念区别。

虽然对 θ 和 p 的更新是分别进行的，但二者通过损失 $\mathcal{L}(\theta, p)$ 相互耦合。p 的变化会调节有效参数 $\tilde{\theta}_i = p_i \theta_i$，进而影响 θ的梯度。反过来，θ 的变化会改变损失景观，从而影响适应度信号 ϕ_i。这种双向交互意味着学习和选择共同演化，但二者并未进行显式协调，这与去中心化进化视角一致。

7.3 种群更新规则

该算法通过模块化的种群更新步骤支持多种演化动力学。对于复制子动力学：

$$p_i \leftarrow p_i + \eta_p \, p_i (\phi_i - \bar{\phi}), \qquad \bar{\phi} = \frac{\sum_j p_j \phi_j}{\sum_j p_j} \tag{12}$$

对于归一化增长动力学：

$$p_i \leftarrow p_i + \eta_p \, p_i \left(\phi_i - \frac{1}{Z} \sum_j p_j \phi_j \right) \tag{13}$$

其中，$Z = \sum_j p_j$ 是总种群质量，作为归一化常数，用于保持不同种群之间的相对尺度。对于选择–突变动力学：

$$p_i \leftarrow p_i + \eta_p \left[p_i (\phi_i - \bar{\phi}) + \mu \left(\frac{1}{N} - p_i \right) \right] \tag{14}$$

其中， $\bar{\phi} = \frac{\sum_j p_j \phi_j}{\sum_j p_j}$ 是种群加权平均适应度， $\mu > 0$ 是突变率， N 是参数组的数量。

每次更新后，都会将种群质量投影到非负正交象限，以确保满足可行性要求。

7.4 剪枝准则

种群质量通常会收敛到双峰分布，大多数值接近于零，或保持在适中的尺度。训练后，将移除种群质量可忽略不计的参数组：

$$\text{prune group } i \text{ if } p_i < \epsilon \tag{15}$$

其中， $\epsilon < 1$ 是一个较小的数值阈值。该阈值反映的是数值容差，而非重要性截断值；低于此阈值的参数组已通过演化动力学经历了有效灭绝。

7.5 算法概要

算法 1 基于选择动力学的进化剪枝

要求： 训练数据；参数 θ；初始化 $p_i \leftarrow 1$

对于训练轮次 $= 1, \ldots, T$ **执行**
  对于小批次 **执行**
    通过随机梯度下降在 $\mathcal{L}(\theta, p)$ 上更新 θ
    计算适应度 ϕ_i （贡献减去正则化）
    通过选定的动态机制（复制器、归一化或突变）更新 p
    项目 $p_i \geq 0$
  结束循环
结束循环
移除满足 $p_i < \epsilon$ 的参数组
返回剪枝后的模型

7.6 讨论

该算法强调过程，而非预先规定的方案。演化动力学的选择会影响稀疏性涌现的轨迹，但不会改变其底层机制。复制子动力学提供了一个典型基线，归一化增长提升了数值稳定性，而选择–突变动力学则增强了抗噪性。至关重要的是，该算法不需要明确的剪枝计划、重要性评分或均衡计算。剪枝是训练过程中演化压力产生的副产品。在下一节中，我们将评估不同的演化动力学在实践中如何影响稀疏性涌现、稳定性和剪枝轨迹。

8 实验设置

本节实验旨在探究进化选择动力学的定性行为，特别是灭绝轨迹、稳定性性质，以及对适应度异质性的敏感度，而非最大化稀疏性或压缩比。因此，我们有意采用简单的架构和可控的设置，以隔离理论所预测的影响。

表 1：MNIST 剪枝结果：不同稀疏度下的测试准确率、阈值和准确率下降幅度。未进行剪枝时的基准准确率为 0.9808。

稀疏性	复制器		归一化		突变	
(%)	ϵ	准确率（下降幅度）	ϵ	准确率（下降幅度）	ϵ	准确率（下降幅度）
35.0	0.6313	0.9551 (0.0257)	0.6313	0.9551 (0.0257)	0.6329	0.9548 (0.0257)
40.0	0.6375	0.9289 (0.0519)	0.6375	0.9289 (0.0519)	0.6389	0.9261 (0.0544)
45.0	0.6465	0.9217 (0.0591)	0.6465	0.9217 (0.0591)	0.6483	0.9235 (0.0570)
50.0	0.6676	0.8825 (0.0983)	0.6676	0.8825 (0.0983)	0.6689	0.8857 (0.0948)

我们使用 MNIST 数据集 [LeCun 等人， 1998], 其中包含 60,000 张训练图像和 10,000 张手写数字测试图像。每张图像都是一幅 28×28 灰度图像，我们将其展平为一个 784 维输入向量。我们不进行数据增强。MNIST 是图像分类的标准基准测试，具有足够的复杂度，能够观察到有意义的剪枝动态，同时计算上仍易于处理。

我们采用具有以下结构的多层感知机（MLP）：包含 784 个特征的输入层；包含 512 个神经元、使用种群缩放线性变换的第一隐藏层；包含 256 个神经元、同样使用种群缩放线性变换的第二隐藏层；以及包含 10 个神经元、使用标准线性变换的输出层。该模型包含 768 个可剪枝参数组（隐藏层中的 $512 + 256$ 个神经元）。每个隐藏神经元都关联一个种群质量 p_i，其初始值为 1.0。为保留分类能力，输出层不进行剪枝。

我们使用随机梯度下降训练权重 θ，并使用以下三种动力学之一更新种群质量 p：复制器（方程 6）、归一化增长复制子（方程 7）或选择–突变（方程 8）。我们使用基于贡献的公式并结合正则化 $\phi_i = |\partial \mathcal{L}/\partial p_i| - \lambda p_i$计算适应度。

训练配置采用以下超参数：20 个训练轮次，批量大小为 128，权重的学习率 θ 为 0.01，动量为 0.9，种群质量的学习率 p 为 5×10^{-4}，适应度衰减正则化 λ为 1×10^{-3}，突变率 μ为 2×10^{-3}（用于突变动力学），质量上限 Z 为 768.0（用于归一化增长动力学），选择信号裁剪值为 0.10，随机种子固定为 0。所有实验均保持这些超参数不变，以确保公平比较。除非另有说明，我们报告使用固定随机种子进行单次运行的结果；多随机种子评估留待未来工作。

训练完成后，我们通过设置 $p_i = 0$，对质量最低的种群进行硬剪枝。为在受控的稀疏度下进行比较，我们将阈值 ϵ设为 p 的经验分位数，以达到目标剪枝比例，具体为 35%、40%、45% 和 50%。随后，我们评估剪枝后的测试准确率。这种训练后剪枝方法能够隔离演化动力学对种群质量分布的影响，使我们得以研究不同动力学如何塑造稀疏度—准确率权衡。所有参数组统一采用剪枝阈值，确保选择完全由演化后的种群质量决定，而非特定层启发式方法或人工干预。我们报告稠密基线准确率（采用演化动力学训练但未剪枝的模型），以及各目标稀疏度下剪枝后的准确率。

9 结果

表1报告 s 在目标稀疏度下进行训练后硬剪枝后的测试准确率。阈值
ϵ通过以下方式选择 分位数选择，以达到各目标稀疏百分比。

已剪枝 acc确率随目标稀疏度单调下降，反映了预期的

表 2：固定阈值剪枝结果：应用固定种群质量阈值所达到的准确率和稀疏性。未进行剪枝时的基准准确率为 0.9808

阈值	复制器		归一化		突变	
ϵ	稀疏性	准确率（下降幅度）	稀疏性	准确率（下降幅度）	稀疏性	准确率（下降幅度）
0.6	0.0%	0.9808 (0.0000)	0.0%	0.9808 (0.0000)	0.0%	0.9808 (0.0000)
0.7	54.8%	0.8573 (0.1235)	54.8%	0.8573 (0.1235)	54.6%	0.8743 (0.1062)
0.8	63.5%	0.7976 (0.1832)	63.5%	0.7976 (0.1832)	63.5%	0.7942 (0.1863)
0.9	67.2%	0.7725 (0.2083)	67.2%	0.7725 (0.2083)	67.3%	0.8140（0.1665）

准确率–稀疏度权衡。从这些结果中可以观察到几个关键现象。在我们的超参数设置下，复制子动力学与归一化增长动力学产生了相同的结果，这表明在此条件下，两种归一化方案会产生相同的选择行为。这说明在这些条件下，纯复制子动力学已经足够稳定，归一化变体也能提供等效性能。在 50% 剪枝率下，突变动力学的鲁棒性略高（88.57% 对比 88.25%），这与理论预测一致：突变可防止过早灭绝，并在选择压力较高时维持多样性。随着稀疏度增加，准确率平稳下降（从剪除 35% 时的 95.5% 降至剪除 50% 时的 88.3–88.6% ），这表明演化动力学能够对神经元重要性进行有序排名。最冗余的神经元会最先被淘汰，只有在关键神经元开始被移除时，性能才会下降。这些结果表明，演化选择会产生有意义的稀疏信号（p分布），训练后可通过阈值处理来获得剪枝模型，而无需在训练期间制定明确的剪枝计划。在种群间梯度几乎同质的情况下，选择压力较弱，稀疏性可能不会涌现，这与理论分析一致。

为进一步探究种群质量阈值与剪枝有效性之间的关系，我们通过改变阈值 ϵ在一系列取值（0.6、0.7、0.8、0.9）中的表现，并测量由此产生的稀疏性和准确率，开展了固定阈值实验。表2 报告了这些结果：对于低于0.7的阈值，在所有动力学中均没有神经元被剪枝，这表明演化后的种群质量仍显著高于这些取值。当阈值为0.7时，约55%的神经元被剪枝，导致复制器和归一化动力学的准确率显著下降12.35%，而突变动力学的准确率下降10.62%。在更高阈值下，突变动力学展现出更强的稳健性，在进行激进剪枝时仍能保持更高的准确率。这些固定阈值结果补充了基于分位数的剪枝方法，揭示了演化种群质量的分布，并证实关键神经元与冗余神经元之间出现了有意义的分离。

演化框架为观察到的剪枝现象提供了自然的解释。与其将稀疏性视为有意优化的结果，不如将其理解为竞争动力学在训练过程中逐步演变的结果。这一视角与基于梯度的学习所具有的局部随机特性相吻合，也解释了剪枝为何呈现渐进且具有路径依赖性的行为。该框架还揭示了为何某些参数组得以保留，而另一些则逐渐衰减。适应度由对学习的贡献来定义，并决定相对增长率。持续有助于降低损失的群体会维持或增强其影响力，而冗余群体则会逐渐失去其代表性。这一机制无需明确协调或全局信息，便能从局部梯度信号中自然涌现。

演化框架为多种现有剪枝启发式方法提供了统一的解释。基于幅值的剪枝对应于适应度与参数范数成正比的情形，此时幅值较小的参数适应度较低。基于梯度的剪枝则与贡献项对 $|\partial\mathcal{L}/\partial p_i|$ 的依赖相吻合，倾向于保留梯度贡献较大的参数。

冗余感知剪枝源于相对适应度机制，该机制通过对高于平均水平的适应度进行竞争，隐式地惩罚相关参数。通过明确揭示这些联系，演化表述为理解这些启发式方法为何有效提供了有原则的基础，也为设计新的剪枝算法提供了指导。

我们的实验评估有意仅限于采用简单多层感知机架构的 MNIST。这样的选择使我们能够清晰地考察演化动力学，同时避免深层架构或复杂数据带来的混杂影响。然而，扩展到残差网络和 Transformer 等更深的网络，以及 ImageNet 和语言建模语料库等更大型的数据集，会带来额外挑战。当种群质量趋近于零时，有效权重矩阵可能会变得病态，从而引发数值稳定性方面的担忧。在非常深的网络中，各层的梯度幅值可能差异显著，因此需要采用分层学习率来应对适应度异质性。扩展到滤波器级或块级剪枝，需要谨慎设计用于结构化剪枝的种群分组和适应度信号。这些扩展是未来工作的重要方向，但超出了本初步研究的范围。

10 结论

我们提出了一种剪枝的演化框架，其中参数组如同种群，在选择压力下其影响力不断演变。通过采用复制器、归一化增长和选择–突变动力学，我们表明，训练后的种群质量为训练后剪枝提供了自然依据，并在 MNIST 上形成了可测量的准确率–稀疏度权衡。在基准准确率约为 98% 的情况下，我们的实验表明，随着稀疏性增加，这三种动力学下的准确率都会逐渐下降；在 35–50% 的剪枝水平下仍保持稳健。使用我们的超参数时，复制器动力学与归一化增长动力学产生了相同的结果，而在高稀疏性下，选择–突变动力学表现出略强的稳健性。

本研究支持这样一种观点：剪枝可以被理解为学习动力学涌现出的结果，而非外部强加的剪枝计划或平衡解。我们将参数组视为承受选择压力的种群，从过程层面解释稀疏性涌现，这一解释与基于梯度的训练所具有的去中心化、随机性特征自然契合。演化视角为理解现有剪枝启发式方法提供了统一框架，并为设计新算法提供了有理论依据的指导。

未来研究的一个重要方向，是将显式的冗余感知适应度信号融入演化框架。尽管当前的表述捕捉了由基于梯度的贡献驱动的选择，但在纯粹选择作用下，高度相关的种群可能形成中性平衡。引入对冗余敏感的适应度项，例如成对相关性惩罚项或互信息度量，可在保留演化解释的同时，提供一种有理论依据的机制来消除这类退化现象。其他拓展方向包括：采用随训练推进而降低的自适应突变率，以平衡探索与利用；引入兼顾准确率和效率指标的多目标适应度；将方法应用于包括残差网络和 Transformer 在内的更大型架构；对不同适应度景观下的灭绝率进行理论分析；通过种群级选择与神经架构搜索建立联系；在卷积网络中对滤波器和模块进行结构化剪枝；将方法拓展至Transformer架构中的注意力头剪枝；以及研究多个网络层之间的协同演化动力学。

参考文献

Han, S., Pool, J., Tran, J., and Dally, W. (2015). 学习权重和连接，以构建高效的神经网络。神经信息处理系统进展（*NeurIPS*）。

Frankle, J. and Carbin, M. (2019). 彩票假设：寻找稀疏且可训练的神经网络。国际表示学习会议（*ICLR*）。LeCun, Y., Denker, J. S., and Solla, S. A. (1990). Optimal Brain Damage（最优脑损伤）。神经信息处理系统进展（*NeurIPS*）。Hassibi, B. and Stork, D. G. (1993). 网络剪枝的二阶导数：Optimal Brain Surgeon（最优脑外科）。神经信息处理系统进展（*NeurIPS*）。Louizos, C., Welling, M., and Kingma, D. P. (2018). 通过 L_0正则化学习稀疏神经网络。国际表示学习会议（*ICLR*）。Zhu, M. and Gupta, S. (2017). 剪枝还是不剪枝：探索剪枝在模型压缩中的有效性。*arXiv*预印本 *arXiv:1710.01878*。Hanson, S. J. and Pratt, L. Y. (1988). 比较反向传播中用于最小化网络构建的偏置。神经信息处理系统进展（*NeurIPS*），第1卷。Mozer, M. C. and Smolensky, P. (1989). 利用相关性自动缩小网络规模。连接科学，1(1):3–16。Li, T., Wu, B., Yang, Y., Fan, Y., Zhang, Y., and Liu, W. (2019). 通过分解卷积滤波器压缩卷积神经网络。*IEEE/CVF*计算机视觉与模式识别会议论文集。Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. (2017). 剪枝滤波器以构建高效卷积网络。国际表示学习会议（*ICLR*）。He, Y., Kang, G., Dong, X., Fu, Y., and Yang, Y. (2018). 用于加速深度卷积神经网络的软滤波器剪枝。国际人工智能联合会议（*IJCAI*），第2234–2240页。Mocanu, D. C., Mocanu, E., Stone, P., Nguyen, P. H., Gibescu, M., and Liotta, A. (2018). 受网络科学启发，利用自适应稀疏连接实现人工神经网络的可扩展训练。《自然·通讯》，9(1)。Evci, U., Ioannou, Y. A., Keskin, C., and Dauphin, Y. (2022). 稀疏神经网络中的梯度流，以及彩票假设为何奏效。*AAAI*人工智能会议。Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. (2021). 彩票假设为何奏效？剪枝神经网络样本复杂度的理论视角。神经信息处理系统进展（*NeurIPS*）。Chen, T., Sui, Y., Chen, X., Zhang, A., and Wang, Z. (2021). 图神经网络的统一彩票假设。国际机器学习会议（*ICML*）。Frantar, E. and Alistarh, D. (2023). SparseGPT：大语言模型可以通过一次性剪枝实现精准压缩。国际机器学习会议，PMLR。Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. (2023). 一种简单有效的大语言模型剪枝方法。*arXiv*预印本 *arXiv:2306.11695*。Li, Y., Yu, Y., Zhang, Q., Liang, C., He, P., Chen, W., and Zhao, T. (2023). LoSparse：基于低秩和稀疏近似的大语言模型结构化压缩。国际机器学习会议（*ICML*），第202卷，第20336–20350页。PMLR。Saxe, A. M., Nelli, S., and Summerfield, C. (2019). 深度神经网络中语义发展的数学理论。美国国家科学院院刊，116(23):11537–11546。

Hofbauer, J. 和 Sigmund, K.（1998）。进化博弈与种群动力学。剑桥大学出版社。Taylor, P. D. 和 Jonker, L. B.（1978）。进化稳定策略与博弈动力学。数学与生物科学，40(1-2):145–156。Sandholm, W.H.（2010）。种群博弈与进化动力学。麻省理工学院出版社。Gunasekar, S.、Lee, J. D.、Soudry, D. 和 Srebro, N.（2018）。矩阵分解中的隐式正则化。神经信息处理系统进展（*NeurIPS*）。Rumelhart, D. E.、Hinton, G. E. 和 Williams, R. J.（1986）。通过反向传播误差学习表征。《自然》，323(6088):533–536。Bottou, L.（2010）。使用随机梯度下降进行大规模机器学习。*COMPSTAT*论文集。Li, H.、Xu, Z.、Taylor, G.、Studer, C. 和 Goldstein, T.（2018）。神经网络损失景观的可视化。神经信息处理系统进展（*NeurIPS*）。Wen, W.、Wu, C.、Wang, Y.、Chen, Y. 和 Li, H.（2016）。深度神经网络中的结构化稀疏性学习。神经信息处理系统进展（*NeurIPS*）。Liu, Z.、Li, J.、Shen, Z.、Huang, G.、Yan, S. 和 Zhang, C.（2017）。通过网络瘦身学习高效卷积网络。国际计算机视觉会议（*ICCV*）。LeCun, Y.、Bottou, L.、Bengio, Y. 和 Haffner, P.（1998）。应用于文档识别的基于梯度的学习。*IEEE*论文集，86(11):2278–2324。