arXiv:2302.02922v1 [cs.LG] 6 Feb 2023

联合边-模型稀疏学习对图神经网络具有可证明的高效性

Shuai Zhang
伦斯勒理工学院
zhangs29@rpi.edu

Meng Wang伦斯勒理工学院
wangm7@rpi.edu

Pin-Yu ChenIBM研究院
Pin-Yu.Chen@ibm.com

Sijia Liu
密歇根州立大学
liusiji5@msu.edu

Songtao LuIBM研究院
songtao@ibm.com

Miao LiuIBM研究院
miao.liu1@ibm.com

摘要

由于训练大规模图神经网络（GNN）的计算挑战巨大，人们采用了各种稀疏学习技术来降低内存和存储成本。例如，图稀疏化 通过采样子图来减少数据聚合量，而模型稀疏化则通过剪枝神经网络来减少可训练权重的数量。尽管这些方法在降低训练成本的同时保持测试准确率方面取得了实证上的成功，但针对图神经网络稀疏学习的理论泛化分析仍不明朗。据我们所知，本文首次从样本复杂度和收敛速率的角度，对实现零泛化误差的联合边-模型稀疏学习进行了理论刻画。本文通过分析证明，采样重要节点以及剪除幅值最低的神经元，均可在不损害测试准确率的前提下降低样本复杂度并加快收敛。尽管分析聚焦于具有数据结构约束的两层图神经网络，但其见解适用于更一般的设置，并通过合成数据集和实际引文数据集得到了验证。

1 引言

图神经网络（GNN）能够有效表示图结构数据，并应用于目标检测（Shi & Rajkumar, 2020；Yan et al., 2018）、推荐系统（Ying et al., 2018；Zheng et al., 2021）、关系学习（Schlichtkrull et al., 2018）和机器翻译（Wu et al., 2020；2016）。然而，直接在大规模图上训练图神经网络，例如科学引文网络（Hull & King, 1987；Hamilton 等，2017；Xu et al., 2018）、社交网络（Kipf 与 Welling，2017；Sandryhaila & Moura, 2014；Jackson, 2010）和符号网络（Riegel et al., 2020），在计算上会变得极具挑战，甚至无法实现。这是由于邻居特征的指数级聚合以及过高的模型复杂度所致。例如，在包含 232,965 个节点、平均度为 492 的 Reddit 数据（Tailor et al., 2020）上训练两层 GNN，所需计算资源可能是使用 ImageNet （ Canziani et al., 2016 ）训练 ResNet-50 的两倍。

加速图神经网络训练的方法可分为两种范式：(i) 对图拓扑进行稀疏化（Hamilton 等，2017；Chen 等，2018；Perozzi 等，2014；Zou 等，2019），以及 (ii) 对网络模型进行稀疏化（Chen 等，2021b；You 等，2022）。对图拓扑进行稀疏化，是指选择一个子图而非原始图，以减少邻域聚合的计算量。可以使用固定子图（例如保留图类型学（Hübler 等，2008）、图移位算子（Adhikari 等，2017；Chakeri 等，2016）或度分布（Leskovec & Faloutsos，2006；Voudigari 等，2016；Eden 等，2018）），也可以应用边采样算法（如边稀疏化（Hamilton 等，2017））或节点稀疏化（Chen 等，2018；Zou 等，2019），在每次迭代中选择不同的子图。对网络模型进行稀疏化，是指降低神经网络模型的复杂度，包括移除非线性激活（Wu 等，2019；He 等，2020）、量化神经元权重（Tailor 等，2020；Bahri 等，2021）和中间层的输出（Liu 等，2021）、对网络进行剪枝（Frankle & Carbin, 2019），或

应用知识蒸馏（Yang 等，2020；Hinton 等，2015；Yao 等，2020；Jaiswal 等，2021）。两种稀疏化框架可以结合使用，例如 Chen 等，2021b 和 You 等，2022 中的联合边采样与网络模型剪枝。

尽管在不牺牲测试准确率的情况下加速图神经网络训练已经取得了许多实证上的成功，但利用稀疏化技术训练图神经网络的理论评估仍在很大程度上未被探索。大多数理论分析都集中于采样图的表达能力（Hamilton 等，2017；Cong 等，2021；Chen et al., 2018；Zou 等，2019；Rong 等，2019）或剪枝网络（Malach 等，2020；Zhang 等，2021；da Cunha 等，2022）。然而，对于泛化分析，即学习到的模型在测试数据上的表现是否良好，相关研究仍然有限。现有的大多数泛化分析仅限于两层情形，即使对于最简单形式的前馈神经网络（神经网络）也是如此，例如可参见（Zhang 等，2020a；Oymak 和 Soltanolkotabi，2020；Huang 等，2021；Shi et al., 2022）。据我们所知，只有 Li et al. (2022)；Allen-Zhu et al. (2019a) 分别通过研究三层图神经网络和神经网络，超越了两层的范围。然而，Li et al. (2022) 需要一个无法通过实证或理论加以论证的强假设，即采样图确实呈现了从数据到标签的映射。此外，Li et al. (2022)；Allen-Zhu et al. (2019a) 都关注于初始化附近的线性化模型，而学习到的权重仅保持在初始化附近（Allen-Zhu & Li, 2022）。线性化模型无法说明使用多层（图）神经网络和网络剪枝的优势。据我们所知，即使对于两层图神经网络，也不存在针对联合稀疏化的有限样本泛化分析。

贡献。本文首次对训练图神经网络时联合进行拓扑稀疏化与模型稀疏化的泛化性能进行了理论分析，包括：（1）明确给出准确预测未知标签所需的已知标签数量（称为样本复杂度）的界，以及随机梯度下降（SGD）的收敛速率；（2）从样本复杂度和收敛速率两个角度，通过定量证明说明，联合进行拓扑稀疏化与模型稀疏化是一种能够提升学习性能的双赢策略。

我们考虑以下问题设置，以建立理论分析：在一个
单隐藏层图神经网络，假设部分节点特征与类别相关（Shi et al., 2022），并决定标签；而部分节点特征与类别无关，仅包含对标注无关的信息，且节点的标签会受到其邻居的类别相关特征影响。带有这一结构约束的数据模型刻画了某些节点比其他节点更具影响力的现象，例如社交网络中的情况（Chen et al., 2018;Veli˘ckovi´c et al., 2018），或图中包含冗余信息的情况（Zheng et al., 2020）。

具体而言，样本复杂度与 $(1-\beta)/\alpha$ 成二次关系，其中 α in $(0,1]$ 是采样到具有类别相关特征的节点的概率，α 越大，表示类别相关特征被采样得越频繁。 β in $[0,1)$ 是使用基于幅值的剪枝方法（如 Frankle & Carbin, 2019）对网络模型进行剪枝时，剪枝神经元所占的比例。达到理想模型所需的随机梯度下降迭代次数与 $(1-\beta)/\alpha$ 成线性关系。因此，我们的结果从形式上证明，如果具有类别相关特征的节点被采样得更频繁，图采样就能更显著地降低样本复杂度和迭代次数。其直观原因是，重要性采样有助于算法更高效地学习类别相关特征，从而减少所需样本量和收敛时间。只要 β 不超过接近 1 的某个阈值，随着剪枝率提高，也能观察到同样的学习效果提升。

2 图神经网络：形式化与算法

2.1 问题形式化

给定一个无向图 $\mathcal{G}(\mathcal{V},\ \mathcal{E})$，其中 $\mathcal{V}$ 是节点集合，$\mathcal{E}$ 是边集合。令 R 表示最大节点度。对于任意节点 $v \in \mathcal{V}$，令 $\boldsymbol{x}_v \in \mathbb{R}^d$ 和 $y_v \in \{+1,-1\}$ 分别表示其输入特征和对应标签[1]，。给定所有节点特征 $\{\boldsymbol{x}_v\}_{v\in\mathcal{V}}$ 以及部分已知的标签$\{y_v\}_{v\in\mathcal{D}}$，这些标签对应于节点集合 $\mathcal{D}\subset\mathcal{V}$中的节点，半监督

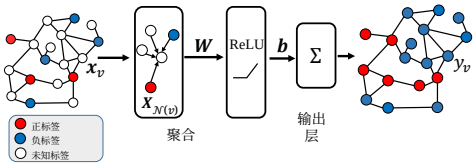


图1：图神经网络中的节点分类示意图

节点分类问题旨在预测 $v\in\mathcal{V}/\mathcal{D}$ 的所有未知标签 y_v。

[1]该分析可以扩展到多类分类，参见附录 I。

本文考虑一种具有非线性聚合函数的图神经网络，如图1所示。隐藏层中 K 个神经元的权重记为 $\{\boldsymbol{w}_k \in \mathbb{R}^d\}_{k=1}^K$，线性层中的权重记为 $\{b_k \in \mathbb{R}\}_{k=1}^K$。令 $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ 和 $\boldsymbol{b} \in \mathbb{R}^K$ 分别为 $\{\boldsymbol{w}_k\}_{k=1}^K$ 和 $\{b_k\}_{k=1}^K$ 的拼接。对于任意节点 v，令 $\mathcal{N}(v)$ 表示其（1跳）邻居集合（包含自连接），而 $\boldsymbol{X}_{\mathcal{N}(v)} \in \mathbb{R}^{d\times|\mathcal{N}(v)|}$ 包含 $\mathcal{N}(v)$ 中的特征。

因此，节点 v 的图神经网络输出可写为：

$$g(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{X}_{\mathcal{N}(v)}) = \frac{1}{K}\sum_{k=1}^{K} b_k \cdot \mathrm{AGG}(\boldsymbol{X}_{\mathcal{N}(v)}, \boldsymbol{w}_k), \tag{1}$$

其中，$\mathrm{AGG}(\boldsymbol{X}_{\mathcal{N}(v)}\ \boldsymbol{w})$ 表示使用特征 $\boldsymbol{X}_{\mathcal{N}(v)}$ 和权重 $\boldsymbol{w}$ 的通用聚合器，例如邻居节点的加权和（Veli˘ckovi´c 等，2018年）、最大池化（Hamilton 等，2017年）或最小池化（Corso 等，2020），并结合某种非线性激活函数。我们将 ReLU 视为 $\phi(\cdot) = \max\{\cdot, 0\}$。给定图神经网络，节点 v 的标签 y_v 由 $\mathrm{sign}(g(\boldsymbol{W}, \boldsymbol{b};\ \boldsymbol{X}_{\mathcal{N}(v)}))$ 预测。

由于 ReLU函数的齐次性，我们仅更新 $\boldsymbol{W}$，这是简化分析的常见做法（Allen-Zhu 等，2019a；Arora 等，2019；Oymak 和 Soltanolkotabi，2020；Huang 等，2021）。训练问题最小化以下经验风险函数（ERF）：

$$\min_{\boldsymbol{W}}: \quad \hat{f}_{\mathcal{D}}(\boldsymbol{W}, \boldsymbol{b}^{(0)}) := -\frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}} y_v \cdot g\big(\boldsymbol{W}, \boldsymbol{b}^{(0)}; \boldsymbol{X}_{\mathcal{N}(v)}\big). \tag{2}$$

测试误差通过以下泛化误差函数进行评估：

$$I(g(\boldsymbol{W}, \boldsymbol{b})) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \max\big\{1 - y_v \cdot g\big(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{X}_{\mathcal{N}(v)}\big), 0\big\}. \tag{3}$$

如果 $I(g(\boldsymbol{W},\ \boldsymbol{b})) = 0,\ \ y_v = \mathrm{sign}(g(\boldsymbol{W},\ \boldsymbol{b};\ \boldsymbol{X}_{\mathcal{N}(v)}))$ 对所有 v 均成立，则表示零测试误差。

尽管不同于实际应用中的图神经网络，本论文所研究的模型可以视为一种单隐藏层图神经网络，这是结构化数据泛化与收敛分析中的最先进实践（Brutzkus & Globerson, 2021；Damian et al., 2022；Shi et al., 2022；Allen-Zhu & Li, 2022）。此外，由于 ReLU 函数的非线性，式（2）的优化问题已经具有高度的非凸性 。例如，正如（Liang et al., 2018；Safran & Shamir, 2018）所指出的，单隐藏层（G）神经网络包含数量难以处理的伪局部极小值。此外，已有证明表明，本论文所考虑的图神经网络模型及数据分布的 VC 维数至少是数据维度的指数函数（证明见附录 G）。该模型具有很强的表达能力，而要获得多项式样本复杂度极其不容易；这是本论文的贡献之一。

2.2 基于边与模型联合稀疏化的 GNN 学习算法

通过小批量随机梯度下降算法求解 GNN 学习问题 (2)，具体如算法1所示。系数 b_k's 从 $+1$ 或 -1 中随机选取，并在训练过程中保持不变。隐藏层中的权重 $\boldsymbol{w}_k$ 初始化为多元高斯分布 $\mathcal{N}(\boldsymbol{0}, \delta^2\boldsymbol{I}_d)$，其方差为较小的常数 δ，例如 $\delta = 0.1$。训练数据被划分为互不相交的子集，每次迭代使用其中一个子集，通过随机梯度下降更新 $\boldsymbol{W}$。算法1包含两个训练阶段：在 $\boldsymbol{W}$ 上进行预训练（第 1–4 行），迭代次数较少；以及在剪枝模型 $\boldsymbol{M} \odot \boldsymbol{W}$ 上重新训练（第 5–8 行），其中 $\odot$ 表示逐元素乘法。这里使用逐神经元幅值剪枝得到权重掩码 $\boldsymbol{M}$，并通过节点采样实现图拓扑稀疏化（第 8 行）。每次迭代中，在每个迭代步骤里，仅将部分邻居节点输入公式 (1) 中的聚合函数，其中 $\mathcal{N}^{(s)}(t)$ 表示迭代 t 时节点 v 的邻居采样子集。

边稀疏化对每个节点采样一部分邻居，而不是全部 $\mathcal{N}(v)$，以降低计算 (1) 时每次迭代的计算复杂度。本文采用 GraphSAGE 框架（Hamilton 等，2017），在该框架中，每次迭代都会为每个节点采样 r （$r \ll R$）个邻居（如果 $|\mathcal{N}(v)| \le r$，则会采样所有这些邻居）。在迭代 t 时，梯度为

$$\nabla\hat{f}_{\mathcal{D}}^{(t)}(\boldsymbol{W}, \boldsymbol{b}^{(0)}) = -\frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}} y_v \cdot \nabla_{\boldsymbol{W}} g\big(\boldsymbol{W}, \boldsymbol{b}^{(0)}; \boldsymbol{X}_{\mathcal{N}_s^{(t)}(v)}\big). \tag{4}$$

算法1通过联合边与模型稀疏化训练图神经网络

输入：节点特征 $\boldsymbol{X}$，已知节点标签$\{y_v\}_{v\in\mathcal{D}}$及 $\mathcal{D}\subseteq\mathcal{V}$，步长 $c_\eta=10\delta$及常数 δ，采样边的数量 r，剪枝率 β，预训练迭代次数 $T'=\|\boldsymbol{X}\|_\infty/c_\eta$，迭代次数 T。初始化： $\boldsymbol{W}(0)\,\boldsymbol{b}^{(0)}$为 $\boldsymbol{w}_k^{(0)}\sim\mathcal{N}(0,\ \delta^2\boldsymbol{I}_d)$，且对于 $k\in[K]$， $b_k^{(0)}\sim$服从 Uniform($\{-1,+1\}$)；预训练：通过带边采样的小批量随机梯度下降更新模型权重 $\boldsymbol{W}$。1：将 $\mathcal{D}$ 划分为互不相交的子集 $\{\mathcal{D}^{(t')}\}_{t'=1}^{T'}$ 2：对于 $t'=0,1,2,\cdots,T'-1$，执行 3：为每个节点 v 从 $\mathcal{D}^{(t')}$ 中采样 $\mathcal{N}^{(t')}{}_s\,(v)$；4：$\boldsymbol{W}(_{t'+1})=\boldsymbol{W}(_{t'})-c_\eta\cdot\nabla\boldsymbol{w}\hat{}f_{\mathcal{D}}^{(t)}(\boldsymbol{W}^{(t')}\boldsymbol{b}^{(0)})$；5：结束循环。剪枝：将 $\boldsymbol{W}^{(T')}$ 中权重幅值最低的 β 比例神经元置为 0，并得到相应的二值掩码 $\boldsymbol{M}$；重新训练：将权重回退至原始初始化 $\boldsymbol{M}\odot\boldsymbol{W}(0)$，并通过带边采样的随机梯度下降更新模型权重；6：将 $\mathcal{D}$ 划分为互不相交的子集 $\{\mathcal{D}^{(t)}\}_{t=1}^{T}$7：对于 $t=0,1,2,\cdots,T$，执行 8：为每个节点 v 从 $\mathcal{D}^{(t)}$ 中采样 $\mathcal{N}_s^{(t)}(v)$；9：$\boldsymbol{W}(_{t+1})=\boldsymbol{W}(_t)-c_\eta\cdot\nabla\boldsymbol{w}\hat{}f_{\mathcal{D}\ t}^{(t)}(\,)(\boldsymbol{M}\odot\boldsymbol{W}^{(t)},\boldsymbol{b}^{(0)})$；10：结束循环。返回：$\boldsymbol{W}^{(T)}$ 和 $\boldsymbol{b}^{(0)}$。

其中， $\mathcal{D}\subseteq\mathcal{V}$ 是带有标签的训练节点子集。本文使用的聚合函数是最大池化函数，即，

$$\mathrm{AGG}(\boldsymbol{X}_{\mathcal{N}_s^{(t)}(v)},\boldsymbol{w})=\max_{n\in\mathcal{N}_s^{(t)}(v)}\phi(\langle\,\boldsymbol{w}\,,\,\boldsymbol{x}_n\,\rangle),\tag{5}$$

该方法已广泛应用于 GraphSAGE（Hamilton 等，2017）及其变体（Guo et al., 2021；Oh et al., 2019；Zhang et al., 2022b；Lo et al., 2022）。本文采用一种重要性采样策略，其理念是某些节点比其他节点具有更高的采样概率，类似于（Chen et al., 2018；Zou 等，2019；Chen 等，2021b）中的采样策略。

模型稀疏化 首先对神经网络进行预训练（通常只需迭代几次），然后将部分神经元权重置零，以剪枝网络。接着，使用更少的参数对剪枝模型重新训练，其训练计算开销也更低。现有的剪枝方法包括神经元剪枝和权重剪枝。前者同时将神经元 $\boldsymbol{w}_k$ 的所有条目置零，后者则独立地将 $\boldsymbol{w}_k$ 的条目置零。

本文研究神经元剪枝。与（Chen 等，2021b）类似，我们首先训练原始图神经网络，直到算法收敛。然后，通过移除范数最小的 $\beta(\beta\in[0,1))$ 比例的神经元，对神经元进行幅值剪枝。设 $\boldsymbol{M}\in\{0,1\}^{d\times K}$为二值掩码矩阵；若移除神经元 k，则矩阵的第 k 列全为零。接着，我们将剩余的图神经网络恢复到原始初始化（即 $\boldsymbol{M}\odot\boldsymbol{W}^{(0)}$），并在模型 $\boldsymbol{M}\odot\boldsymbol{W}$上重新训练。

3 理论分析

3.1 理论发现概要

在正式介绍数据模型和理论结果之前，我们先简要概述这些结果的关键要点。我们考虑这样一种一般设定：节点特征由一些类别相关特征和类别无关特征的含噪实现构成，且 δ 表示加性噪声的上界。节点 v 的标签 y_v 由 $\mathcal{N}$（v）中的类别相关特征决定。为简化起见，我们假设 $\mathcal{N}$（v）恰好包含一个类别的类别相关特征。表 1 总结了若干主要参数。主要结论包括：

（T1）零泛化误差的样本复杂度与收敛分析。我们证明，对于初始化和随机梯度下降步骤中的随机性，学习到的模型（无论是否经过稀疏化）都能以高概率实现零泛化误差。样本复杂度与 σ^2 和 K^{-1} 呈线性关系。迭代次数与 σ 和 $K^{-1/2}$ 呈线性关系。因此，学习

表 1：一些重要记号

K	隐藏层中的神经元数量；	σ	输入特征中加性噪声的上界；
r	每个节点的采样边数量；	R	原始图的最大度 $\mathcal{G}$；
L	类别相关模式和类别无关模式的数量；		
α	一个节点的采样邻居中包含类别相关节点的概率；		
β	模型权重的剪枝率； $\beta \in [0, 1-1/L)$ ； $\beta = 0$ 表示不进行剪枝；		

如果神经网络稍微过参数化，那么学习性能会得到提升，具体表现为样本复杂度更低、收敛更快。

(T2) 边稀疏化和重要性采样能够提升学习性能。 样本复杂度是 r的二次函数，这表明边稀疏化可以降低样本复杂度。其直觉在于，边稀疏化降低了类别相关特征与类别无关特征的聚合程度，使得针对所考虑的数据模型学习类别相关模式变得更加容易，从而提升学习性能。样本复杂度和迭代次数分别关于 α^{-1}呈二次和线性关系。由于较大的 α 意味着以更高概率采样类别相关特征，因此这一结果符合这样的直觉：成功的重要性采样策略有助于更快地学习类别相关特征，并减少所需样本数。

(T3) 基于幅值的模型剪枝能够提升学习性能。 样本复杂度和计算时间均与$(1-\beta)^2$, 呈线性关系，这表明，剪除更多幅值较小的神经元可以降低样本复杂度和计算时间。其直觉在于，能够准确学习类别相关特征的神经元往往比其他神经元具有更大的幅值；移除其他神经元则能提高学习效率。

（T4）边和模型稀疏化是GNN学习中的双赢策略。我们的定理为联合边-模型稀疏化的成功提供了理论验证。样本复杂度和迭代次数分别关于 $\frac{1-\beta}{\alpha}$ 呈二次和线性关系，这表明两种技术可以结合使用，从而有效提升学习性能。

3.2 正式理论结果

数据模型。令 $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{L}$ $(\forall L \leq d)$ 表示 $\mathbb{R}^d$ 中任意的正交向量集合[2] 。令$\boldsymbol{p}_+ := \boldsymbol{p}_1$ 和 $\boldsymbol{p}_- := \boldsymbol{p}_2$ 分别表示正类模式和负类模式，它们与标签具有
因果关系。 $\mathcal{P}$ 中的其余向量是类别无关模式。每个节点 v 的节点特征 $\boldsymbol{x}_v$ 是这些模式之一的带噪版本，即 $\boldsymbol{x}_v = \boldsymbol{p}_v + \boldsymbol{z}_v$，其中 $\boldsymbol{p}_v \in \mathcal{P}$，而 $\boldsymbol{z}_v$ 是节点 v 处的任意噪声，并且对于某个 σ 满足 $\|\boldsymbol{z}\|_2 \leq \sigma$ 。

y_v 当节点 v 或其任意邻居包含 p_+ （或 p_-）时，称其为 $+1$ （或 -1）。具体而言，根据节点特征是否相关（下标中的 N）以及标签，将 $\mathcal{V}$ 划分为四个互不相交的集合 $\mathcal{V}_+$、 $\mathcal{V}_-$、 $\mathcal{V}_{N+}$ 和 $\mathcal{V}_{N-}$，即 $\boldsymbol{p}_v = \boldsymbol{p}_+, \ \forall v \in \mathcal{V}_+$; $\boldsymbol{p}_v = \boldsymbol{p}_-, \ \forall v \in \mathcal{V}_-$; 以及 $\boldsymbol{p}_v \in \{\boldsymbol{p}_3, ..., \boldsymbol{p}_L\}, \ \forall v \in \mathcal{V}_{N+} \cup \mathcal{V}_{N-}$。于是，我们有 $y_v = 1, \ \forall v \in \mathcal{V}_+ \cup \mathcal{V}_{N+}$，且 $y_v = -1, \ \forall v \in \mathcal{V}_- \cup \mathcal{V}_{N-}$。

我们假设

（**A1**） $\mathcal{V}_{N+}$ （或 $\mathcal{V}_{N-}$）中的每个 v 都至少与 $\mathcal{V}_+$（或 $\mathcal{V}_-$）中的一个节点相连。 $\mathcal{V}_+$ 与$\mathcal{V}_{N-}$之间没有边。 $\mathcal{V}_-$ 与 $\mathcal{V}_{N+}$之间没有边。

(A2) $\mathcal{D}$ 中的正负标签是平衡的，即$\left||\mathcal{D} \cap (\mathcal{V}_+ \cup \mathcal{V}_{N+})| - |\mathcal{D} \cap (\mathcal{V}_- \cup \mathcal{V}_{N-})|\right| = O(\sqrt{|\mathcal{D}|})$。

(A1) 表示图中相连的节点往往具有相同的标签，并排除了节点 v 同时与 p_+ 和 p_- 相连的情况，以简化分析。对此类假设在 Cora 中的数值论证

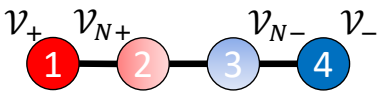


图 2：数据模型的玩具示例。节点 1 和 2 的标签为 $+1$。节点 3 和 4 的标签为 -1。节点 1 和 4 具有与类别相关的特征。节点 2 和 3 具有与类别无关的特征。 $\mathcal{V}_+ = \{1\}$, $\mathcal{V}_{N+} = \{2\}$,$\mathcal{V}_{N-} = \{3\}$, $\mathcal{V}_- = \{4\}$

数据集见附录 F.2。（**A2**）可放宽至观测标签不平衡的情况。只需相应地提高 ERF 中少数类的权重（见式（2））， 这是类别不平衡图神经网络学习中的常见技巧（Chen 等， 2021a）， 我们的分析只需稍作

[2]正交性约束简化了分析，并已在 (Brutzkus & Globerson, 2021) 中采用。我们在第4节的合成数据实验中放宽了这一约束。

修改即可。正交模式的数据模型由 Brutzkus & Globerson（2021）提出，用于分析卷积神经网络相较于全连接神经网络的优势。该模型通过消除类别相关模式与类别无关模式之间的相互作用，简化了分析。在此，我们将其推广到节点特征包含额外噪声且不再正交的情况。

为定量分析重要性采样的影响，令 α 表示如下概率的下界：对于任意节点 v，v 的采样邻居中至少包含一个属于 $\mathcal{V}_+$ 或 $\mathcal{V}_-$ 的节点（见表 1）。显然，$\alpha = r/R$ 是均匀采样 [3]的下界。较大的 α 表明该采样策略确实更频繁地选取具有类别相关特征的节点。

定理 1 给出了算法1通过图稀疏化学习图结构数据时的样本复杂度（**C1**）和收敛速率（**C2**）。具体而言，在样本数量（**C1**）和迭代次数（**C2**）足够多时，返回的模型可实现零泛化误差（由式 **(8)** 得出）。

定理 1. 令步长 c_η 为某个正数常量，剪枝率 $\beta \in [0, 1 - 1/L)$。假设有界噪声满足 $\sigma < 1/L$，且模型规模足够大，满足 $K > L^2 \cdot \log q$，其中 $q > 0$为某个常量。则以至少 $1 - q^{-10}$的概率，当

(C1) 带标签节点的数量满足

$$|\mathcal{D}| = \Omega\big((1 + L^2\sigma^2 + K^{-1}) \cdot \alpha^{-2} \cdot (1 + r^2) \cdot (1 - \beta)^2 \cdot L^2 \cdot \log q\big), \tag{6}$$

(C2) 重新训练阶段的迭代次数满足

$$T = \Omega\big(c_\eta^{-1}(1 + |\mathcal{D}|^{-1/2}) \cdot (1 + L\sigma + K^{-1/2}) \cdot (1 - \beta) \cdot \alpha^{-1} \cdot L\big), \tag{7}$$

算法*1*返回的模型实现了零泛化误差，即，

$$I\big(g(\boldsymbol{W}^{(T)}, \boldsymbol{U}^{(T)})\big) = 0. \tag{8}$$

3.3 技术贡献

尽管我们的数据模型受到用于分析卷积神经网络的特征学习框架（Brutzkus & Globerson, 2021）的启发，但分析学习动态的技术框架在以下方面不同于现有框架。

首先，我们的工作给出了式（6）中的首个多项式阶样本复杂度界，该界定量刻画了参数依赖关系，并实现了零泛化误差。在 Brutzkus & Globerson, 2021 中，泛化界是通过更新第二层（线性层）的权重、同时固定非线性层的权重得到的。然而，神经网络的强大表达能力主要来自非线性层的权重。更新隐藏层的权重比更新输出层的权重能够实现更小的泛化误差。因此，本文通过刻画非线性层（隐藏层）中的权重更新，得到了一个多项式阶样本复杂度界，而这无法从 Brutzkus & Globerson, 2021 推导出来。此外，更新隐藏层权重是一个非凸问题，相较于将更新输出层权重视为凸问题的情形，更具挑战性。

其次，本文的理论框架能够刻画基于幅值的剪枝方法，而（Brutzkus & Globerson, 2021）中的方法则无法做到这一点。具体而言，我们的分析给出了更紧的界，使得“幸运神经元”的下界可能远大于“不幸神经元”的上界（见引理2–5）。这构成了刻画模型剪枝优势的理论基础，而（Brutzkus & Globerson, 2021）并未提供这一基础（见引理5.3和5.5）。一方面，（Brutzkus & Globerson, 2021）仅给出了适用于所有方向的“不幸神经元”统一界，而本文的引理4则给出了不同方向上的具体界。另一方面，本文考虑了样本数量的影响，并需要刻画正类和负类之间的梯度偏移。由于类别无关模式的存在，以及边采样会破坏标签与模式分布之间的依赖关系，这一问题颇具挑战性，并会导致意料之外的分布偏移。我们将一些特殊的数据组作为参照，使其与标签之间保持固定的依赖关系，并且相对于采样数据具有可控的分布偏移。

第三，本文的理论框架能够刻画边采样，而（Brutzkus & Globerson, 2021）中的方法则不能。（Brutzkus & Globerson, 2021）要求数据样本

[3]某些采样策略的 α 下界见附录 G.1。

通过间隔泛化界确保训练样本中包含类别相关模式（Shalev- Shwartz & Ben-David, 2014; Bartlett & Mendelson, 2002）。然而，在进行数据采样时，采样后的数据可能不再包含类别相关模式。因此，更新第二层并不稳健，但我们的理论结果表明，更新隐藏层能够抵御边采样造成的离群点。

3.4 证明概要

在介绍证明的正式路线图之前，我们借鉴 (Brutzkus & Globerson, 2021) 中“幸运神经元”的概念，对证明作一个高层次的说明，其中的节点具有良好的初始权重。我们强调，我们借鉴的仅仅是这一概念；“幸运神经元”的所有性质，例如 (10) 至 (13)，都是独立推导得出的，并排除了其他论文中的理论发现。本文证明，“幸运神经元”的幅值以类别相关特征的采样率为速率增长，而“不幸神经元”的幅值则以上界为训练数据规模的倒数（见命题2）。当训练数据足够多时，“幸运神经元”具有较大的幅值，并主导输出值。通过剪枝去除幅值较小的神经元，我们可以保留“幸运神经元”，并有可能移除“不幸神经元”（见命题 3）。此外，我们证明，“幸运神经元”的主方向与类别相关模式一致，且“幸运神经元”的比例足够大（见命题 1）。因此，输出由“幸运神经元”的主方向决定，而该方向就是相应的类别相关模式。具体而言，我们将证明，对于每个具有 $y_v = 1$ 的节点 v，学习到的权重 $\boldsymbol{W}_{(T)}$ 所作出的预测是准确的，即 $g(\boldsymbol{M} \odot \boldsymbol{W}_{(T)}\, \boldsymbol{b}^{(0)};\ \boldsymbol{X}_{\mathcal{N}(v)}) > 1$。对于具有负标签的节点，论证过程相同。随后，根据式 (3) 中定义的泛化误差，可得零测试误差。

将神经元分为两个子集 $\mathcal{B}_+ = \{k \mid b_k^{(0)} = +1\}$ 和 $\mathcal{B}_- = \{k \mid b_k^{(0)} = -1\}$。我们首先证明，存在一些神经元 i 属于 $\mathcal{B}_+$，其权重 $\boldsymbol{w}_i^{(t)}$ 在所有迭代次数$t \geq 0$中都接近 $\boldsymbol{p}_+$。这些神经元被称为“幸运神经元”，在对 v进行分类时起主导作用，且这类神经元所占比例至少接近 $1/L$。形式化地，

命题 1. 设 $\mathcal{K}_+ \subseteq \mathcal{B}_+$ 表示“幸运神经元”的集合，对于任意 i属于 $\mathcal{K}_+$,

$$\min_{\|\boldsymbol{z}\|_2 \leq \sigma} \langle\, \boldsymbol{w}_i^{(t)}\,,\, \boldsymbol{p}_+ + \boldsymbol{z}\,\rangle \geq \max_{\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_+, \|\boldsymbol{z}\|_2 \leq \sigma} \langle\, \boldsymbol{w}_i^{(t)}\,,\, \boldsymbol{p} + \boldsymbol{z}\,\rangle \quad \textit{for all} \quad t. \tag{9}$$

那么有

$$|\mathcal{K}_+|/K \geq (1 - K^{-1/2} - L\sigma)/L. \tag{10}$$

接下来，我们将在命题2中证明，当 $|\mathcal{D}|$ 足够大时，幸运神经元 i 的权重 $\boldsymbol{w}_i^{(t)}$ 在 $\boldsymbol{p}_+$ 上的投影以 $c_\eta \alpha$ 的速率增长。因此，较大的 α 对应较高的速率。在所有方向上，$\mathcal{B}_-$ 中的神经元增长都要慢得多，唯独在 $\boldsymbol{p}_-$方向上例外。

命题 2.

$$\begin{aligned} &\langle\, \boldsymbol{w}_i^{(t)}\,,\, \boldsymbol{p}_+ \,\rangle \geq c_\eta\big(\alpha - \sigma\sqrt{(1+r^2)/|\mathcal{D}|}\big)t,\ \ \forall i \in \mathcal{K}_+, \forall t \\ &|\langle\, \boldsymbol{w}_j^{(t)}\,,\, \boldsymbol{p}\,\rangle| \leq c_\eta(1+\sigma)\sqrt{(1+r^2)/|\mathcal{D}|}\cdot t,\ \ \forall j \in \mathcal{B}_-, \forall \boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_-, \forall t. \end{aligned} \tag{11}$$

命题 3 表明，$\mathcal{K}_+$ 中“幸运神经元”的权重幅值大于 $\mathcal{B}_+/\mathcal{K}_+$中神经元的权重幅值。结合式 (10)，只要满足 $\beta < 1 - 1/L$，“幸运神经元”就不会被幅值剪枝。令 $\mathcal{K}_\beta$ 表示在 $|\mathcal{K}_\beta| = (1-\beta)K$ 后得到的神经元集合。

命题 3。 存在一个小的正整数 C，使得

$$\|\boldsymbol{w}_i^{(t)}\|_2 > \|\boldsymbol{w}_j^{(t)}\|_2,\ \ \forall i \in \mathcal{K}_+, \forall j \in \mathcal{B}_+/\mathcal{K}_+, \forall t \geq C. \tag{12}$$

此外，对于所有 $\beta \leq 1 - 1/L$，都有 $\mathcal{K}_\beta \cap \mathcal{K}_+ = \mathcal{K}_+$。

因此，当样本数量足够大时，幸运神经元的幅值会比其他神经元增长得快得多（根据命题 2）。当幸运神经元的比例足够大时（根据命题 1），学习到的模型的输出将严格为正。此外，采用适当的剪枝率可以进一步提高幸运神经元的比例（根据命题 3），从而降低样本复杂度并加快收敛速率。

最后，我们考虑无特征噪声的情况，以说明主要计算过程

$$
\begin{aligned}
& g(\boldsymbol{M} \odot \boldsymbol{W}^{(T)}, \boldsymbol{M} \odot \boldsymbol{b}^{(0)}; \boldsymbol{X}_{\mathcal{N}(v)}) \\
&= \tfrac{1}{K}\big(\textstyle\sum_{i \in \mathcal{B}_+ \cap \mathcal{K}_\beta} \max_{u \in \mathcal{N}(v)} \phi(\langle\, \boldsymbol{w}_i^{(T)}\,,\, \boldsymbol{x}_u \,\rangle) - \sum_{j \in \mathcal{B}_- \cap \mathcal{K}_\beta} \max_{u \in \mathcal{N}(v)} \phi(\langle\, \boldsymbol{w}_j^{(T)}\,,\, \boldsymbol{x}_u \,\rangle)\big) \\
&\geq \tfrac{1}{K} \sum_{i \in \mathcal{K}_+} \max_{u \in \mathcal{N}(v)} \phi(\langle\, \boldsymbol{w}_i^{(T)}\,,\, \boldsymbol{x}_u \,\rangle) - \tfrac{1}{K} \sum_{j \in \mathcal{B}_- \cap \mathcal{K}_\beta} \max_{\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_-} |\langle\, \boldsymbol{w}_j^{(T)}\,,\, \boldsymbol{p} \,\rangle| \\
&\geq \big[\alpha|\mathcal{K}_+|/K - (1-\beta)(1+\sigma)\sqrt{(1+r^2)/|\mathcal{D}|}\big] c_\eta T \quad > 1,
\end{aligned}
\tag{13}
$$

其中，第一个不等式由以下事实得出： $\mathcal{K}_\beta \cap \mathcal{K}_+ = \mathcal{K}_+$、 ϕ 是非负 ReLU函数，且节点 v （v）不包含 $\boldsymbol{p}_-$ ，其中 $y_v = +1$。第二个不等式由命题2得出。最后一个不等式由式（10）以及结论（C1）和（C2）得出。证明至此完成。详情请参阅补充材料。

4 数值实验

4.1 合成数据实验

我们生成一个包含 10000 个节点的图，节点度为 30。独热向量 $\boldsymbol{e}_1$ 和 $\boldsymbol{e}_2$ 分别选作 $\boldsymbol{p}_+$ 和 $\boldsymbol{p}_-$。类别无关模式从 $\boldsymbol{p}_+$ 和 $\boldsymbol{p}_-$的零空间中随机选取。这是对数据模型中正交性约束的放宽。所有模式的 $\|\boldsymbol{p}\|_2$ 均归一化为 1。噪声 $\boldsymbol{z}_v$ 服从高斯 $\mathcal{N}(0, \sigma^2)$。节点特征和标签满足 (A1) 和 (A2)，构造细节见附录 F。测试误差是未知标签预测错误所占的百分比。如果返回的模型实现了零测试误差，则认为学习过程成功 。

样本复杂度。 我们首先验证式（6）中的样本复杂度界。每个结果均取100次独立试验的均值。白色方块表示所有试验均成功，黑色方块表示全部失败。在这些实验中，我们改变一个参数并固定其他所有参数。在图3中， $r = 15$ ，我们改变重要性采样概率 α。样本复杂度与α^{-2}呈线性关系。图4表明，在达到某个上界之前，样本复杂度与 $(1-\beta)^2$ 几乎呈线性关系，其中 β 为剪枝率。所有这些结果均与式（6）中的理论预测一致。

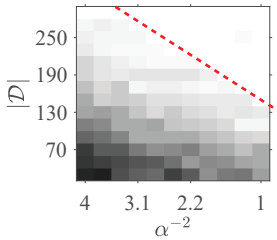


图3： $|\mathcal{D}|$ 与重要性采样概率 α的关系

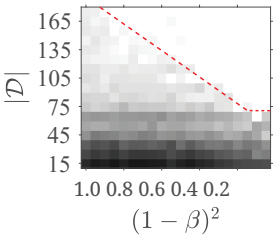


图4： $|\mathcal{D}|$ 与剪枝率 β的关系

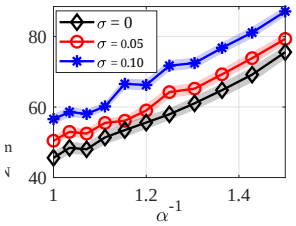


图5：迭代次数与 α的关系

训练收敛速率。 接下来，我们评估采样和剪枝如何减少达到零泛化误差所需的迭代次数。图5展示了不同 α 在不同噪声水平 σ下所需的迭代次数。每个点均取1000次独立实现的均值，低透明度区域表示标准差为一个标准差的误差条。我们可以看到，迭代次数与 $1/\alpha$呈线性关系，这验证了式（6）中的理论发现。因此，重要性采样减少了收敛所需的迭代次数。图6展示了不同剪枝率下收敛所需的迭代次数。基线是训练稠密网络的平均迭代次数。幅值剪枝所需的迭代次数与 β几乎呈线性关系，这验证了式（7）中的理论发现。相比之下，随机剪枝的性能有所下降，因为其达到收敛所需的迭代次数多于基线。

幅值剪枝会移除包含无关信息的神经元。图7展示了算法收敛后神经元权重的分布。通过收集100次独立试验中的神经元，得到 10^4 个点。纵轴表示神经元权重的范数 $\boldsymbol{w}_k$，纵轴表示神经元权重与 $\boldsymbol{p}_+$ （底部）或 $\boldsymbol{p}_-$ （顶部）之间的角度。十字形的蓝色点表示具有 $b_k = 1$的 $\boldsymbol{w}_k$，圆形的红色点表示具有 $b_k = -1$的 $\boldsymbol{w}_k$。在这两种情况下，范数较小的 $\boldsymbol{w}_k$ 确实与 $\boldsymbol{p}_+$ （或 $\boldsymbol{p}_-$）的夹角较大，因此包含用于分类类别 $+1$ （或 -1）的类别无关信息。图7验证了命题 3，表明幅值剪枝会移除包含类别无关信息的神经元。

联合边-模型稀疏化带来的性能提升。图8展示了重要性采样概率 α 和剪枝率 β 变化时的学习成功率。对于每一对 α 和 β，结果取100次独立试验的平均值。我们可以观察到，当 α 或β 增加时，学习到理想模型的可能性会提高。

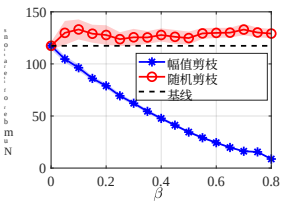


图6：迭代次数与剪枝率的关系 β

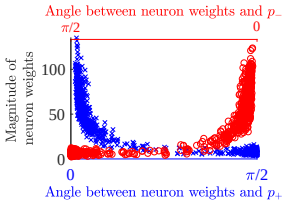


图7：神经元权重的分布

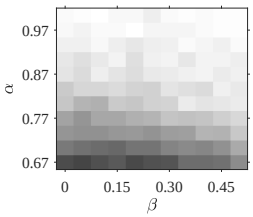


图8：联合边-模型稀疏化的学习成功率

4.2 真实引文数据集上的联合稀疏化

我们在真实引文数据集（Cora、Citeseer 和 Pubmed）上评估联合边-模型稀疏化算法（Sen 等，2008），使用标准图卷积网络（GCN）（一种双消息传递图神经网络）（Kipf 与 Welling，2017）。本文将 Chen 等，2021b 中的统一图神经网络稀疏化（*UGS*）实现为边采样方法，并采用基于幅值的剪枝作为模型剪枝方法。

图9展示了在 Cora 数据集上进行节点分类的性能。可以看到，联合稀疏化有助于降低达到与原始模型相同测试误差所需的样本复杂度。例如， P_2的采样边和剪枝神经元联合比例为（0.90,0.49）， P_3的采样边和剪枝神经元联合比例为（0.81,0.60），它们返回的模型在测试性能上优于在更大数据集上训练的原始模型（P_1）。通过改变训练样本量，我们发现了所提出理论的特征性行为：随着联合稀疏化，样本复杂度降低。

图10展示了 Citeseer 数据集在不同稀疏化率下的测试误差，颜色越深表示误差越低。在这两幅图中，我们观察到，即使剪枝超过 90% 的神经元并移除 25% 的边，联合边采样与剪枝仍能降低测试误差，这证明了联合边-模型稀疏化的有效性。此外，使用较少训练样本的联合模型-边稀疏化，性能可与不进行稀疏化时相当，甚至更好。例如，使用 120 个训练样本时，不进行任何稀疏化的测试误差为 30.4%。然而，仅使用 96 个训练样本，联合稀疏化就能将测试误差降至 28.7%。受篇幅限制，我们仅展示部分结果。更多关于合成数据集和真实数据集的实验，请参阅补充材料。

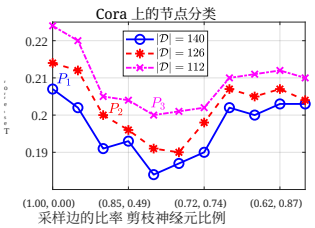


图9：Cora 上的测试误差。

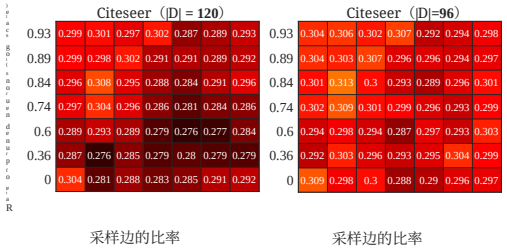


图10：展示 Citeseer 上采用 96 个和 120 个训练节点时测试误差的热力图。

5 结论

受稀疏学习器在加速图神经网络训练方面取得的实证成效所鼓舞，本文定量刻画了图采样和神经元剪枝对达到理想测试准确率所需样本复杂度及收敛速率的影响。据我们所知，这是首个针对图神经网络训练中联合边-模型稀疏化的理论泛化分析。未来的研究方向包括将分析推广到多层情形、其他图采样策略（如 FastGCN（Chen et al., 2018）），以及基于链接的分类问题。

致谢

本研究得到美国空军科学研究办公室 FA9550-20-1-0122、AROW911NF-21-1-0255、美国国家科学基金会 1932196，以及伦斯勒-IBM人工智能研究合作（http://airc.rpi.edu）的资助；该合作是 IBM人工智能视界网络（http://ibm.biz/AIHorizons）的一部分。感谢伦斯勒理工学院的 Kevin Li 和 Sissi Jian 帮助设计数值实验。感谢所有匿名审稿人提出的建设性意见。

可复现性声明

对于第 3.2 节的理论结果，我们在附录 D 中给出了必要的引理，并在附录 E 中基于这些引理给出了主要定理的完整证明。所有引理的证明均收录于附录 H。对于第4节的实验，附录 F 总结了数据和图表的生成实现细节，源代码可在补充材料中找到。

参考文献

BijayaAdhikari、YaoZhang、Sorour E Amiri、Aditya Bharadwaj 和 B Aditya Prakash。基于传播的时序网络摘要。*IEEE*知识与数据工程汇刊，30(4):729–742，2017年。

Zeyuan Allen-Zhu 和 Yuanzhi Li。特征净化：对抗训练如何实现鲁棒深度学习。发表于*2021*年 *IEEE*第*62*届计算机科学基础年会（*FOCS*），第977–988页。IEEE，2022。

Zeyuan Allen-Zhu、Yuanzhi Li 和 Yingyu Liang。过参数化神经网络中的学习与泛化：超越两层网络。载于神经信息处理系统进展，第6158–6169页，2019a。

Zeyuan Allen-Zhu、Yuanzhi Li 和 Zhao Song。通过过参数化实现深度学习的收敛理论。载于国际机器学习会议，第242–252页。PMLR，2019b。

Sanjeev Arora、Simon Du、Wei Hu、Zhiyuan Li 和 Ruosong Wang。过参数化两层神经网络的优化与泛化的细粒度分析。载于国际机器学习会议，第322–332页。PMLR，2019。

Mehdi Bahri、Gaétan Bahl 和 Stefanos Zafeiriou。二值图神经网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9492–9501页，2021年。

Peter L Bartlett 和 Shahar Mendelson。Rademacher 复杂度与高斯复杂度：风险界与结构性结果。机器学习研究期刊，3（11月）：463–482，2002年。

Alon Brutzkus 和 Amir Globerson。最大池化网络的优化与泛化分析。发表于人工智能中的不确定性，第 1650–1660 页。PMLR，2021年。

Alfredo Canziani、Adam Paszke 和 Eugenio Culurciello。面向实际应用的深度神经网络模型分析。*arXiv* 预印本 *arXiv:1605.07678*，2016。

Alireza Chakeri、Hamidreza Farhidzadeh 和 Lawrence O Hall。谱聚类中的谱稀疏化。发表于*2016*年第*23*届国际模式识别会议（*ICPR*），第 2301–2306 页。IEEE，2016年。

Deli Chen、Yankai Lin、Guangxiang Zhao、Xuancheng Ren、Peng Li、Jie Zhou 和 XuSun。面向半监督节点分类的拓扑不平衡学习。神经信息处理系统进展，34：29885–29897，2021a。

Jie Chen、Tengfei Ma 和 Cao Xiao。FastGCN：通过重要性采样实现图卷积网络的快速学习。载于国际学习表征会议，2018年。

Tianlong Chen、Yongduo Sui、Xuxi Chen、Aston Zhang 和 Zhangyang Wang。图神经网络的统一彩票假设。载于国际机器学习会议，第 1695–1706 页。PMLR，2021b。

Weilin Cong、Morteza Ramezani 和 Mehrdad Mahdavi。论深度对训练图卷积网络的可证明益处。载于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，2021年。网址：https: //openreview.net/forum?id=r-oRRT-ElX。

Gabriele Corso、Luca Cavalleri、Dominique Beaini、Pietro Liò 和 Petar Veli˘ckovi´c。面向 graphnets 的主邻域聚合。神经信息处理系统进展，33：13260–13271，2020。

Arthur da Cunha、Emanuele Natale 和 LaurentViennot。证明卷积神经网络的强彩票券假设。载于国际学习表征会议，2022。

Alexandru Damian、Jason Lee 和 Mahdi Soltanolkotabi。神经网络能够通过梯度下降学习表征。载于学习理论会议，第 5413–5452 页。PMLR，2022。

Amit Daniely 和 Eran Malach。使用神经网络学习奇偶函数。神经信息处理系统进展，33:20356–20365，2020。

Simon S Du、Xiyu Zhai、Barnabas Poczos 和 Aarti Singh。梯度下降可证明能够优化过参数化神经网络。发表于国际学习表征会议，2018年。

Simon S Du、Kangcheng Hou、Russ R Salakhutdinov、Barnabas Poczos、Ruosong Wang 和 Keyulu Xu。图神经切核：将图神经网络与图核融合。发表于神经信息处理系统进展，第5724–5734页，2019。

Talya Eden、Shweta Jain、Ali Pinar、Dana Ron 和 C Seshadhri。使用次线性图样本对度分布进行可证明且实用的近似。发表于*2018*年万维网大会论文集，第449–458页，2018年。

Jonathan Frankle 和 Michael Carbin。《彩票假设：寻找稀疏可训练神经网络》。载于国际学习表征会议，2019年。网址：https: //openreview.net/forum?id=rJl-b3RcF7。

Vikas Garg、Stefanie Jegelka 和 Tommi Jaakkola。图神经网络的泛化与表征能力界限。载于国际机器学习会议，第 3419–3430 页。PMLR，2020。

Fangtai Guo、Zaixing He、Shuyou Zhang、Xinyue Zhao、Jinhui Fang 和 Jianrong Tan。用于基于骨架的手势识别的归一化边卷积网络。模式识别，118:108044，2021年。

Will Hamilton、Zhitao Ying 和 Jure Leskovec。大型图上的归纳式表示学习。载于神经信息处理系统进展，第 1024–1034 页，2017年。

Xiangnan He、Kuan Deng、Xiang Wang、Yan Li、Yongdong Zhang 和 Meng Wang。《LightGCN：简化并增强用于推荐的图卷积网络》。载于第 *43* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，第 639–648 页，2020 年。

Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。《神经网络中的知识蒸馏》。统计，1050：9，2015 年。

Weihua Hu、Matthias Fey、Marinka Zitnik、Yuxiao Dong、Hongyu Ren、Bowen Liu、Michele Catasta 和 Jure Leskovec。开放图基准：用于图机器学习的数据集。神经信息处理系统进展，33：22118–22133，2020。

Baihe Huang、Xiaoxiao Li、Zhao Song 和 Xin Yang。Fl-ntk：一种基于神经切线核的联邦学习分析框架。载于国际机器学习会议，第 4423–4434 页。PMLR，2021年。

Christian Hübler、Hans-Peter Kriegel、Karsten Borgwardt 和 Zoubin Ghahramani。《用于代表性子图采样的 Metropolis算法》。载于*2008*年第八届*IEEE*国际数据挖掘会议，第283–292页。IEEE，2008年。

RichardHulland Roger King。《语义数据库建模：综述、应用与研究问题》。*ACM*计算调查，19(3)：201–260，1987年。

Matthew O Jackson。《社会与经济网络》。普林斯顿大学出版社，2010年。

Arthur Jacot、Franck Gabriel 和 Clément Hongler。《神经切线核：神经网络中的收敛与泛化》。载于第*32*届神经信息处理系统国际会议论文集，2018年。

Ajay Kumar Jaiswal、Haoyu Ma、Tianlong Chen、Ying Ding 和 Zhangyang Wang。《抽中中奖彩票后，如何更好地使用它》。*arXiv*预印本 *arXiv:2101.03255*，2021年。

Svante Janson。部分依赖随机变量之和的大偏差。随机结构与算法，24(3)：234–248，2004年。

Stefani Karp、Ezra Winston、Yuanzhi Li 和 Aarti Singh。局部信号自适应性：超越核方法的神经网络可证明特征学习。神经信息处理系统进展，34：24883–24897，2021年。

Thomas N. Kipf 和 Max Welling。利用图卷积网络进行半监督分类。载于国际学习表征会议，2017年。

Jaehoon Lee、Yasaman Bahri、Roman Novak、Samuel S Schoenholz、Jeffrey Pennington 和 Jascha Sohl-Dickstein。作为高斯过程的深度神经网络。载于 国际学习表征会议，2018 年。

Jure Leskovec 和 Christos Faloutsos。大型图上的采样。载于 第*12*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第631–636页，2006年。

Hongkang Li、Meng Wang、Sijia Liu、Pin-Yu Chen 和 Jinjun Xiong。采用图拓扑采样训练图卷积网络的泛化保证。载于国际机器学习会议，第 13014–13051 页。PMLR，2022。

Hongkang Li、Meng Wang、Sijia Liu 和 Pin-Yu Chen。视觉 Transformer 的理论理解：学习、泛化与样本复杂度。载于国际学习表征会议，2023年。网址：https://openreview.net/forum?id=jClGv3Qjhb。

Yuanzhi Li 和 Yingyu Liang。通过随机梯度下降在结构化数据上学习过参数化神经网络。载于神经信息处理系统进展，第 8157–8166 页，2018年。

Shiyu Liang、Ruoyu Sun、Yixuan Li 和 Rayadurgam Srikant。理解用于二分类的神经网络损失曲面。载于国际机器学习会议，第 2835–2843 页。PMLR，2018年。

Zirui Liu、Kaixiong Zhou、Fan Yang、Li Li、Rui Chen 和 Xia Hu。Exact：通过极端激活压缩实现可扩展的图神经网络训练。载于国际学习表征会议，2021年。

Wai Weng Lo、Siamak Layeghy、Mohanad Sarhan、Marcus Gallagher、Marius Portmann。E-GraphSAGE：一种基于图神经网络的物联网入侵检测系统。载于*NOMS 2022-2022 IEEE/IFIP*网络运营与管理研讨会，第 1–9 页。IEEE，2022。

Eran Malach、Gilad Yehudai、Shai Shalev-Schwartz 和 Ohad Shamir。《证明彩票假设：剪枝就是你所需要的一切》。载于国际机器学习会议，第 6682–6691 页。PMLR，2020年。

Jihun Oh、Kyunghyun Cho 和 Joan Bruna。《利用数据驱动的节点采样推进 GraphSAGE》。*arXiv* 预印本 *arXiv:1904.12935*，2019。Samet Oymak 和 Mahdi Soltanolkotabi。《迈向适度过参数化：训练浅层神经网络的全局收敛保证》。*IEEE*信息论选刊，1(1):84–105，2020。Shirui Pan、Ruiqi Hu、Guodong Long、Jing Jiang、Lina Yao 和 Chengqi Zhang。《用于图嵌入的对抗正则化图自编码器》。*arXiv* 预印本 *arXiv:1802.04407*，2018年。Bryan Perozzi、Rami Al-Rfou 和 Steven Skiena。《DeepWalk：社交表征的在线学习》。载于第*20*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 701–710 页，2014年。Ryan Riegel、Alexander Gray、Francois Luus、Naweed Khan、Ndivhuwo Makondo、Ismail Yunus Akhalwaya、Haifeng Qian、Ronald Fagin、Francisco Barahona、Udit Sharma 等。《逻辑神经网络》。*arXiv* 预印本 *arXiv:2006.13155*，2020。

Yu Rong、Wenbing Huang、Tingyang Xu 和 Junzhou Huang。DropEdge：面向节点分类的深层图卷积网络。载于国际学习表征会议，2019。

Itay Safran 和 Ohad Shamir。伪局部极小值在双层ReLU神经网络中很常见。载于国际机器学习会议，第4430–4438页，2018年。

Aliaksei Sandrykhaila 和 Jose MF Moura。利用图上的信号处理进行大数据分析：具有不规则结构的海量数据集的表示与处理。*IEEE*信号处理杂志，31(5)：80–90，2014年。

Franco Scarselli、Ah Chung Tsoi 和 Markus Hagenbuchner。图神经网络与递归神经网络的Vapnik–Chervonenkis维数。神经网络，108：248–259，2018年。

Michael Schlichtkrull、Thomas N Kipf、Peter Bloem、Rianne van den Berg、Ivan Titov 和 Max Welling。利用图卷积网络对关系数据建模。载于欧洲语义网会议论文集，第593–607页。施普林格，2018年。

Prithviraj Sen、Galileo Namata、MustafaBilgic、Lise Getoor、BrianGalligher 和 TinaEliassi-Rad。《网络数据中的集体分类》。《人工智能杂志》，29(3):93–93，2008年。

Shai Shalev-Shwartz 和 Shai Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。

Weijing Shi 和 Raj Rajkumar。Point-gnn：用于点云三维目标检测的图神经网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1711–1719页，2020。

Zhenmei Shi、Junyi Wei 和 Yingyu Liang。神经网络中特征学习的理论分析：从输入中涌现及其相对于固定特征的优势。载于国际学习表征会议，2022。

Shyam Anil Tailor、Javier Fernandez-Marques 和 Nicholas Donald Lane。Degree-quant：面向图神经网络的量化感知训练。载于国际学习表征会议，2020。

Petar Veli˘ckovi´c、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。国际学习表征会议，2018年。

Saurabh Verma 和 Zhi-Li Zhang。图卷积神经网络的稳定性与泛化。载于第 *25* 届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 1539–1548 页，2019 年。Elli Voudigari、Nikos Salamanos、Theodore Papageorgiou 和 Emmanuel J Yannakoudakis。Rank degree：一种高效的图采样算法。载于*2016 IEEE/ACM*社会网络分析与挖掘进展国际会议，第 120–129 页。IEEE，2016 年。

Kuansan Wang、Zhihong Shen、Chiyuan Huang、Chieh-Han Wu、Yuxiao Dong 和 Anshul Kanakia。微软学术图谱：专家不够时怎么办。定量科学研究，1(1)：396–413，2020 年。

Zixin Wen 和 Yuanzhi Li。探讨自监督对比学习中的特征学习过程。载于国际机器学习会议，第 11112–11122 页。PMLR，2021年。

Felix Wu、Amauri Souza、Tianyi Zhang、Christopher Fifty、Tao Yu 和 Kilian Weinberger。简化图卷积网络。载于国际机器学习会议，第 6861–6871 页，2019。

Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey 等。谷歌的神经机器翻译系统：弥合人类翻译与机器翻译之间的差距。*arXiv* 预印本 *arXiv:1609.08144*，2016。

Zonghan Wu、Shirui Pan、Fengwen Chen、Guodong Long、Chengqi Zhang 和 S Yu Philip。图神经网络综合综述。*IEEE*神经网络与学习系统汇刊，32(1):4–24，2020。

Keyulu Xu、Chengtao Li、Yonglong Tian、Tomohiro Sonobe、Ken-ichi Kawarabayashi 和 Stefanie Jegelka。利用跳跃知识网络进行图上的表示学习。载于国际机器学习会议，第 5453–5462 页。PMLR，2018年。

Keyulu Xu、Mozhi Zhang、Stefanie Jegelka 和 Kenji Kawaguchi。图神经网络优化：跳跃连接与更深网络带来的隐式加速。载于国际机器学习会议，第 11592–11602 页。PMLR，2021年。

Sijie Yan、Yuanjun Xiong 和 Dahua Lin。用于基于骨架的动作识别的时空图卷积网络。载于人工智能促进协会会议论文集，2018年。

Yiding Yang、Jiayan Qiu、Mingli Song、Dacheng Tao 和 Xinchao Wang。从图卷积网络中提炼知识。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 7074–7083 页，2020。

Huaxiu Yao、Chuxu Zhang、Ying Wei、Meng Jiang、Suhang Wang、Junzhou Huang、Nitesh Chawla 和 Zhenhui Li。通过知识迁移实现图小样本学习。载于*AAAI*人工智能会议论文集，第34卷，第6656–6663页，2020年。

Rex Ying、Ruining He、Kaifeng Chen、Pong Eksombatchai、William L Hamilton 和 Jure Leskovec。面向网络规模推荐系统的图卷积神经网络。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第974–983页，2018年。

Haoran You、Zhihan Lu、Zijian Zhou、Yonggan Fu 和 Yingyan Lin。Early-birdgcns：通过抽取早鸟彩票，实现面向更高效GCN训练与推理的图网络协同优化。*AAAI*人工智能会议，2022。

Shuai Zhang、Meng Wang、Sijia Liu、Pin-Yu Chen 和 Jinjun Xiong。具有泛化能力保证的图神经网络快速学习：单隐藏层情形。载于国际机器学习会议，2020a。

Shuai Zhang、Meng Wang、Jinjun Xiong、Sijia Liu 和 Pin-Yu Chen。具有泛化能力保证的卷积神经网络训练线性收敛性的改进：单隐藏层情形。*IEEE*神经网络与学习系统汇刊，32(6):2622–2635，2020b。

Shuai Zhang、Meng Wang、Sijia Liu、Pin-Yu Chen 和 Jinjun Xiong。无标签数据如何提升自训练中的泛化能力？单隐藏层的理论分析。载于国际学习表征会议，2022a。

Tao Zhang、Hao-Ran Shan 和 Max A Little。Causal GraphSAGE：一种基于因果采样的稳健图分类方法。模式识别，128:108696，2022b。

Zeru Zhang、Jiayin Jin、Zijie Zhang、Yang Zhou、Xin Zhao、Jiaxiang Ren、Ji Liu、Lingfei Wu、RuomingJin 和 Dejing Dou。利用惯性流形理论验证彩票假设。神经信息处理系统进展，34，2021年。

Cheng Zheng、Bo Zong、Wei Cheng、Dongjin Song、Jingchao Ni、Wenchao Yu、Haifeng Chen 和 Wei Wang。通过神经稀疏化实现鲁棒的图表示学习。载于国际机器学习会议，第 11458–11468 页。PMLR，2020。

Liyuan Zheng、Zhen Zuo、Wenbo Wang、Chaosheng Dong、Michinari Momma 和 Yi Sun。采用邻居相似度注意力机制的异构图神经网络，用于替代产品推荐。*AAAI*人工智能会议，2021年。

Kai Zhong、Zhao Song、Prateek Jain、Peter L Bartlett 和 Inderjit S Dhillon。单隐层神经网络的恢复保证。载于国际机器学习会议，第 4140–4149 页，2017年。

Xianchen Zhou 和 Hongxia Wang。随着层数增加，图卷积网络的泛化误差可能增大。*Neurocomputing*，424:97–106，2021年。

Difan Zou、Ziniu Hu、Yewen Wang、Song Jiang、Yizhou Sun 和 Quanquan Gu。用于训练深层大规模图卷积网络的层依赖重要性采样。神经信息处理系统进展论文集，32，2019。

补充材料：

联合边-模型稀疏学习对图神经网络而言已被证明是高效的

在下文中，相关工作收录于附录 A。附录 B 概述了证明技术的基本思路。附录 C 总结了证明所用的记号，有用的引理则收录于附录 D。附录 E 给出了定理 2 的详细证明；定理 2 是定理 1 的正式版本。附录 F 介绍了第4节中合成数据实验的详细信息，并收录了若干其他实验结果，以弥补正文篇幅有限的不足。附录 G 证明了 VC 维度的下界。附录 G.1 给出了针对某些边采样策略的 α 界。附录 H 总结了有用引理的补充证明。此外，我们在附录 I 中概述了将本文框架扩展到多类分类问题的基本思路。

A 相关工作

图神经网络的泛化分析。 近期有两篇论文（Du 等，2019；Xu 等，2021）利用神经切线核（NTK）框架（Malach 等，2020；Allen-Zhu 等，2019b；Jacot 等，2018；Du 等，2018；Lee 等，2018）分析图神经网络的泛化能力。文献（Du 等，2019）表明，只有当标签由某种特殊函数生成时，图神经切线核（GNTK）才能实现有界的泛化误差，例如，该函数需要是线性函数或偶函数。文献（Xu 等，2021）分析了带跳跃连接的深度线性图神经网络的泛化能力。NTK 方法考虑模型充分过参数化的情形，即神经元数量是样本数量的多项式函数，从而使风险函数在任意初始化附近的景观几乎呈凸形。所需的模型复杂度远高于实际情况，而且这些结果与数据分布无关。由于神经网络的学习过程与输入结构密切相关（Shi et al., 2022），无分布分析（例如 NTK）可能无法准确解释具有特殊结构的数据上的学习性能。沿用模型恢复框架（Zhong 等，2017；Zhang 等，2022a；2020b），Zhang 等人（2020a）在假设特征服从高斯分布的前提下分析了单隐藏层图神经网络的泛化能力，但该分析要求采用特殊的张量初始化方法，也无法解释随机初始化下随机梯度下降的实际成功。除此之外，文献（Scarselli 等，2018；Garg 等，2020）通过经典的拉德马赫复杂度，以及文献（Verma 与 Zhang，2019；Zhou 与 Wang，2021）通过一致稳定性框架，刻画了训练误差与测试误差之间的泛化间隙。

具有数据结构约束的泛化分析。 假设数据来自良好分离的分布的混合分布，Li & Liang, 2018 分析了单隐层全连接神经网络的泛化。近期研究（Shi et al., 2022；Brutzkus & Globerson, 2021；Allen-Zhu & Li, 2022；Karp et al., 2021；Wen & Li, 2021；Li et al., 2023）在假设数据可以划分为判别性特征和背景模式的情况下，分析了单隐层神经网络。具有非线性激活函数的神经网络能够记忆判别性特征，并且在具有相同结构约束的未见数据上保证泛化，而采用随机初始化的线性分类器无法在多项式规模和时间内学习数据映射（Shi et al., 2022；Daniely & Malach, 2020）。然而，这些研究都未考虑图神经网络或稀疏化。

B 技术概述

在介绍证明细节之前，我们将在本节概述证明技术。首先，我们先不涉及边和模型稀疏化方法，概述证明概要，以作铺垫

。然后，我们将阐述推导边和模型稀疏化方法相关结果时面临的主要挑战。

B.1 带结构约束的数据上的图神经网络学习 S

为方便表述，我们分别用 $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 表示 $\mathcal{D}$ 中带正标签和负标签的节点集合，其中 $\mathcal{D}_+ = (\mathcal{V}_+ \cup \mathcal{V}_{N+}) \cap \mathcal{D}$，且 $\mathcal{D}_- = (\mathcal{V}_- \cup \mathcal{V}_{N-}) \cap \mathcal{D}$。回顾一下，$\boldsymbol{p}_+$ 只存在于节点 $v \in \mathcal{D}_+$ 的邻居中，而 $\boldsymbol{p}_-$ 只存在于节点 $v \in \mathcal{D}_-$ 的邻居中。相比之下，对于 $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 中的数据，类别无关模式的分布相同。此外，对于某些神经元，当 $v \in \mathcal{D}_+$ 时，其梯度方向始终接近 $\boldsymbol{p}_+$；而由 $v \in \mathcal{D}_-$ 得出的梯度始终几乎与 $\boldsymbol{p}_+$ 正交。这样的神经元即为第 3.4 节命题 1 中定义的幸运神经元，我们将在附录 C 中从另一个角度正式定义幸运神经元。以 $\mathcal{B}_+$ 中的神经元为例，其中 $\mathcal{B}_+ = \{k \mid b_k = +1\}$ 表示线性层中具有正系数的神经元集合。对于幸运神经元 $k \in \mathcal{K}_+$，权重在$\boldsymbol{p}_+$ 上的投影严格增加（参见附录 D 中的引理 2）。对于其他神经元，即不幸神经元，类别无关模式服从相同分布且与 y_v独立，由 $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 产生的梯度相近。具体而言，由于 $\mathcal{D}_+$ 与 $\mathcal{D}_-$之间存在偏移，总体梯度的阶为 $\sqrt{(1+R^2)/|\mathcal{D}|}$，其中 R 为图的度。当训练样本数量足够大时，权重在类别无关模式上的投影增长速度远慢于其在类别相关模式上的投影。关于神经元权重更新的示意图，请参见图11。

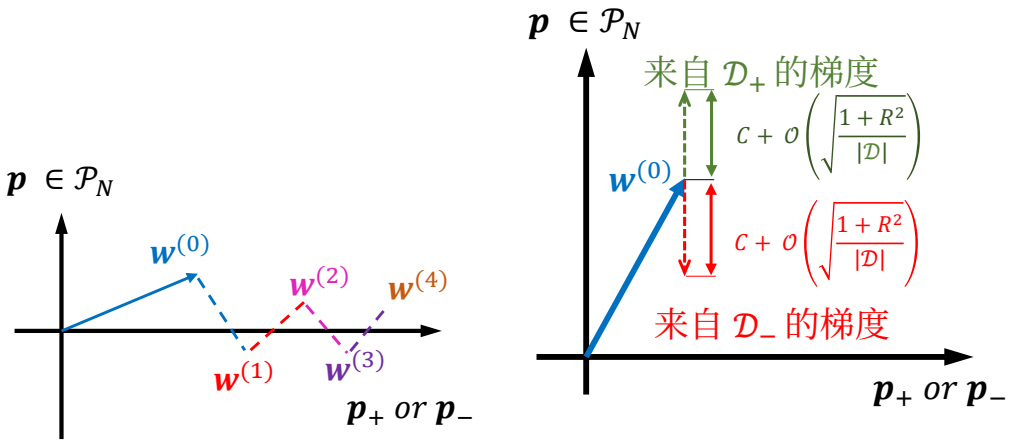


图11：迭代次数 $\{\boldsymbol{w}^{(t)}\}_{t=1}^{T}$示意图：（a）幸运神经元，以及（b）不幸神经元。

与针对 $k \in \mathcal{B}_+$的 $\boldsymbol{w}_k^{(t)}$ 推导类似，我们可以证明，$\mathcal{K}_-$中的神经元（即相对于 $k \in \mathcal{B}_-$而言的幸运神经元）其权重更新主要沿着 $\boldsymbol{p}_-$的方向。回想一下，图神经网络模型的输出写作式（20），$k \in \mathcal{B}_+$ 在线性层中的相应系数均为正。有了这些结论，我们知道，与其他模式相比，$\mathcal{B}_+$ 中的神经元在 $\boldsymbol{p}_+$ 方向上的幅度相对较大，且相应系数 b_k为正。因此，对于节点 $v \in \mathcal{V}_+ \cup \mathcal{V}_{N+}$，计算得到的标签将严格为正。与上述推导类似，节点 $v \in \mathcal{V}_- \cup \mathcal{V}_{N-}$ 的计算标签将严格为负。

边稀疏化。图11(b)表明，在不进行采样时，沿任意类别无关模式方向的梯度是 $\sqrt{1+R^2}$ 的线性函数。对边进行采样可以显著降低图的度，即仅采样 r 个邻居节点时，图的度会降至 r。因此，神经元权重在任意 $\boldsymbol{p} \in \mathcal{P}_N$ 上的投影会变小，神经元也就不太可能学习类别无关特征。此外，由于只需遍历边的一个子集，每次迭代的计算复杂度也会降低。然而，采样

图边可能会导致某些训练节点的类别相关特征遗漏（α 较小），从而降低收敛速率，并需要更多迭代次数。

模型稀疏化。比较图11(a)和图11(b)可以看出，幸运神经元的幅值增长速度远快于不幸神经元。此外，根据引理 6，我们知道，初始化时的幸运神经元在后续迭代中始终会是幸运神经元。因此，对原始稠密模型采用基于幅值的剪枝方法，会移除不幸神经元，同时保留幸运神经元。幸运神经元比例提高后，神经元能更快地学习类别相关特征。此外，该算法能够容忍输入中类别无关模式所导致的更大梯度噪声，其量级为图11(b)中的 $1/\sqrt{|\mathcal{D}|}$。因此，收敛所需的样本数可以大幅减少。

噪声因子 z。噪声因子 $\boldsymbol{z}$ 主要从以下几个方面影响泛化。首先，在 Brutzkus & Globerson（2021）中，样本复杂度取决于 $\{\boldsymbol{x}_v\}_{v\in\mathcal{D}}$ 的大小；然而，在存在噪声时，该大小可能高达 $|\mathcal{D}|$。其次，幸运神经元比例会随着噪声水平的变化而降低。幸运神经元比例越低，收敛所需的训练样本和迭代次数就越多。

C 记号

在本节中，我们将具体说明第 3.2 节所述的数据模型和问题形式化，并定义一些重要记号，以简化证明的表述。此外，以下证明中使用的所有记号均汇总于表2和表3。

C.1 带结构约束的数据模型

回顾第 3.2 节中的定义，节点 v 的节点特征写作

$$\boldsymbol{x}_v = \boldsymbol{p}_v + \boldsymbol{z}_v, \tag{14}$$

其中，$\boldsymbol{p}_v \in \mathcal{P}$，且 $\boldsymbol{z}_v$ 是有界噪声，满足 $\|\boldsymbol{z}_v\|_2 \le \sigma$。此外，$\mathcal{P}$中有 L 个正交模式，记作 $\{\boldsymbol{p}_\ell\}_{\ell=1}^{L}$。$\boldsymbol{p}_+ := \boldsymbol{p}_1$ 是正类相关模式，$\boldsymbol{p}_- := \boldsymbol{p}_2$ 是负类相关模式，其余模式记作 $\mathcal{P}_N$，为类别无关模式。对于节点 v，如果其邻居中包含 $\boldsymbol{p}_+$ 或 $\boldsymbol{p}_-$，则其标签 y_v 为正或负。称节点 v包含类别相关特征，表示 $\boldsymbol{p}_v = \boldsymbol{p}_+$ 或 $\boldsymbol{p}_-$。

根据 $\boldsymbol{x}_v$ 和 y_v，我们将 $\mathcal{V}$ 中的节点划分为四个互不相交的分区，即 $\mathcal{V} = \mathcal{V}_+ \cup \mathcal{V}_- \cup \mathcal{V}_{N+} \cup \mathcal{V}_{N-}$，其中

$$\begin{aligned}
\mathcal{V}_+ &:= \{v \mid \boldsymbol{p}_v = \boldsymbol{p}_+\};\\
\mathcal{V}_- &:= \{v \mid \boldsymbol{p}_v = \boldsymbol{p}_-\};\\
\mathcal{V}_{N+} &:= \{v \mid \boldsymbol{p}_v \in \mathcal{P}_N, y_v = +1\};\\
\mathcal{V}_{N-} &:= \{v \mid \boldsymbol{p}_v \in \mathcal{P}_N, y_v = -1\}.
\end{aligned} \tag{15}$$

然后，我们考虑如下模型：(i) $\boldsymbol{p}_+$ 和 $\boldsymbol{p}_-$ 的分布相同，即 y,

$$\text{Prob}(\boldsymbol{p}_v = \boldsymbol{p}_+) = \text{Prob}(\boldsymbol{p}_v = \boldsymbol{p}_-), \tag{16}$$

以及 (ii) $\boldsymbol{p} \in \mathcal{P}_N$ 在 $\mathcal{V}_{N+}$ 和 $\mathcal{V}_{N-}$中的分布相同，即

$$\text{Prob}(\boldsymbol{p} \mid v \in \mathcal{V}_{N+}) = \text{Prob}(\boldsymbol{p} \mid v \in \mathcal{V}_{N-}) \quad \text{for any} \quad \boldsymbol{p} \in \mathcal{P}_N. \tag{17}$$

容易验证，当 (16) 和 (17) 成立时，$\mathcal{D}$ 中正标签和负标签的数量是平衡的，即

$$\text{Prob}(y_v = +1) = \text{Prob}(y_v = -1) = \frac{1}{2}. \tag{18}$$

如果 $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 高度不平衡，即 $\big||\mathcal{D}_+| - |\mathcal{D}_-|\big| \gg \sqrt{|\mathcal{D}|}$，则 (2) 中的目标函数可以修改为

$$\hat{f}_{\mathcal{D}} := -\frac{1}{2|\mathcal{D}_+|}\sum_{v\in\mathcal{D}_+} y_v \cdot g\big(\boldsymbol{W}; \boldsymbol{X}_{\mathcal{N}(v)}\big) - \frac{1}{2|\mathcal{D}_-|}\sum_{v\in\mathcal{D}_-} y_v \cdot g\big(\boldsymbol{W}; \boldsymbol{X}_{\mathcal{N}(v)}\big), \tag{19}$$

并且将式（6）中所需的样本数量 $|\mathcal{D}|$ 替换为 $\min\{|\mathcal{D}_+|, |\mathcal{D}_-|\}$。

C.2 图神经网络模型

容易验证，（1）等价于以下模型

$$g(\boldsymbol{W},\boldsymbol{U};\boldsymbol{x})=\frac{1}{K}\sum_{k}\text{AGG}(\boldsymbol{X}_{\mathcal{N}(v)},\boldsymbol{w}_k)-\frac{1}{K}\sum_{k}\text{AGG}(\boldsymbol{X}_{\mathcal{N}(v)},\boldsymbol{u}_k),\tag{20}$$

其中，(20) 中的神经元权重 $\{\boldsymbol{w}_k\}_{k=1}^K$ 对应于 (1) 中的神经元权重 $b_k=+1$，而 (20) 中的神经元权重 $\{\boldsymbol{u}_k\}_{k=1}^K$ 对应于 (1) 中的神经元权重 $b_k=-1$。这里，我们滥用符号 K 来表示 $\{k|b_k=+1\}$ 或 $\{k|b_k=-1\}$中的神经元数量，它与 (1) 中的 K 相差 2 倍。由于本文旨在提供阶数分析，因此 (1) 和 (20) 中 K 的界相同。

与 (20) 中的模型相对应，我们将 $\boldsymbol{M}_+$ 记为相对于 $\boldsymbol{W}$ 剪枝后的掩码矩阵，将 $\boldsymbol{M}_-$ 记为相对于 $\boldsymbol{U}$剪枝后的掩码矩阵。为便于分析，我们考虑在 $\boldsymbol{W}$和 $\boldsymbol{U}$中进行均衡剪枝，即 $\|\boldsymbol{M}_+\|_0=\|\boldsymbol{M}_-\|_0=(1-\beta)Kd$。

C.3 证明的补充记号

模式函数 $\mathcal{M}(v)$. 现在，回顾一下，在迭代 t时，节点 v 的聚合函数写作

$$\text{AGG}(\boldsymbol{X}_{\mathcal{N}^{(t)}(v)},\boldsymbol{w}_k^{(t)})=\max_{n\in\mathcal{N}^{(t)}(v)}\ \phi(\langle\ \boldsymbol{w}_k^{(t)}\ ,\ \boldsymbol{x}_n\ \rangle).\tag{21}$$

然后，在迭代 t时，我们将模式函数 $\mathcal{M}^{(t)}$ 定义为：$\mathcal{V}\to\mathcal{P}\bigcup\{\boldsymbol{0}\}$ 在迭代 t 时为

$$\mathcal{M}^{(t)}(v;\boldsymbol{w})=\begin{cases}\boldsymbol{0}, & \text{if}\quad \max\limits_{n\in\mathcal{N}^{(t)}(v)}\ \phi(\langle\ \boldsymbol{w}\ ,\ \boldsymbol{x}_n\ \rangle)\le 0\\ \underset{\{\boldsymbol{x}_n|n\in\mathcal{N}^{(t)}(v)\}}{\text{argmax}}\ \phi(\langle\ \boldsymbol{w}\ ,\ \boldsymbol{x}_n\ \rangle), & \text{otherwise}\end{cases}.\tag{22}$$

类似于式（14）中 $\boldsymbol{p}_v$ 的定义，我们定义 $\mathcal{M}_p^{(t)}$和 $\mathcal{M}_z^{(t)}$，使得 $\mathcal{M}_p^{(t)}$是相对于 $\mathcal{M}^{(t)}$ 的无噪声模式，而 $\mathcal{M}_z^{(t)}$是相对于 $\mathcal{M}^{(t)}$的噪声。

此外，对于不进行边采样的情况，我们定义 $\mathcal{M}$：$\mathcal{V}\to\mathcal{P}\bigcup\{\boldsymbol{0}\}$，使得

$$\mathcal{M}(v;\boldsymbol{w})=\begin{cases}\boldsymbol{0}, & \text{if}\quad \max\limits_{n\in\mathcal{N}(v)}\ \phi(\langle\ \boldsymbol{w}\ ,\ \boldsymbol{x}_n\ \rangle)\le 0\\ \underset{\{\boldsymbol{x}_n|n\in\mathcal{N}(v)\}}{\text{argmax}}\ \phi(\langle\ \boldsymbol{w}\ ,\ \boldsymbol{x}_n\ \rangle), & \text{otherwise}\end{cases}.\tag{23}$$

幸运神经元的定义。当且仅当某个神经元在迭代 t 时的权重向量 $\{\boldsymbol{w}_k^{(t)}\}_{k=1}^K$ 满足以下条件时，我们称其为幸运神经元

$$\mathcal{M}_p(v;\boldsymbol{w}_k^{(t)})\equiv\boldsymbol{p}_+\quad\text{for any}\quad v\in\mathcal{V}\tag{24}$$

或其权重向量满足 $\{\boldsymbol{u}_k^{(t)}\}_{k=1}^K$

$$\mathcal{M}_p(v;\boldsymbol{u}_k^{(t)})\equiv\boldsymbol{p}_-\quad\text{for any}\quad v\in\mathcal{V}.\tag{25}$$

设 $\mathcal{W}(t)$、$\mathcal{U}(t)$ 为位于 t次迭代中的幸运神经元，使得

$$\begin{aligned}\mathcal{W}(t)&=\{k\mid\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})=\boldsymbol{p}_+\quad\text{for any}\quad v\in\mathcal{V}\},\\ \mathcal{U}(t)&=\{k\mid\mathcal{M}_p^{(t)}(v;\boldsymbol{u}_k^{(t)})=\boldsymbol{p}_-\quad\text{for any}\quad v\in\mathcal{V}\}\end{aligned}\tag{26}$$

其余所有神经元，记为 $\mathcal{W}^c(t)$ 和 $\mathcal{U}^c(t)$，都是第 t 次迭代中的不幸神经元。与第 3.4 节命题 1 中“幸运神经元”的定义相比，我们有 $\cap_{t=1}^T\mathcal{W}(t)$ $\mathcal{K}_+$ 6 $\mathcal{K}_+=\cap_{t=1}^T\mathcal{W}$ t $=\mathcal{W}$ 0 。根据下文的上下文（见引理），可以验证 ( ) ( )。

幸运神经元和不幸神经元的梯度 **以及** 不幸神经元。 我们可以将式 (4) 中的梯度下降改写为

$$\begin{aligned}&\frac{\partial \hat{f}_{\mathcal{D}}^{(t)}}{\partial \boldsymbol{w}_k^{(t)}}\\ =&-\frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}}\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}^{(t)}(v)})}{\partial \boldsymbol{w}_k^{(t)}}\\ =&-\frac{1}{2|\mathcal{D}_+|}\sum_{v\in\mathcal{D}_+}\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}^{(t)}(v)})}{\partial \boldsymbol{w}_k^{(t)}}+\frac{1}{2|\mathcal{D}_-|}\sum_{v\in\mathcal{D}_-}\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}^{(t)}(v)})}{\partial \boldsymbol{w}_k^{(t)}},\end{aligned}\tag{27}$$

其中， $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 分别表示 $\mathcal{D}$ 中具有正标签和负标签的节点集合。根据 (22) 中 $\mathcal{M}^{(t)}$ 的定义，容易验证

$$\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}(v)})}{\partial \boldsymbol{w}_k^{(t)}}=\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}),\tag{28}$$

而 $\boldsymbol{w}_k$ 的更新为

$$\begin{aligned}\boldsymbol{w}_k^{(t+1)}=&\ \boldsymbol{w}_k^{(t)}+c_\eta\cdot\mathbb{E}_{v\in\mathcal{D}}\ y_v\cdot\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}),\\ \text{or}\quad \boldsymbol{w}_k^{(t+1)}=&\ \boldsymbol{w}_k^{(t)}+c_\eta\cdot\mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)})-c_\eta\cdot\mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}),\end{aligned}\tag{29}$$

其中，我们滥用符号 $\mathbb{E}_{v\in\mathcal{S}}$ 来表示

$$\mathbb{E}_{v\in\mathcal{S}}f(v)=\frac{1}{|\mathcal{S}|}\sum_{v\in\mathcal{S}}f(v)\tag{30}$$

对于任意集合 $\mathcal{S}$ 和某个函数 f。此外，不失一般性，我们不考虑满足 $\max_{\boldsymbol{p}\in\mathcal{P}}\langle\ \boldsymbol{w}_k^{(0)}\ ,\ \boldsymbol{p}\ \rangle<0$ 的神经元，因为：(1) 该神经元完全不会被更新；(2) 当 2^{-L}时，该神经元出现的概率可以忽略不计。

最后，由于本文关注的是阶数分析，在部分证明中可以忽略一些常数。具体而言，我们用 $h_1(L)\gtrsim h_2(L)$ 表示存在某个正数常量C，使得当 $L\in\mathbb{R}$ 足够大时， $h_1(L)\geq C\cdot h_2(L)$。对于 $h_1(L)\eqsim h_2(L)$和 $h_1(L)\lesssim h_2(L)$，也可以作类似定义。

D 有用引理

引理 1 说明了神经元数量与幸运神经元比例之间的关系。当隐藏层中的神经元数量足够大，满足式 (31) 时，根据式 (32)，幸运神经元比例至少为 $(1-\varepsilon_K-L\sigma/\pi)/L$，其中 σ 为噪声水平， L 为模式数量。

引理 1. 假设初始化 $\{\boldsymbol{w}_k^{(0)}\}_{k=1}^K$ 和 $\{\boldsymbol{u}_k^{(0)}\}_{k=1}^K$ 由独立同分布的高斯分布生成。那么，若神经元数量 K 足够大，满足

$$K\geq\varepsilon_K^{-2}L^2\log q,\tag{31}$$

式(24) 和(25)中定义的幸运神经元比例满足

$$\rho\geq(1-\varepsilon_K-\frac{L\sigma}{\pi})\cdot\frac{1}{L}\tag{32}$$

引理 2 和引理 3 展示了幸运神经元在类别相关模式和类别无关模式方向上的权重投影。

引理 2. 对于幸运神经元 $k\in\mathcal{W}(t)$，令 $\boldsymbol{w}_k^{(t+1)}$ 为算法 *1* 返回的下一次迭代。那么，神经元权重满足以下不等式：

1. 在 $\boldsymbol{p}_+$方向上，我们有

$$\langle\ \boldsymbol{w}_k^{(t+1)}\ ,\ \boldsymbol{p}_+\ \rangle\geq\langle\ \boldsymbol{w}_k^{(t)}\ ,\ \boldsymbol{p}_+\ \rangle+c_\eta\left(\alpha-\sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right);$$

表2：集合的重要记号

$[Z], Z \in \mathbb{N}_+$	$\{1, 2, 3, \cdots, Z\}$的集合
$\mathcal{V}$	图 $\mathcal{G}$中的节点集合
$\mathcal{E}$	图 $\mathcal{G}$中的边集合
$\mathcal{P}$	类别相关模式和类别无关模式的集合
$\mathcal{K}_+$	相对于 $\boldsymbol{W}^{(0)}$ 的幸运神经元集合
$\mathcal{K}_-$	相对于 $\boldsymbol{U}^{(0)}$ 的幸运神经元集合
$\mathcal{K}_{\beta+}(t)$	相对于 $\boldsymbol{W}^{(t)}$ 的未剪枝神经元集合
$\mathcal{K}_{\beta-}(t)$	相对于 $\boldsymbol{U}^{(t)}$ 的未剪枝神经元集合
$\mathcal{P}_N$	类别无关模式的集合
$\mathcal{D}$	训练数据的集合
$\mathcal{D}_+$	带正标签的训练数据集合
$\mathcal{D}_-$	带负标签的训练数据集合
$\mathcal{N}(v)$ $v \in \mathcal{V}$	图 $\mathcal{G}$中节点 v 的邻居节点（包括 v 其本身）
$\mathcal{N}_s^{(t)}(v)$ $v \in \mathcal{V}$迭代 t 时节点 v 的采样节点	
$\mathcal{W}(t)$	迭代 t 时关于权重 $\boldsymbol{W}^{(t)}$ 的幸运神经元集合
$\mathcal{U}(t)$	迭代 t 时关于权重 $\boldsymbol{U}^{(t)}$ 的幸运神经元集合
$\mathcal{W}^c(t)$	迭代 t 时关于权重 $\boldsymbol{W}^{(t)}$ 的不幸神经元集合
$\mathcal{U}^c(t)$	在第 t 次迭代中，相对于权重 $\boldsymbol{U}^{(t)}$ 的不幸神经元集合

2. 在 $\boldsymbol{p}_-$ 方向上，或对于类别无关模式，使得对于任意 $\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_+$，我们有

$$\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p} \rangle \geq -c_\eta - \sigma - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}},$$

and

$$\langle \boldsymbol{w}_k^{(t+1)}, \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p} \rangle \leq c_\eta \cdot (1+\sigma) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.$$

引理3. 对于幸运神经元 $k \in \mathcal{U}(t)$，令 $\boldsymbol{u}_k^{(t+1)}$ 为算法*1*返回的下一次迭代。则神经元权重满足以下不等式:

1. 沿着 $\boldsymbol{p}_-$ 的方向，我们有

$$\langle \boldsymbol{u}_k^{(t+1)}, \boldsymbol{p}_- \rangle \geq \langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p}_- \rangle + c_\eta \left(\alpha - \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \right);$$

2. 沿着 $\boldsymbol{p}_+$ 的方向，或对于类别无关模式，使得对于任意 $\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_-$，我们有

$$\langle \boldsymbol{u}_k^{(t+1)}, \boldsymbol{p} \rangle - \langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p} \rangle \geq -c_\eta - \sigma - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}},$$

and

$$\langle \boldsymbol{u}_k^{(t+1)}, \boldsymbol{p} \rangle - \langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p} \rangle \leq c_\eta \cdot (1+\sigma) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.$$

引理 4 和 5 展示了不幸神经元的权重沿着类别相关模式和类别无关模式方向的更新。

表 3：标量和矩阵的重要记号

π	圆的周长与直径之比这一数学常数
d	输入特征的维度
K	隐藏层中的神经元数量
$\boldsymbol{x}_v,\ v \in \mathcal{V}$	节点 v 在 $\mathbb{R}^d$ 中的输入特征
$\boldsymbol{p}_v,\ v \in \mathcal{V}$	节点 v 在 $\mathbb{R}^d$ 中的无噪声输入特征
$\boldsymbol{z}_v, v \in \mathcal{V}$	节点 v 在 $\mathbb{R}^d$中的噪声因子
$y_v,\ v \in \mathcal{V}$	节点 v 在 $\{-1,+1\}$中的标签
R	原始图的度 $\mathcal{G}$
$\boldsymbol{W},\ \boldsymbol{U}$	隐藏层中的神经元权重
$\boldsymbol{X}_{\mathcal{N}}$	$\{\boldsymbol{x}_n\}_{n\in\mathcal{N}}$ 在 $\mathbb{R}^{d\times\lvert\mathcal{N}\rvert}$中的集合
r	采样邻居节点的数量
L	类别相关特征和类别无关特征的数量
$\boldsymbol{p}_+$	与正标签对应的类别相关模式
$\boldsymbol{p}_-$	与负标签对应的类别相关模式
$\boldsymbol{z}$	输入特征中的加性噪声
σ	$\lVert\boldsymbol{z}\rVert_2$ 噪声因子的上界 $\boldsymbol{z}$
$\boldsymbol{M}_+$	剪枝模型的 $\boldsymbol{W}$ 掩码矩阵
$\boldsymbol{M}_-$	剪枝模型的 $\boldsymbol{U}$ 掩码矩阵

引理4。 对于不幸神经元 $k \in \mathcal{W}^c(t)$，设 $\boldsymbol{w}_k^{(t+1)}$ 为算法*1*返回的下一次迭代。则该神经元的权重满足以下不等式。

1. 沿 $\boldsymbol{p}_+$方向，我们有

$$\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p}_+ \,\rangle \geq \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}_+ \,\rangle - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}};$$

2. 沿 $\boldsymbol{p}_-$方向，我们有

$$\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p}_- \,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}_- \,\rangle \geq -c_\eta - \sigma - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}},$$

and

$$\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p}_- \,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}_- \,\rangle \leq c_\eta \cdot \sigma \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}};$$

3. 沿类别无关模式的方向，使得任意 $\boldsymbol{p} \in \mathcal{P}_N$，我们有

$$\left|\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p} \,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p} \,\rangle\right| \leq c_\eta \cdot (1+\sigma)\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.$$

引理 5. 对于不幸神经元 $k \in \mathcal{U}^c$（t），设 $\boldsymbol{u}_k^{(t+1)}$ 为算法*1*返回的下一次迭代。则，神经元权重满足以下不等式。

1. 在 $\boldsymbol{p}_-$的方向上，我们有

$$\langle\, \boldsymbol{u}_k^{(t+1)}\,,\, \boldsymbol{p}_- \,\rangle \geq \langle\, \boldsymbol{u}_k^{(t)}\,,\, \boldsymbol{p}_- \,\rangle - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}};$$

2. 在 $\boldsymbol{p}_+$的方向上，我们有

$$\langle\, \boldsymbol{u}_k^{(t+1)}\,,\, \boldsymbol{p}_+\,\rangle - \langle\, \boldsymbol{u}_k^{(t)}\,,\, \boldsymbol{p}_+\,\rangle \geq -c_\eta - \sigma - c_\eta \cdot \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}},$$

and

$$\langle\, \boldsymbol{u}_k^{(t+1)}\,,\, \boldsymbol{p}_+\,\rangle - \langle\, \boldsymbol{u}_k^{(t)}\,,\, \boldsymbol{p}_+\,\rangle \leq c_\eta \cdot \sigma \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}};$$

3. 对于类别无关模式的方向，只要任意 $\boldsymbol{p} \in \mathcal{P}_N$，我们有

$$\big|\langle\, \boldsymbol{u}_k^{(t+1)}\,,\, \boldsymbol{p}\,\rangle - \langle\, \boldsymbol{u}_k^{(t)}\,,\, \boldsymbol{p}\,\rangle\big| \leq c_\eta \cdot (1+\sigma) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.$$

Le引理 6 表明，初始化时的幸运神经元在迭代过程中仍然是幸运神经元，且 d
the 幸运神经元的数量至少与初始化时相同。

引理 6. 设 $\mathcal{W}(t)$、$\mathcal{U}(t)$ 为(26)中第 t 次迭代时的幸运神经元集合。则我们有 *e*

$$\mathcal{W}(t) \subseteq \mathcal{W}(t+1), \qquad \mathcal{U}(t) \subseteq \mathcal{U}(t+1) \tag{33}$$

若样本数量 $|\mathcal{D}| \gtrsim \alpha^{-2}(1+r^2)\log\, q.$

引理 7 表明，某些幸运神经元的幅值始终大于所有不幸神经元的幅值。

引理 7. 设 $\{\boldsymbol{W}_{(t')}\}_{t'=1}^{T'}$ 为剪枝前算法*1*返回的迭代结果。令 $\mathcal{W}^c(t')$为第 t' 次迭代中的不幸神经元集合，则有

$$\langle\, \boldsymbol{w}_{k_1}^{(t')}\,,\, \boldsymbol{w}_{k_1}^{(t')}\,\rangle < \langle\, \boldsymbol{w}_{k_2}^{(t')}\,,\, \boldsymbol{w}_{k_2}^{(t')}\,\rangle \tag{34}$$

对任意 $k_1 \in \mathcal{W}(t')$和 $k_2 \in \mathcal{W}^c(0)$，均有。

引理 8 给出了部分依赖随机变量的矩生成函数界，可用于刻画图结构数据的 Chernoff 界。

引理 8 (引理 7，Zhang 等人 (2020a))。 给定一个包含 N 个部分依赖但同分布随机变量的 $\mathcal{X} = \{x_n\}_{n=1}^N$。对于每个 $n \in [N]$，假设 x_n 至多与 $\mathcal{X}$ 中的 $d_{\mathcal{X}}$ 个随机变量相关（包括 x_n 自身），且 x_n 的矩生成函数满足 $\mathbb{E}_{x_n} e^{s x_n} \leq e^{C s^2}$，其中常数 C 可依赖于 x_n 的分布。那么，$\sum_{n=1}^N x_n$ 的矩生成函数有如下界

$$\mathbb{E}_{\mathcal{X}} e^{s \sum_{n=1}^N x_n} \leq e^{C d_{\mathcal{X}} N s^2}. \tag{35}$$

E 主定理的证明

接下来，我们给出主定理（定理 2）的正式表述及其证明。

定理 2. 设 $\varepsilon_N \in (0,1)$ 和 $\varepsilon_K \in (0, 1 - \frac{\sigma L}{\pi})$ 为某些正数常量。则假设训练样本数量满足

$$|\mathcal{D}| \gtrsim \varepsilon_N^{-2} \cdot \alpha^{-2} \cdot (1-\beta)^2 \cdot (1+\sigma)^2 \cdot \big(1 - \varepsilon_K - \frac{\sigma L}{\pi}\big)^{-2} \cdot (1+r^2) \cdot L^2 \cdot \log q, \tag{36}$$

且神经元数量满足

$$K \gtrsim \varepsilon_K^{-2} L^2 \log q, \tag{37}$$

其中某个正数常量为 q。则经过满足以下条件的 T 次迭代后

$$T \gtrsim \frac{c_\eta \cdot (1-\beta) \cdot L}{\alpha \cdot (1 - \varepsilon_N - \sigma) \cdot (1 - \varepsilon_K - \sigma L/\pi)}, \tag{38}$$

式(3) 中的泛化误差函数满足

$$I\big(g(\boldsymbol{W}^{(T)}, \boldsymbol{U}^{(T)})\big) = 0 \tag{39}$$

以至少为 $1 - q^{-C}$ 的概率，其中 $C > 0$ 为某个常数。

定理 *2* 的证明。令 $\mathcal{K}_{\beta+}$ 和 $\mathcal{K}_{\beta-}$ 分别为关于 $\boldsymbol{M}_+ \odot \boldsymbol{W}$ 和 $\boldsymbol{M}_- \odot \boldsymbol{U}$ 的神经元索引。那么，对于任意标签为 $y_v = 1$ 的节点 $v \in \mathcal{V}$，我们有

$$
\begin{aligned}
& g(\boldsymbol{M}_+ \odot \boldsymbol{W}^{(t)}, \boldsymbol{M}_- \odot \boldsymbol{U}^{(t)}; v) \\
=& \frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{K}_{\beta+}} \max_{n \in \mathcal{N}(v)} \phi\big(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{x}_n \rangle\big) - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} \phi\big(\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{x}_n \rangle\big) \\
\geq& \frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{K}_{\beta+}} \max_{n \in \mathcal{N}(v)} \phi\big(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{x}_n \rangle\big) \\
& - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p}_n \rangle| - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{z}_n \rangle| \\
=& \frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{W}(t)} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ + \boldsymbol{z}_n \rangle + \frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{K}_{\beta+}/\mathcal{W}(t)} \max_{n \in \mathcal{N}(v)} \phi\big(\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{x}_n \rangle\big) \\
& - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p}_n \rangle| - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{z}_n \rangle| \\
\geq& \frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{W}(t)} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ + \boldsymbol{z}_n \rangle \\
& - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p}_n \rangle| - \frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{z}_n \rangle|,
\end{aligned}
\tag{40}
$$

其中， $\mathcal{W}(t)$ 是第 t 次迭代时的幸运神经元集合。

于是，我们有

$$
\begin{aligned}
\frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{W}(t)} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ + \boldsymbol{z}_n \rangle &\geq \frac{1-\sigma}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{W}(0)} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ \rangle \\
&\geq \frac{1-\sigma}{|\mathcal{K}_{\beta+}|} \cdot |\mathcal{W}(0)| \cdot c_\eta \cdot \Big(\alpha - \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\Big) \cdot t \\
&\gtrsim \frac{1}{|\mathcal{K}_{\beta+}|} \cdot |\mathcal{W}(0)| \cdot c_\eta \cdot \alpha \cdot t
\end{aligned}
\tag{41}
$$

其中，第一个不等式由引理 6 得出，第二个不等式由引理 2 得出。一方面，根据引理 2，我们知道，在 $\{w_k^{(0)}\}_{k=1}^K$ 个神经元中，至少有 $(1 - \varepsilon_K - \frac{\sigma L}{\pi})K$ 个是幸运神经元。另一方面，我们知道，剪枝前 $\mathcal{W}(0)$ 中神经元的幅值始终大于所有其他神经元。因此，这些神经元在基于幅值的剪枝后不会被剪枝。因此，我们有

$$
|\mathcal{W}(0)| = (1 - \varepsilon_K - \frac{\sigma L}{\pi})K = (1 - \varepsilon_K - \frac{\sigma L}{\pi})|\mathcal{K}^+|/(1-\beta) \tag{42}
$$

因此，我们有

$$
\frac{1}{|\mathcal{K}_{\beta+}|} \sum_{k \in \mathcal{W}(t)} \langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ \rangle \gtrsim (1 - \varepsilon_K - \frac{\sigma L}{\pi}) \cdot \frac{1}{(1-\beta)L} \cdot \alpha \cdot t. \tag{43}
$$

In a此外，当 $y_v = 1$ 时， $\boldsymbol{X}_{\mathcal{N}(v)}$ 不包含 $\boldsymbol{p}_-$ 。于是，根据引理 3 和引理 5，我们有

$$
\begin{aligned}
\frac{1}{|\mathcal{K}_{\beta-}|} \sum_{k \in \mathcal{K}_{\beta-}} \max_{n \in \mathcal{N}(v)} |\langle \boldsymbol{u}_k^{(t)}, \boldsymbol{p}_n \rangle| &\lesssim c_\eta (1+\sigma) \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} t \\
&\lesssim \varepsilon_N \cdot (1 - \varepsilon_K - \frac{\sigma L}{\pi}) \cdot \alpha \cdot \frac{1}{(1-\beta)L} t,
\end{aligned}
\tag{44}
$$

其中，最后一个不等式由式（36）得出。

此外，我们有

$$\begin{aligned}\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{K}_{\beta-}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(t)}\,,\,\boldsymbol{z}_n\,\rangle| &\le \mathbb{E}_{k\in\mathcal{K}_{\beta-}}\|\boldsymbol{u}_k^{(t)}\|_2\cdot\|\boldsymbol{z}_n\|_2\\ &\le \sigma\cdot\mathbb{E}_{k\in\mathcal{K}_{\beta-}}\sum_{\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_-}\langle\,\boldsymbol{u}_k^{(t)}\,,\,\boldsymbol{p}\,\rangle\\ &\lesssim \sigma\cdot c_\eta\cdot L(1+\sigma)\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\cdot t.\end{aligned}\tag{45}$$

结合 (43)、(44) 和 (45)，我们得到

$$g(\boldsymbol{W}^{(T)},\boldsymbol{U}^{(T)};v)\ge(1-\varepsilon_N)(1-\varepsilon_K-\sigma L/\pi)(1-\sigma L)\frac{\alpha}{(1-\beta)L}\cdot c_\eta\cdot T.\tag{46}$$

因此，当迭代次数满足

$$T>\frac{c_\eta\cdot(1-\beta)\cdot L}{\alpha\cdot(1-\varepsilon_N-\sigma)\cdot(1-\varepsilon_K-\sigma L/\pi)\cdot(1-\sigma L)},\tag{47}$$

我们有 $g(\boldsymbol{W}^{(t)}\boldsymbol{U}^{(t)};\,v)\,>1$。

与 (46) 的证明类似，对于任意带标签 $v\in\mathcal{V}$ with label $y_v=-1$，我们有

$$\begin{aligned}&g(\boldsymbol{M}_+\odot\boldsymbol{W}^{(T)},\boldsymbol{M}_-\odot\boldsymbol{U}^{(T)};v)\\ =&\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}\phi\big(\langle\,\boldsymbol{w}_k^{(T)}\,,\,\boldsymbol{x}_n\,\rangle\big)-\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{K}_{\beta-}}\max_{n\in\mathcal{N}(v)}\phi\big(\langle\,\boldsymbol{u}_k^{(T)}\,,\,\boldsymbol{x}_n\,\rangle\big)\\ \le&-\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{K}_{\beta-}}\max_{n\in\mathcal{N}(v)}\phi\big(\langle\,\boldsymbol{w}_k^{(T)}\,,\,\boldsymbol{x}_n\,\rangle\big)\\ &+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(T)}\,,\,\boldsymbol{p}_n\,\rangle|+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta-}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(T)}\,,\,\boldsymbol{z}_n\,\rangle|\\ =&-\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{W}(T)}\langle\,\boldsymbol{w}_k^{(T)}\,,\,\boldsymbol{p}_-+\boldsymbol{z}_n\,\rangle-\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{K}_{\beta-}/\mathcal{W}(t)}\max_{n\in\mathcal{N}(v)}\phi\big(\langle\,\boldsymbol{w}_k^{(T)}\,,\,\boldsymbol{x}_n\,\rangle\big)\\ &+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(T)}\,,\,\boldsymbol{p}_n\,\rangle|+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(T)}\,,\,\boldsymbol{z}_n\,\rangle|\\ \ge&-\frac{1}{|\mathcal{K}_{\beta-}|}\sum_{k\in\mathcal{W}(T)}\langle\,\boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-+\boldsymbol{z}_n\,\rangle\\ &+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(t)}\,,\,\boldsymbol{p}_n\,\rangle|+\frac{1}{|\mathcal{K}_{\beta+}|}\sum_{k\in\mathcal{K}_{\beta+}}\max_{n\in\mathcal{N}(v)}|\langle\,\boldsymbol{u}_k^{(t)}\,,\,\boldsymbol{z}_n\,\rangle|,\\ \le&-\frac{1}{|\mathcal{K}_{\beta-}|}|\mathcal{U}(T)|\cdot c_\eta\cdot\alpha T\\ &+c_\eta\cdot(1+\sigma)\cdot T\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}+c_\eta\cdot\sigma\cdot(1+\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\cdot T\\ \lesssim&-(1-\varepsilon_N)(1-\varepsilon_K-\sigma L/\pi)\cdot(1-\sigma L)\cdot\frac{\alpha}{(1-\beta)L}\\ \le&-1.\end{aligned}\tag{48}$$

因此，对于任意带标签 v with label $y_v=-1$，我们有 $g(\boldsymbol{W}^{(t)}\boldsymbol{U}^{(t)};\,v)\,<-1$。

总之，当条件 (36)、(37) 和 (38) 成立时，式 (3) 中的泛化函数取值为零。

□

F 数值实验

F.1 实验实现

合成图结构数据的生成。第4节中使用的合成数据按以下方式生成。首先，我们随机生成三组节点，分别记为 $\mathcal{V}_+$、$\mathcal{V}_-$和 $\mathcal{V}_N$。为 $\mathcal{V}_+$ 中的节点赋予带噪声的 $\boldsymbol{p}_+$，为 $\mathcal{V}_-$ 中的节点赋予带噪声的 $\boldsymbol{p}_-$。$\mathcal{V}_N$ 中节点的模式是类别无关模式。对于 $\mathcal{V}_N$中的任意节点 v， v 将均匀地连接到 $\mathcal{V}_+$ 或 $\mathcal{V}_-$ 中的一些节点。如果该节点连接到 $\mathcal{V}_+$中的节点，则其标签为 $+1$。否则，其标签为 -1。最后，我们将在 $\mathcal{V}_+$、 $\mathcal{V}_-$ 或 $\mathcal{V}_N$中的节点之间添加随机连接。为验证我们的定理， $\mathcal{V}_N$ 中的每个节点将恰好连接到 $\mathcal{V}_+$ 或 $\mathcal{V}_-$中的一个节点，并通过从 $\mathcal{V}_N$中随机选择 $M-1$ 个其他节点，使 $\mathcal{V}_N$ 中节点的度恰好为 M。我们对合成数据使用全批量梯度下降。如果训练误差变为零，或达到最大迭代次数500，算法1就会终止。除非另有说明，$\delta=0.1$、 $c_\eta=1$、 $r=20$、 $\alpha=r/R$、 $\beta=0.2$、 $d=50$、 $L=200$、 $\sigma=0.2$和 $|\mathcal{D}|=100$，其余节点作为测试数据。

边上的重要性采样的实现。 给定重要性边的采样率 α 以及采样邻居节点的数量 r。首先，以 α 的采样率为每个节点采样重要性边。然后，以无放回方式随机采样剩余的边，直到采样节点的数量达到 r。

模型权重幅值剪枝的实现。 剪枝算法完全遵循算法1中的伪代码。迭代次数 T' 设定为5。

误差条示意图。 图中，低透明度区域表示误差条。误差条的上包络基于均值加一个标准差的值，下包络基于均值减一个标准差的值。

F.2 数据模型假设的经验性论证

在本部分中，我们将以 Cora 数据集为例，说明我们的数据假设能够模拟一些真实的应用场景。

Cora 数据集是一个包含 2708 个节点的引文网络，节点分属七个类别，分别为“神经网络”“规则学习”“强化学习”“概率方法”“理论”“遗传算法”和“基于案例”。为便于表述，我们将上述标签依次记为“类别1”至“类别7”。对于每个节点，我们通过聚合其一跳邻居节点来计算聚合特征向量，每个特征向量的维度为 1433。接着，我们收集标签相同节点的聚合特征向量，构建矩阵，并在表 4 中列出各类别对应矩阵的最大奇异值。随后，我们计算不同类别对应的第一主成分之间的余弦相似度，即 $\arccos \frac{<z_1,\ z_2>}{\|z_1\|_2\cdot\|z_2\|_2}$。这里的第一主成分是具有最大奇异值的右奇异向量。结果汇总于表 5。

表 4：类别汇集特征矩阵中最大的 5 个奇异值（SV）

标签	第一个奇异值	第二个奇异值	第三个奇异值	第四个奇异值	第五个奇异值
类别1	8089.3	101.6	48.1	46.7	42.5
类别2	4451.2	48.7	28.0	26.7	22.4
类别3	4634.5	53.1	29.4	28.1	24.1
类别4	6002.5	72.9	37.0	35.6	31.4
类别5	5597.8	67.4	33.3	33.1	29.1
类别6	5947.9	72.1	36.7	35.5	31.3
类别7	5084.6	62.0	32.3	30.7	27.2

表 5：类别之间的余弦相似度

	类别1	类别2	类别3	类别4	类别5	类别6	类别7
类别1	0	89.3	88.4	89.6	87.4	89.4	89.5
类别2	89.3	0	89.1	89.5	88.7	89.7	88.7
类别3	88.4	89.1	0	89.7	89.9	89.0	89.1
类别4	89.6	89.5	89.7	0	88.0	88.9	88.7
类别5	87.4	88.7	89.9	88.0	0	89.8	88.2
类别6	89.4	89.7	89.0	88.9	89.8	0	89.5
类别7	89.5	88.7	89.1	88.7	88.3	89.5	0

从表 4 可以看出，收集到的特征矩阵均近似为秩一矩阵。此外，从表 5 可以看出，不同类别的第一主成分几乎相互正交。例如，对应“类别1”“类别2”和“类别3”的特征矩阵之间的两两角度分别为 89.3、88.4 和 89.2。表 4 表明，节点特征高度集中于一个方向；表 5 表明，不同类别的主成分几乎相互正交且彼此独立。因此，我们可以将特征矩阵的主方向视为类别相关特征。如果节点经常连接到两个类别的类别相关特征，那么特征矩阵至少会有两个主方向，从而使矩阵的秩达到二或更高。如果某一类别中的节点经常连接到两个类别的类别相关特征，那么该类别的第一主成分将是至少两个类别相关特征的混合，不同类别的主成分之间也就不可能几乎相互正交。因此，我们可以得出结论：不同类别的大多数节点连接到不同的类别相关特征，并且这些类别相关特征几乎相互正交。

此外，沿用上述相同的实验设置，我们在大规模数据集 Ogbn-Arxiv 上实施了另一项数值实验，以进一步验证我们的数据模型。前10个类别的估计类别相关特征之间的余弦相似度汇总于表 6。可以看到，大多数角度在65到90之间，这表明不同类别的类别相关模式之间存在足够大的距离。此外，为了验证 V_+(V_-) 和 V_{N+}(V_{N-}) 中节点特征的存在，我们比较了同一类别的节点特征和估计的类别相关特征。图 12 展示了同一类别的节点特征与估计的类别相关特征之间的余弦相似度。可以看到，角度小于40的节点可视为 V_+(V_-)，其他节点则可视为 V_{N+}(V_{N-})，这验证了类别相关模式和类别无关模式的存在。通过对节点特征进行聚类，我们也观察到了与我们类似的结果，即节点嵌入分布在一个较小的空间中（Pan 等，2018）。

F.3 合成数据上的补充实验

图 13 展示了样本复杂度随 K 变化时的相变。样本复杂度几乎是 $1/K$ 的线性函数。图 14 表明，样本复杂度随 σ^2 线性增长，其中 σ^2 是特征中的噪声水平。在图 15 中，α 固定为 0.8，同时增加采样邻居的数量 r。样本复杂度与 r^2 呈线性关系。

图16展示了不同采样边数量及 $R = 30$ 下所需的迭代次数。拟合曲线（以黑色虚线表示）是 $\alpha = r/R$ 时关于 α^{-1} 的线性函数。可以看到，对于 $\alpha = r/R$，拟合曲线与经验结果相符，这验证了式（7）中的界。此外，应用重要性采样后，当 α 较大时，迭代次数会显著减少。

图17展示了不同剪枝率下收敛所需的迭代次数 β。所有结果均取自100次独立试验的平均值。黑色虚线代表基线，即训练原始稠密网络所需的平均迭代次数。带圆圈标记的蓝线表示对神经元权重进行幅值剪枝的表现。迭代次数几乎是剪枝率的线性函数，这验证了我们在式（7）中的理论发现。带星形标记的红线

表 6：各类别之间的余弦相似度

	类别1	类别2	类别3	类别4	类别5	类别6	类别7	类别8	类别9	类别10
类别1	0	84.72	77.39	88.61	86.81	89.88	87.38	86.06	83.42	72.54
类别2	84.72	0	59.15	69.91	54.85	48.86	58.39	72.35	67.11	57.93
类别3	77.38	59.15	0	67.62	63.26	46.55	54.32	66.43	57.17	45.48
类别4	88.61	69.91	67.62	0	75.27	73.23	33.58	68.16	38.80	84.71
类别5	86.81	54.85	63.26	75.27	0	35.00	66.68	52.44	72.37	65.64
类别6	71.22	70.83	65.44	66.64	61.99	0.00	67.58	71.60	64.31	67.88
类别7	71.90	68.03	68.43	69.68	67.64	67.58	0.00	68.24	70.38	60.50
类别8	61.19	70.86	70.43	62.79	63.24	71.60	68.24	0.00	64.58	70.65
类别9	62.25	70.73	64.69	71.03	71.70	64.31	70.38	64.58	0.00	71.88
类别10	69.69	67.38	68.49	64.92	67.17	67.88	60.50	70.65	71.88	0.00

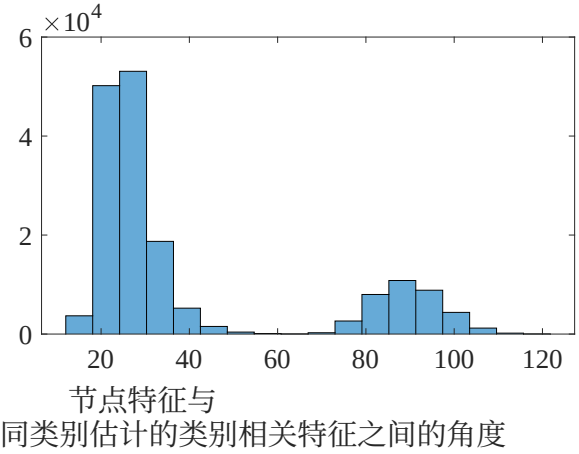


图 12：节点特征与从同一类别估计的类别相关特征之间的余弦相似度分布

表示未回溯至原始初始化时的收敛情况，其结果差于基线，因为剪枝后的随机初始化会导致模型的过参数化程度降低。

图18展示了对1000次独立试验取平均后，不同样本数量下的测试误差。带圆圈标记的红线显示了训练原始稠密模型的表现。带星形标记的蓝线对应幅值剪枝后的模型，其测试误差低于训练原始模型时的测试误差。使用随机剪枝进行的补充实验以带菱形标记的黑线概括，其测试误差始终高于训练原始模型时的测试误差。

此外，我们在 Cora 数据上验证了理论刻画。与合成数据相比，α 并非边的采样率，而是类别相关特征的采样率；对于真实数据集而言，这一采样率是未知的。此外，目前没有标准方法来定义实际数据训练实验的成功。因此，我们对实验做了一些修改，使其符合我们的理论框架。我们利用附录 F.2 中估计的类别相关特征来判断节点是否与类别相关；也就是说，如果节点特征与类别相关特征之间的角度小于 30 度，我们就将该节点特征称为类别相关特征。对于合成数据，如果一次试验达到零泛化误差，我们就将其定义为成功。相较之下，对于实际数据，若测试准确率大于 80%，我们就认为试验成功。图 19 展示了样本复杂度随 α^{-2} 的变化情况，短语转变曲线几乎是一条直线，这验证了我们的

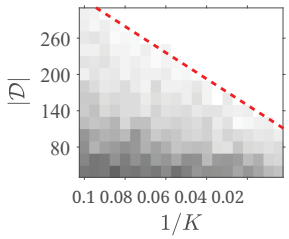


图 13： $|\mathcal{D}|$ 与神经元数量的关系 K

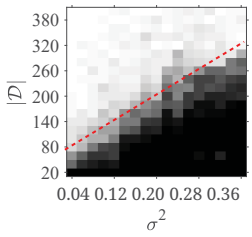


图 14： $|\mathcal{D}|$ 与噪声水平的关系 σ^2。

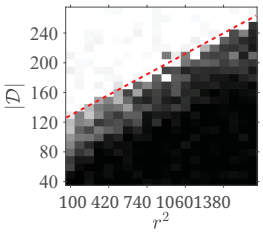


图 15： $|\mathcal{D}|$ 与采样邻居数量的关系 r

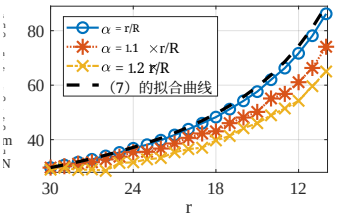


图16：随机采样与重要性采样下，迭代次数随边采样率的变化。

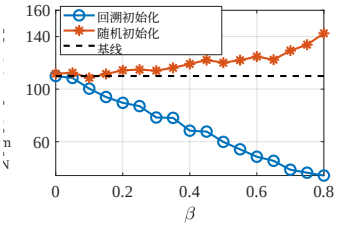


图17：迭代次数随模型权重剪枝率的变化。

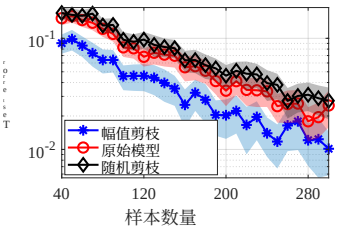


图18：测试误差随样本数量的变化。

式（6）中的理论结果。图20和图21展示了在 Cora测试集上训练时前 50 个训练轮次的收敛速率。可以看到，正如式（6）中的理论结果所示，收敛速率是剪枝率和 α^{-1} 的线性函数。

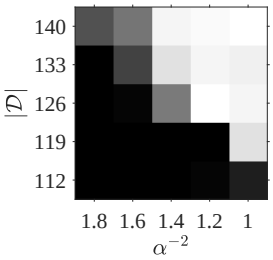


图19： $\mathcal{D}$ 随估计的重要性采样概率 α的变化。

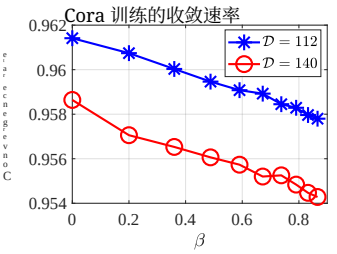


图20： $\mathcal{D}$ 随剪枝率 β的变化。

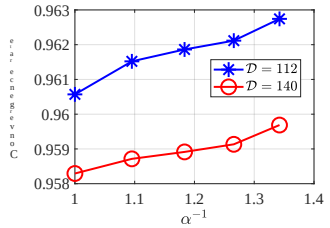


图21：收敛速率随估计的重要性采样概率 α的变化。

F.4 放宽数据假设的补充实验

在以下数值实验中，我们通过在 (1) V_+ 与 $V_{\mathcal{N}-}$、(2) V_- 与 $V_{\mathcal{N}_+}$ 以及 (3) V_+ 与 V_- 之间添加额外的边，放宽假设（A1）。我们随机选择 γ 比例的节点：如果其标签为正（或负），则将其连接到 $\mathcal{V}_-$（或 $\mathcal{V}_+$）中的一个节点。我们将选出的节点称为“异常值节点”，其余节点称为“干净节点”。请注意，集合“异常值节点”中的两个相同节点可能具有不同的标签，这意味着无法找到从“异常值节点”到其标签的任何映射。因此，我们仅在这些“干净节点”上评估泛化性能，但使用“干净节点”和“异常值节点”的混合数据训练图神经网络。

图23展示了当 γ 变化时样本复杂度的相变。采样边的数量 $r = 20$、剪枝率 $\beta = 0.2$、数据维度 $d = 50$以及模式的数量$L = 200$。我们可以看到，当 γ 小于0.3时，样本复杂度几乎保持不变，这表明即使放宽假设（A1），我们的理论洞见仍然成立。

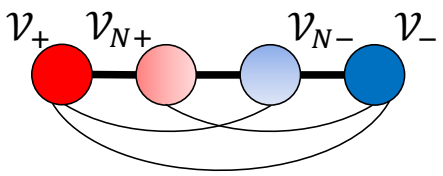


图22：数据模型的玩具示例。节点1和2的标签为 $+1$。节点3和4的标签为 -1。节点1和4具有类别相关特征。节点2和3具有类别无关特征。 $\mathcal{V}_+ = \{1\}$, $\mathcal{V}_{N+} = \{2\}$, $\mathcal{V}_{N-} = \{3\}$,$\mathcal{V}_- = \{4\}$。

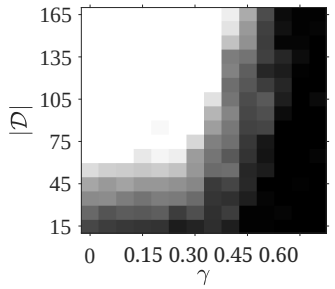


图23：训练节点数量 $|\mathcal{D}|$与 γ。

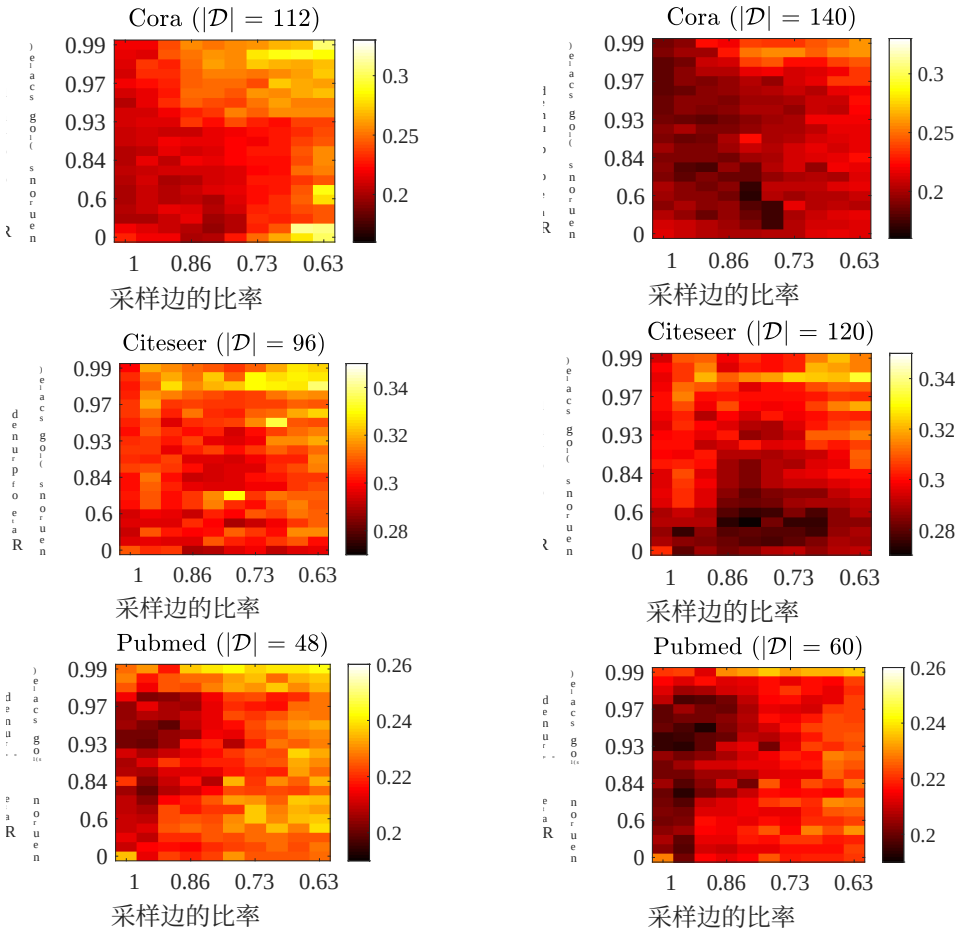


图24：展示不同带标签节点数量下 Cora（第一行子图）、Citeseer（第二行子图）和 Pubmed（第三行子图）测试误差的热力图。

F.5 小规模真实数据上的补充实验

在图 26–25 中，我们将边采样方法与基于幅度的神经元剪枝方法结合实施。在图 26 和图 24 中，Chen 等人（2021b）提出的统一图神经网络稀疏化（*UGS*）被用作边采样方法[4]，此处使用的图神经网络模型是标准图

[4]实验使用 https://github.com/VITA-Group/ Unified-LTH-GNN 中的代码实现。

图卷积神经网络（GCN）Kipf 和 Welling（2017）（两层消息传递）。在图25中，我们将采用池化聚合作为采样方法的 GraphSAGE 实现为 [5]。除图24外，我们使用 140 个（Cora）、120 个（Citeseer）和 60 个（Pubmed）带标签数据进行训练，使用 500 个节点进行验证，并使用 1000 个节点进行测试。在图24中，训练数据的数量标注在每个子图的标题中。

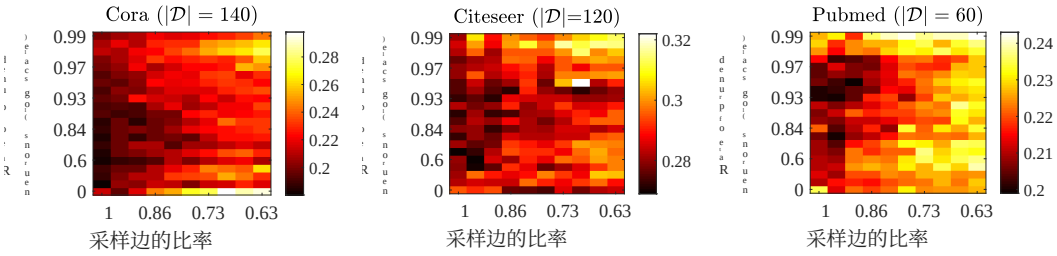


图25：联合 GraphSAGE-池化与模型剪枝的节点分类性能。

图26显示，使用 UGS 进行边采样以及基于幅值的神经元剪枝，都可以降低 Cora、Citeseer 和 Pubmed 数据集上的测试误差，这验证了我们的理论发现，即联合稀疏化能够提升泛化能力。相比之下，随机剪枝会降低性能，其测试误差大于基线。

图24和图25展示了在不同采样率和剪枝率下，Cora、Citeseer 和 Pubmed 数据集上的测试误差，颜色越深表示误差越低。在所有子图中，我们都可以观察到，联合边采样与剪枝能够降低测试误差，这验证了联合边-模型稀疏化的有效性。对于图24，通过比较同一数据集上使用不同训练样本时联合稀疏化的性能，可以得出结论：与不进行稀疏化相比，使用更少训练样本的联合模型-边稀疏化能够实现相近甚至更好的性能。

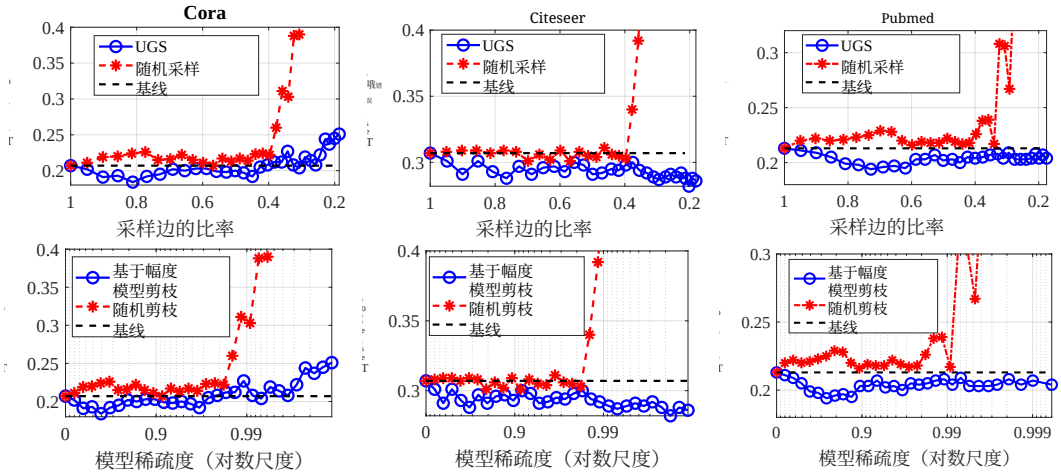


图26：图卷积网络在 Cora（第一列子图）、Citeseer（第二列子图）和 Pubmed（第三列子图）上的节点分类性能。

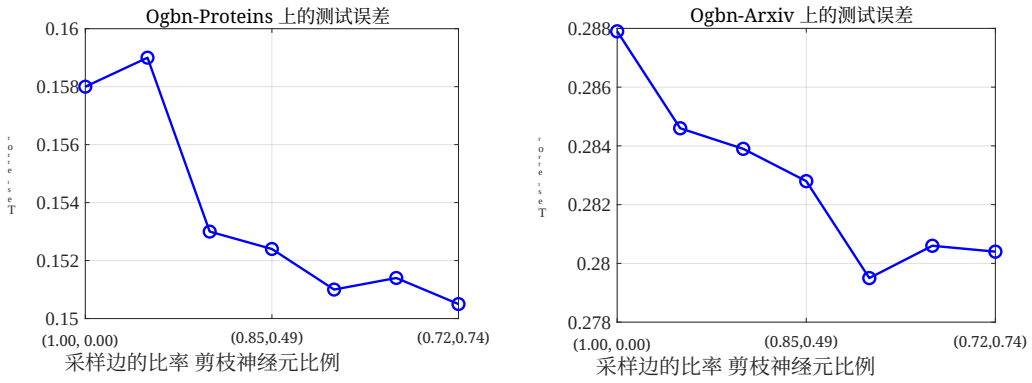
F.6 大规模真实数据上的补充实验

在这一部分，我们通过一个 28 层 ResGCN，在名为 Ogbn-Proteins Hu et al. (2020) 的大规模蛋白质数据集和名为 Obgn-Arxiv Wang et al. (2020) 的引文数据集上开展了实验。实验采用与 Chen 等人（2021b）相同的设置，将统一图神经网络稀疏化（UGS）和基于幅值的神经元剪枝方法作为训练

[5]实验使用 https://github.com/williamleif/ GraphSAGE 中的代码实现

算法。Ogbn-proteins 数据集是一个包含 132,534 个节点和 39,561,252 条边的无向图。节点表示蛋白质，边表示蛋白质之间不同类型的、具有生物学意义的关联。每条边都附带 8 维特征，其中每一维表示某一种关联类型的近似置信度，取值范围为 0 到 1。
Ogbn-Arxiv 数据集是一个包含 169,343 个节点和 1,166,243 条边的引文网络。每个节点是一篇 arXiv 论文，每条边表示一篇论文引用了另一篇论文。每个节点都附带一个 128 维特征向量，该向量由论文标题和摘要中词语的嵌入均值计算得到。

图 27 总结了使用联合边-模型稀疏化在上述数据集上训练 ResGCN 的测试误差。我们可以看到，联合模型稀疏化能够降低泛化误差，这印证了我们的理论洞见：联合稀疏化可以降低样本复杂度。



(a) Ogbn-Proteins。　　(b) Ogbn-Arxiv。

图 27：在 (a) Ogbn-Proteins 和 (b) Ogbn-Arxiv 上的测试误差。

G 图结构数据和两层 GNN 的 VC 维数

按照 Scarselli 等人（2018）的框架，我们将输入数据定义为 $(\mathcal{G}_v,\ v)$，其中 $\mathcal{G}_v$ 是关于节点 v 的子图。本文考虑的输入数据分布记为 $\mathcal{G}\times\mathcal{V}$。为简化分析，对于任意 $(\mathcal{G}_v,\ v)\in\mathcal{V}$，$\mathcal{N}(v)$ 恰好包含来自 $\{\boldsymbol{p}_+,\boldsymbol{p}_-\}$ 的一个特征，而所有其他节点特征均来自 $\mathcal{P}_N$。具体而言，我们总结如下图神经网络模型的 VC 维数。

定义 1（第4节 4.1，Scarselli 等人（2018））。设 g 为图神经网络模型，数据集 $\mathcal{D}$ 来自 $\mathcal{G}\times\mathcal{V}$。若对于任意一组二元赋值 $\boldsymbol{y}\in\{0,1\}^{|\mathcal{D}|}$，都存在参数$\boldsymbol{\Theta}$，使得 g $\mathcal{D}$ 可以被打散，则称 g 能打散 $\mathcal{D}$；具体而言，当 $y_v=1$ 时，$g(\boldsymbol{\Theta};\ (\mathcal{G}_v,v))>0$，而当 $y_v=0$ 时，$g(\boldsymbol{\Theta};\ (\mathcal{G}_v,v))<0$。随后，*GNN* 的 *VC* 维数定义为可被打散的最大集合的大小，即

$$VC\text{-}dim(g)=\max_{\mathcal{D}\ is\ shattered\ by\ g}|\mathcal{D}|. \tag{49}$$

基于上述定义，我们推导出本文所考虑的数据分布上图神经网络模型的 VC 维度，并在定理 3 中进行了总结。定理 3 表明，该数据分布上图神经网络模型的 VC 维度至少为 $2^{L/2-1}$，这是关于 L 的指数函数。受 Brutzkus & Globerson（2021）关于卷积神经网络的研究启发，主要思路是构造一个神经网络能够打散的特殊数据集。在这里，我们通过构造如式（51）所示的一组图结构数据来扩展这一证明。接着，对于任意标签组合，根据式（54），都可以构造一个可解的线性系统来拟合这些标签。

定理 3（VC 维数）。假设正交特征的数量为 L，则图神经网络模型在数据 $\mathcal{G}\times\mathcal{V}$ 上的 *VC* 维数满足

$$VC\text{-}dim(\mathcal{H}_{GNN}(\mathcal{G}\times\mathcal{V}))\geq 2^{\frac{L}{2}-1}. \tag{50}$$

证明。不失一般性，记与类别无关的特征为 $\{\boldsymbol{p}_i\}_{i=1}^{L-2}$。因此，$\mathcal{P}=\{\boldsymbol{p}_i\}_{i=1}^{L-2}\bigcup\boldsymbol{p}_+\bigcup\boldsymbol{p}_-$。定义集合 $\mathcal{J}$，使得

$$\mathcal{J}=\{\boldsymbol{J}\in\{0,1\}^{L/2-1}\mid J_i=0 \text{ or } 1, 1\le i\le L/2-1\}.$$

接下来，我们基于集合 $\mathcal{J}$定义一个大小为 $2^{L/2-1}$ 的子集 $\mathcal{D}\in\mathcal{G}\times\mathcal{V}$。

对于任意的 $(\mathcal{G}_v,\ v)$，我们考虑与恰好 $L/2-1$ 个邻居相连的图结构数据 $\mathcal{G}_v$。回想一下，对于每个节点 v，特征矩阵 $\boldsymbol{X}_{\mathcal{N}(v)}$ 必须包含 $\boldsymbol{p}_+$ 或 $\boldsymbol{p}_-$，我们随机选取一个节点 $u\in\mathcal{N}(v)$，并令 $\boldsymbol{x}_u$ 为 $\boldsymbol{p}_+$ 或 $\boldsymbol{p}_-$。接着，我们对 $\mathcal{N}(v)$中的所有其他节点排序，并将它们标记为 $\{v_i\}_{i=1}^{L/2-1}$。然后，给定一个元素 $\boldsymbol{J}\in\mathcal{J}$，我们将节点 v_i 的特征设为

$$\boldsymbol{x}_{v_i}=J_i\boldsymbol{p}_{2i-1}+(1-J_i)\boldsymbol{p}_{2i}, \tag{51}$$

其中，J_i是向量 $\boldsymbol{J}$的第 i个元素。我们将这样构造的数据 $(\mathcal{G}_v,\ v)$记为 $D_{\boldsymbol{J}}$。因此，数据集 $\mathcal{D}$ 定义为

$$\mathcal{D}=\{D_{\boldsymbol{J}}\mid\boldsymbol{J}\in\mathcal{J}\}. \tag{52}$$

接下来，我们考虑具有 $K=2^{L/2-1}$的式 (20) 形式的图神经网络模型。给定每个数据 $D_{\boldsymbol{J}}\in\mathcal{D}$对应的标签 $y_{\boldsymbol{J}}$，我们通过求解以下方程组得到一组 $\alpha_{\boldsymbol{J}}\in\mathbb{R}$

$$\sum_{\boldsymbol{J}'\in\mathcal{J}/\boldsymbol{J}^\dagger}\alpha_{\boldsymbol{J}'}=y_{\boldsymbol{J}},\forall\boldsymbol{J}\in\mathcal{J}, \tag{53}$$

其中 $\boldsymbol{J}^\dagger=\mathbf{1}-\boldsymbol{J}$。我们可以看出，(53) 可重写为

$$\begin{bmatrix}0&1&\cdots&1\\1&0&\cdots&1\\\vdots&\vdots&\ddots&\vdots\\1&1&\cdots&0\end{bmatrix}\cdot\begin{bmatrix}\alpha_1\\\alpha_2\\\vdots\\\alpha_{L/2-1}\end{bmatrix}=\begin{bmatrix}y_{1^\dagger}\\y_{2^\dagger}\\\vdots\\y_{(L/2-1)^\dagger}\end{bmatrix} \tag{54}$$

容易验证，(54) 中的矩阵是可逆的。因此，对于任意固定的 $\{y_{\boldsymbol{J}}\}_{\boldsymbol{J}\in\mathcal{J}}$，$\{\alpha_{\boldsymbol{J}}\}_{\boldsymbol{J}\in\mathcal{J}}$ 都有唯一解。

由于将 K 选为 $2^{L/2-1}$，我们通过定义将神经元权重表示为 $\{\boldsymbol{w}_{\boldsymbol{J}}\}_{\boldsymbol{J}\in\mathcal{J}}$ 和 $\{\boldsymbol{u}_{\boldsymbol{J}}\}_{\boldsymbol{J}\in\mathcal{J}}$

$$\begin{aligned}\boldsymbol{w}_{\boldsymbol{J}}&=\max\{\alpha_{\boldsymbol{J}},0\}\cdot\sum_{v'\in\mathcal{N}(v),v\in D_{\boldsymbol{J}}}\boldsymbol{x}_{v'},\\\boldsymbol{u}_{\boldsymbol{J}}&=\max\{-\alpha_{\boldsymbol{J}},0\}\cdot\sum_{v'\in\mathcal{N}(v),v\in D_{\boldsymbol{J}}}\boldsymbol{x}_{v'}.\end{aligned} \tag{55}$$

然后，对于每个 $D_{\boldsymbol{J}}\in\mathcal{D}$，将 (55) 代入 (20)：

$$g=\sum_{\boldsymbol{J}'\in\mathcal{J}}\Big(\max_{v'\in\mathcal{N}(v)}\ \sigma\big(\langle\ \boldsymbol{w}_{\boldsymbol{J}'}\ ,\ \boldsymbol{x}_{v'}\ \rangle\big)-\max_{v'\in\mathcal{N}(v)}\ \sigma\big(\langle\ \boldsymbol{u}_{\boldsymbol{J}'}\ ,\ \boldsymbol{x}_{v'}\ \rangle\big)\Big). \tag{56}$$

当 $\alpha_{\boldsymbol{J}'}\ge0$时，我们有 $\boldsymbol{u}_{\boldsymbol{J}'}=\mathbf{0}$，并且

$$\begin{aligned}&\max_{v'\in\mathcal{N}(v)}\ \sigma\big(\langle\ \boldsymbol{w}_{\boldsymbol{J}'}\ ,\ \boldsymbol{x}_{v'}\ \rangle\big)\\=&\alpha_{\boldsymbol{J}'}\cdot\max\ \sum_{u'\in\mathcal{N}(u),u\in D_{\boldsymbol{J}}}\boldsymbol{x}_{u'}\sum_{v'\in\mathcal{N}(v),v\in D_{\boldsymbol{J}}}\boldsymbol{x}_{v'}\\=&\alpha_{\boldsymbol{J}'}\cdot\max_i\ (J_i\boldsymbol{p}_{2i-1}+(1-J_i)\boldsymbol{p}_{2i})\cdot(J_i'\boldsymbol{p}_{2i-1}+(1-J_i')\boldsymbol{p}_{2i})\\=&\begin{cases}0, & \text{if}\quad\boldsymbol{J}'=\mathbf{1}-\boldsymbol{J}\\\alpha_{\boldsymbol{J}'}, & \text{otherwise}\end{cases},\end{aligned} \tag{57}$$

其中，第二个等式由（51）得出，最后一个等式则源于$\boldsymbol{p} \in \mathcal{P}$的正交性。与（57）类似，当 $\alpha_{\boldsymbol{J}'} < 0$ 时，我们有

$$\begin{aligned}
&-\max_{v' \in \mathcal{N}(v)} \ \sigma\big(\langle\, \boldsymbol{u}_{\boldsymbol{J}'}\,,\, \boldsymbol{x}_{v'}\,\rangle\big) \\
=&-\left(-\alpha_{\boldsymbol{J}'}\right) \cdot \max \sum_{u' \in \mathcal{N}(u), u \in D_{\boldsymbol{J}}} \boldsymbol{x}_{u'} \sum_{v' \in \mathcal{N}(v), v \in D_{\boldsymbol{J}}} \boldsymbol{x}_{v'} \\
=&\alpha_{\boldsymbol{J}'} \cdot \max_i \ \ (J_i \boldsymbol{p}_{2i-1} + (1-J_i)\boldsymbol{p}_{2i}) \cdot (J_i' \boldsymbol{p}_{2i-1} + (1-J_i')\boldsymbol{p}_{2i}) \\
=&\begin{cases} 0, \ \text{if } \boldsymbol{J}' = \mathbf{1} - \boldsymbol{J} \\ \alpha_{\boldsymbol{J}'}, \text{otherwise} \end{cases}.
\end{aligned} \tag{58}$$

因此，(55) 可计算为

$$\begin{aligned}
g &= \sum_{\boldsymbol{J}' \in \mathcal{J}} \Big(\max_{v' \in \mathcal{N}(v)} \ \sigma\big(\langle\, \boldsymbol{w}_{\boldsymbol{J}'}\,,\, \boldsymbol{x}_{v'}\,\rangle\big) - \max_{v' \in \mathcal{N}(v)} \ \sigma\big(\langle\, \boldsymbol{u}_{\boldsymbol{J}'}\,,\, \boldsymbol{x}_{v'}\,\rangle\big)\Big) \\
&= \sum_{\boldsymbol{J}' \in \mathcal{J}/\boldsymbol{J}^\dagger} \alpha_{\boldsymbol{J}'} = y_{\boldsymbol{I}},
\end{aligned} \tag{59}$$

由此完成证明。 □

G.1 各种采样策略下的重要性采样概率

在本节中，我们给出不同采样策略下 α 的界。为简化起见，我们考虑一个每个节点的度均为 R （$R > r$）的图，并为任意节点 $v \in \mathcal{V}$ 固定采样 r 个邻居节点。

引理9给出了均匀采样策略下 α 的取值，并且 α 的阶为 r/R，且取决于包含类别相关模式的节点的平均度。形式化地，我们有

引理9。 对于均匀采样策略，即以相同概率采样所有邻居节点，我们有

$$1 - \left(1 - \frac{\bar{c}}{R}\right)^r \leq \mathbb{E}\alpha \leq 1 - \left(1 - \frac{\bar{c}}{R - \bar{c} + 1}\right)^r, \tag{60}$$

其中， $\bar{c}$ 是 $\mathcal{D} \cap (\mathcal{V}_+ \cup \mathcal{V}_-)$中节点的度。

推论9.1。 对于均匀采样策略，当 $\mathcal{V}_N$ 中的节点恰好连接到$\mathcal{V}_+ \cap \mathcal{V}_-$中的一个节点时， $\mathbb{E}\alpha = \frac{r}{R}$。

推论9.2。 对于均匀采样策略， $\mathbb{E}\alpha \geq \frac{\bar{c}r}{R}$ 当 $r \ll R$时。

备注 9.1：给定 $\alpha = \frac{r}{R}$, (6)对所有 r都给出相同的样本复杂度界，因此使用图采样并不会带来改进。此外，根据(7)可知，所需迭代次数会增加 R/r倍，而每次迭代中采样的边数会减少 r/R倍。考虑到其他步骤中使用的计算资源，我们预计总计算时间会增加。

备注 9.2：由推论9.2可知，当 $r \ll R$ 时，均匀采样平均可将样本复杂度降低 $\frac{1}{\bar{c}^2}$ 倍。然而，当 r过小时， $\overline{\alpha}$ 的一次实现可能具有较大的方差。实践中， r 宜取中等范围，以兼顾降低样本复杂度和算法稳定性。

对于 FastGCN Chen等人（2018），每个节点都会被赋予一个采样概率。我们考虑一种简化的此类重要性采样方法，将训练数据 $\mathcal{D}$ 划分为两个子集$\mathcal{D}_I$ 和 $\mathcal{D}_U$，并令同一子集中的节点具有相同的采样概率。设 γ 为两组采样概率之比，即，

$$\gamma = \frac{\text{Prob}(v \in \mathcal{D}_I \text{is sampled})}{\text{Prob}(v \in \mathcal{D}_U \text{is sampled})} \ (\gamma \geq 1).$$

引理 10 描述了这样一种重要性采样策略：类别相关节点的采样概率比类别无关节点高 γ 倍。根据推论 10.1，我们知道，与均匀采样相比， α 可以提高 γ 倍。形式化地，有

引理 10。 假设 $\mathcal{D}_I = \mathcal{V}_+ \cup \mathcal{V}_-$，且 $\mathcal{D}_U = \mathcal{V}_N$，则有

$$\mathbb{E}\alpha \geq \frac{\gamma r}{R - r + \gamma}. \tag{61}$$

推论 10.1. 对于引理 *10* 中的重要性采样策略， $\mathbb{E}\alpha \geq \gamma \cdot \frac{r}{R}$ 当 $R \gg r$时。–

注 10.1：从推论 10.1 可以看出，平均而言，引理 10 中采用的重要性采样在 $r \ll R$ 时可将样本复杂度降低 $\frac{1}{\gamma^2}$ 倍。

引理 10 描述了这样一种重要性采样策略：一部分具有类别无关特征的节点被赋予较高的采样概率。从推论 10.1 可知， α 相比均匀采样可提升 $1+\frac{\gamma}{(\gamma-1)\lambda}$ 倍，其中 λ 是比例 $|\mathcal{V}_N \cap \mathcal{D}_I|/|\mathcal{V}_N|$。形式化地，我们有

引理11. 假设 $\mathcal{D}_I = \mathcal{V}_+ \cup \mathcal{V}_- \cup \mathcal{V}_\lambda$，其中 $\mathcal{V}_\lambda \subseteq \mathcal{V}_N$ 且 $|\mathcal{V}_\lambda| = \lambda|\mathcal{V}_N|$。我们有

$$\mathbb{E}\alpha \geq \frac{\gamma r}{\big(1 + (\gamma - 1)\lambda\big)R - r + \gamma}. \tag{62}$$

推论11.1。 对于引理*11*中的重要性采样策略， $\mathbb{E}\alpha \geq 1+\frac{\gamma}{(\gamma-1)\lambda} \cdot \frac{r}{R}$ 当~~$R \gg r$~~时。 –

备注 11.1：从推论11.1可以看出，平均而言，引理11中采用的重要性采样在 $r \ll R$ 时可将样本复杂度降低 $\left(\frac{1+(\gamma-1)\lambda}{\gamma}\right)^2$ 倍。

H 有用引理的证明

H.1 引理1的证明

核心思路是求出神经元成为幸运神经元的概率，如式（66）所示。与无噪声情形相比，后者可通过应用 Brutzkus & Globerson（2021）中的对称性质轻松解决；本文则进一步刻画了特征向量中的噪声所导致的边界偏移，如式（65）所示。

引理 *1* 的证明。 对于随机生成的权重 $\boldsymbol{w}_k$，其服从高斯分布；定义随机变量 i_k，使得

$$i_k = \begin{cases} 1, & \text{if } \boldsymbol{w}_k \text{ is the } \textit{lucky neuron} \\ 0, & \text{otherwise} \end{cases}. \tag{63}$$

接下来，我们推导 i_k 的分布。令 θ_1 为 $\boldsymbol{p}_+$ 与初始权重 $\boldsymbol{w}$ 之间的角度。令 $\{\theta_\ell\}_{\ell=2}^L$ 为 $\boldsymbol{p}_-$ 与 $\boldsymbol{p} \in \mathcal{P}_N$ 之间的角度。于是，很容易验证 $\{\theta_\ell\}_{\ell=1}^L$ 服从区间 $[0, 2\pi]$ 上的均匀分布，因为 $\boldsymbol{p} \in \mathcal{P}$ 彼此正交。于是，给定 $\boldsymbol{w}$ 服从高斯分布，$\langle \boldsymbol{w}, \boldsymbol{p} \rangle$ 与 $\boldsymbol{p} \in \mathcal{P}$ 相互独立。这等价于说， $\{\theta_\ell\}_{\ell=1}^L$ 相互独立。考虑到噪声水平 σ （其阶数为 $1/L$）显著小于 1，幸运神经元的概率可界定为

$$\text{Prob}\Big(\theta_1 + \Delta\theta \leq \theta_\ell - \Delta\theta \leq 2\pi, \quad 2 \leq \ell \leq L\Big), \tag{64}$$

其中 $\Delta\theta \approx \sigma$。于是，我们有

$$\begin{aligned} &\text{Prob}\Big(\theta_1 + \Delta\theta \leq \theta_\ell - \Delta\theta \leq 2\pi, \quad 2 \leq \ell \leq L\Big) \\ =&\prod_{\ell=1}^{L} \text{Prob}\Big(\theta_1 + \Delta\theta \leq \theta_\ell - \Delta\theta \leq 2\pi\Big) \\ =&\Big(\frac{2\pi - \theta_1 - \Delta\theta}{2\pi}\Big)^{L-1}. \end{aligned} \tag{65}$$

接下来，我们可以将该概率界定为

$$\begin{aligned}\text{Prob}(\boldsymbol{w}\text{ is the lucky neuron}) =& \int_0^{2\pi} \frac{1}{2\pi}\cdot\left(\frac{2\pi-\theta_1-\Delta\theta}{2\pi}\right)^{L-1} d\theta_1\\ \simeq& \frac{1}{L}\cdot\left(\frac{2\pi-2\Delta\theta}{2\pi}\right)^{L}\\ \simeq& \frac{1}{L}\cdot\left(1-\frac{L\sigma}{\pi}\right).\end{aligned} \tag{66}$$

我们知道 i_k 服从参数为 $\frac{1}{L}(1-\frac{L\sigma}{\pi})$ 的伯努利分布。根据霍夫丁不等式，我们知道

$$\frac{1}{L}\left(1-\frac{L\sigma}{\pi}\right)-\sqrt{\frac{C\log q}{K}} \le \frac{1}{K}\sum_{k=1}^{K} i_k \le \frac{1}{L}\left(1-\frac{L\sigma}{\pi}\right)+\sqrt{\frac{C\log q}{K}} \tag{67}$$

其成立的概率至少为 q^{-C}。令 $K=\mathcal{D}(\varepsilon_K^{-2}L^2\log q)$，我们有

$$\frac{1}{K}\sum_{k=1}^{K} i_k \ge \left(1-\varepsilon_K-\frac{L\sigma}{\pi}\right)\cdot\frac{1}{L}. \tag{68}$$

为了保证 (68) 中的概率为正，噪声水平需要满足

$$\sigma < \frac{\pi}{L}. \tag{69}$$

□

H.2 引理 2 的证明

梯度的界在 (74) 中分为两部分，其中 I_1 和 I_2 分别涉及无噪声特征和噪声。项 I_2 可以用标准切尔诺夫界进行界定，最终结果见 (80)。对于 I_1，由 $\mathcal{D}_+$ 得到的梯度始终沿着具有幸运神经元的 $\boldsymbol{p}_+$ 方向。因此，根据 (83)，神经元权重会沿着 $\boldsymbol{p}_+$ 的方向持续增大。此外，如果对于某个 $y_v=-1$，$\langle\, \boldsymbol{w}_k\,,\,\boldsymbol{x}_v\,\rangle>0$，那么由 v 得到的梯度会迫使 w_k 朝着与 $\boldsymbol{x}_v$ 相反的方向移动。因此，由 $\mathcal{D}_-$ 得到的梯度在更新 $\boldsymbol{w}_k$ 时只会产生负向贡献。对于 $\{\boldsymbol{p}_v\}_{v\in\mathcal{D}_-}$ 中存在的任意模式，也就是说，对于任意 $\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_+$，$\langle\, \boldsymbol{w}_k\,,\,\boldsymbol{p}\,\rangle$ 的值都有一个上界，见 (90)。

引理 *2* 的证明。 对于任意 $v\in\mathcal{D}_+$ 和权重为 $\boldsymbol{w}_k^{(t)}$ 的幸运神经元，我们有

$$\mathcal{M}(v;\boldsymbol{w}_k^{(t)}) = \boldsymbol{p}_+ + \mathcal{M}_z(v;\boldsymbol{w}_k^{(t)}). \tag{70}$$

如果在 $\mathcal{N}^{(t)}(v)$ 中采样到了具有类别相关模式的邻居节点，则梯度计算为

$$\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}^{(t)}(v)})}{\partial \boldsymbol{w}_k}\Big|_{\boldsymbol{w}_k=\boldsymbol{w}_k^{(t)}} = \boldsymbol{p}_+ + \mathcal{M}_z(v;\boldsymbol{w}_k^{(t)}). \tag{71}$$

对于 $v\in\mathcal{D}_-$ 和权重为 $\boldsymbol{w}_k^{(t)}$ 的幸运神经元，我们有

$$\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}) \in \{\boldsymbol{x}_n\}_{n\in\mathcal{N}(v)}\bigcup\{\boldsymbol{0}\}, \tag{72}$$

并且梯度可以表示为

$$\begin{aligned}\frac{\partial g(\boldsymbol{W}^{(t)},\boldsymbol{U}^{(t)};\boldsymbol{X}_{\mathcal{N}^{(t)}(v)})}{\partial \boldsymbol{w}_k}\Big|_{\boldsymbol{w}_k=\boldsymbol{w}_k^{(t)}} =& \mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)})\\ =& \mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) + \mathcal{M}_z^{(t)}(v;\boldsymbol{w}_k^{(t)}).\end{aligned} \tag{73}$$

定义 I_1 和 I_2。 根据上述分析，我们有

$$\begin{aligned}
&\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p}\,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}\,\rangle \\
&= c_\eta \cdot \mathbb{E}_{v\in\mathcal{D}}\langle\, y_v \cdot \mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle \\
&= c_\eta \cdot \Big(\mathbb{E}_{v\in\mathcal{D}}\langle\, y_v \cdot \mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle + \mathbb{E}_{v\in\mathcal{D}}\langle\, y_v \cdot \mathcal{M}_z^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle\Big) \\
&:= c_\eta(I_1 + I_2).
\end{aligned} \tag{74}$$

边界 I_2。 回顾噪声因子 $\boldsymbol{z}_v$ 彼此相同且与 v独立，不难验证， y_v 和 $\mathcal{M}_z(v;\boldsymbol{w}_k^{(t)})$相互独立。因此，式 (74) 中的 I_2 可界定为

$$\begin{aligned}
|I_2| &= |\mathbb{E}_{v\in\mathcal{D}} y_v \cdot \mathbb{E}_{v\in\mathcal{D}}\langle\, \mathcal{M}_z(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle| \\
&\le |\mathbb{E}_{v\in\mathcal{D}} y_v| \cdot |\mathbb{E}_{v\in\mathcal{D}}\langle\, \mathcal{M}_z(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle| \\
&\le \sigma \cdot |\mathbb{E}_{v\in\mathcal{D}} y_v|.
\end{aligned} \tag{75}$$

回顾可采样的邻居节点数量固定为 r。因此，对于任意固定节点 v， $\big\{u\big|u\in\mathcal{D}\big\}$ 中至多有 $(1+r^2)$ 个元素（包括 v 自身）与 y_v相关。此外，根据假设（**A2**），有 Prob$(y=1)$ = $\frac{1}{2}$ 以及 Prob$(y=-1)$ = $\frac{1}{2}$。根据引理 8， $\sum_{v\in\mathcal{D}}(y_v - \mathbb{E}y_v)$的矩生成函数满足

$$e^{\sum_{v\in\mathcal{D}} s(y_v - \mathbb{E}y_v)} \le e^{C(1+r^2)|\mathcal{D}|s^2}, \tag{76}$$

其中， $\mathbb{E}y_v = 0$和 C 是某个正数常量。

根据 Chernoff 不等式，有

$$\text{Prob}\left\{\left|\sum_{v\in\mathcal{D}}(y_v - \mathbb{E}\, y_v)\right| > th\right\} \le \frac{e^{C(1+r^2)|\mathcal{D}|s^2}}{e^{|\mathcal{D}|\cdot th\cdot s}} \tag{77}$$

对于任意 $s>0$。令 $s = th/\big(C(1+r^2)\big)$ 和 $th = \sqrt{(1+r^2)|\mathcal{D}|\log q}$，则有

$$\left|\sum_{v\in\mathcal{D}}(y_v - \mathbb{E}\, y_v)\right| \le C\sqrt{(1+r^2)|\mathcal{D}|\log q} \tag{78}$$

以至少 $1-q^{-c}$ 的概率。根据第 3.2 节中的（**A2**），我们知道

$$|\mathbb{E}y_v| \lesssim \sqrt{|\mathcal{D}|}. \tag{79}$$

因此， I_2 有界如下:

$$\begin{aligned}
|I_2| &\le \Big|\sum_{v\in\mathcal{D}} \frac{1}{|\mathcal{D}|} y_v\Big| \\
&\le \Big|\sum_{v\in\mathcal{D}} \frac{1}{|\mathcal{D}|}(y_v - \mathbb{E}\, y_v)\Big| + |\mathbb{E}\, y_v| \\
&\lesssim \sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.
\end{aligned} \tag{80}$$

当 I_1 满足 $\boldsymbol{p} = \boldsymbol{p}_+$时的界。 由于带有负标签的节点 v 的邻居不包含 $\boldsymbol{p}_+$，因此（73）中的 $\mathcal{M}_p$ （v； $\boldsymbol{w}_k^{(t)}$）不可能是 $\boldsymbol{p}_+$。此外， $\boldsymbol{p}_+$ 与 $\{\boldsymbol{p}_n\}_{n\in\mathcal{N}(v)}\bigcup\{\boldsymbol{0}\}$ 正交。于是，我们有

$$\Big\langle \boldsymbol{p}_+, \frac{\partial g(\boldsymbol{W},\boldsymbol{U};\boldsymbol{X}_{\mathcal{N}(v)})}{\partial \boldsymbol{w}_k}\Big|_{\boldsymbol{w}_k=\boldsymbol{w}_k^{(t)}}\Big\rangle = \langle\, \boldsymbol{p}_+\,,\, \mathcal{M}_z(v;\boldsymbol{w}_k^{(t)})\,\rangle. \tag{81}$$

我们用 $\mathcal{D}_s^{(t)}$ 表示类别相关模式包含在采样邻居节点中的节点集合。回顾第 3 节中定义的 α, 其表示具有重要边的节点的采样率，我们有

$$|\mathcal{D}_s^{(t)}| = \alpha|\mathcal{D}_+|. \tag{82}$$

结合 (71)、(73) 和 (81)，我们有

$$\begin{aligned} I_1 &= \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}_+\ \rangle \\ &= \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}_+\ \rangle \\ &\geq \langle\ \mathbb{E}_{v\in\mathcal{D}_s}\mathcal{M}_p(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}_+\ \rangle \\ &\geq \alpha. \end{aligned} \tag{83}$$

当 I_1 且 $\boldsymbol{p}\neq\boldsymbol{p}_+$时。 接下来，对于任意 $\boldsymbol{p}\in\mathcal{P}/\mathcal{P}+$，我们将证明以下等式：

$$I_1 \leq \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \tag{84}$$

当 $\langle\ \boldsymbol{w}_k^{(t+1)}\ ,\ \boldsymbol{p}\ \rangle \leq -\sigma$时，我们有

$$\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) \neq \boldsymbol{p}, \quad \text{and} \quad \langle\ \mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle = 0 \quad \text{for any} \quad v\in\mathcal{D}. \tag{85}$$

因此，我们有

$$I_1 = \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle = 0. \tag{86}$$

当 $\langle\ \boldsymbol{w}_k^{(t+1)}\ ,\ \boldsymbol{p}\ \rangle > -\sigma$时，我们有

$$\begin{aligned} I_1 &= \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ &= \langle\ \mathbb{E}_{v\in\mathcal{D}_+/\mathcal{D}_s}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle. \end{aligned} \tag{87}$$

定义一个映射 $\mathcal{H}$: $\mathbb{R}^d \longrightarrow \mathbb{R}^d$，使得

$$\mathcal{H}(\boldsymbol{p}_v) = \begin{cases} \boldsymbol{p}_- & \text{if} \quad \boldsymbol{p}_v = \boldsymbol{p}_+ \\ \boldsymbol{p}_v & \text{otherwise} \end{cases}, \tag{88}$$

于是，我们有

$$\begin{aligned} I_1 =& \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ =& \langle\ \mathbb{E}_{v\in\mathcal{D}_+/\mathcal{D}_s}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ =& \langle\ \mathbb{E}_{v\in\mathcal{D}_+/\mathcal{D}_s}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ \leq& \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ =& \langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \end{aligned} \tag{89}$$

其中，第三个等式成立是因为对于 $v\in\mathcal{D}_+/\mathcal{D}_s$，$\mathcal{N}^{(t)}(v)$ 不包含 $\boldsymbol{p}_+$，且$\mathcal{H}(\boldsymbol{X}_{\mathcal{N}^{(t)}(v)}) = \boldsymbol{X}_{\mathcal{N}^{(t)}(v)}$。此外，最后一个等式成立是因为对于节点 $v\in\mathcal{D}_-$，$\boldsymbol{x}_v$ 不包含 $\boldsymbol{p}_+$。因此，我们有

$$\begin{aligned} &\langle\ \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ =& \mathbb{E}_{v\in\mathcal{D}}\langle\ y_v\mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle \\ \leq& \left|\mathbb{E}_{v\in\mathcal{D}}y_v\right| \cdot \left|\mathbb{E}_{v\in\mathcal{D}}\langle\ \mathcal{M}_p^{(t)}(\mathcal{H}(v);\boldsymbol{w}_k^{(t)})\ ,\ \boldsymbol{p}\ \rangle\right| \\ \leq& \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \end{aligned} \tag{90}$$

以至少 $1-q^{-C}$ 的概率，其中 $1-q^{-C}$ 是某个正数常量。

陈述 1 的证明。 由 (80) 和 (83) 可知，$\boldsymbol{w}_k$ 沿 $\boldsymbol{p}_+$ 方向的更新有如下界：

$$\begin{aligned} \langle\ \boldsymbol{w}_k^{(t+1)}\ ,\ \boldsymbol{p}_+\ \rangle - \langle\ \boldsymbol{w}_k^{(t)}\ ,\ \boldsymbol{p}_+\ \rangle &\geq c_\eta(I_1 - |I_2|) \\ &\geq c_\eta\cdot(\alpha - \sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}). \end{aligned} \tag{91}$$

于是，我们有

$$\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p}_+ \rangle - \langle \boldsymbol{w}_k^{(0)} , \boldsymbol{p}_+ \rangle \geq c_\eta \cdot (\alpha - \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}) \cdot t. \tag{92}$$

Pro陈述 2。 根据（80）和（90）， $\boldsymbol{w}_k$ 沿 $\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_+$ 方向的更新上界为 r
bou^nded as

$$\begin{aligned} \langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle &\leq c_\eta (I_1 + |I_2|) \\ &\leq c_\eta \cdot (1+\sigma) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \end{aligned} \tag{93}$$

因此，我们有

$$\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(0)} , \boldsymbol{p} \rangle \leq c_\eta \cdot (1+\sigma) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \cdot t. \tag{94}$$

$\boldsymbol{w}_k$ 沿 $\boldsymbol{p} \in \mathcal{P}/\boldsymbol{p}_+$ 方向的更新下界为

$$\begin{aligned} \langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle &\geq c_\eta (I_1 - |I_2|) \\ &\geq -c_\eta \cdot \left(1 + \sigma \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right). \end{aligned} \tag{95}$$

为推导下界，我们通过数学归纳法证明以下等式:

$$\langle \boldsymbol{w}^{(t)} , \boldsymbol{p} \rangle \geq -c_\eta \left(1 + \sigma + \sigma t \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right). \tag{96}$$

显然，当 $t = 0$时，（96）成立。假设（96）对 t成立。
当 $\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle \leq -\sigma$时，我们有

$$\mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) \neq \boldsymbol{p}, \quad \text{and} \quad \langle \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p} \rangle = 0 \quad \text{for any} \quad v \in \mathcal{D}. \tag{97}$$

因此，我们有

$$I_1 = \langle \mathbb{E}_{v \in \mathcal{D}_+} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v \in \mathcal{D}_-} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p} \rangle = 0, \tag{98}$$

and

$$\begin{aligned} \langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p} \rangle &= \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle + c_\eta \cdot (I_1 + I_2) \\ &= \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle + c_\eta \cdot I_2 \\ &\geq -c_\eta \cdot \left(1 + \frac{\sigma}{c_\eta} + \sigma t \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right) - c_\eta \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \\ &= -c_\eta \cdot \left(1 + \frac{\sigma}{c_\eta} + \sigma(t+1) \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right). \end{aligned} \tag{99}$$

当 $\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle > -\sigma$时，我们有

$$\begin{aligned} &\langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p} \rangle \\ &= \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle + c_\eta (I_1 + I_2) \\ &= \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle + c_\eta \cdot \langle \mathbb{E}_{v \in \mathcal{D}_+} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v \in \mathcal{D}_-} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p} \rangle + c_\eta I_2 \\ &= \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle - c_\eta \cdot \langle \mathbb{E}_{v \in \mathcal{D}_+} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p} \rangle + c_\eta I_2 \\ &\geq -\sigma - c_\eta - c_\eta \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \end{aligned} \tag{100}$$

因此，根据 (99) 和 (100)，我们知道 (96) 对 $t+1$ 成立。 □

H.3 引理4的证明

(101) 中的梯度界分为两部分，其中 I_3 和 I_4 分别对应无噪声特征和噪声。I_4 的界与证明引理 2 时 I_2 的界类似。然而，对于一个不幸神经元，由 $\mathcal{D}_+$ 得出的梯度并不总是沿着 $\boldsymbol{p}_+$ 的方向。我们需要额外的技术来刻画 $\mathcal{D}_+$ 和 $\mathcal{D}_-$ 之间的偏移。一个关键问题是确保 $\mathcal{M}^{(t)}(v)$ 与 y_v 的独立性，这通过构造匹配数据得以解决，定义见 (108)。根据 (113)，我们证明了 $\boldsymbol{w}_k^{(t)}$ 的尺度按 $\sqrt{(1+r^2)/|\mathcal{D}|}$ 的阶增长。

引理4的证明。**定义 I_1 和 I_2**。与（74）类似，我们定义项 I_3 和 I_4 a s

$$\begin{aligned}
&\langle\, \boldsymbol{w}_k^{(t+1)}\,,\, \boldsymbol{p}\,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}\,\rangle \\
&= c_\eta \cdot \mathbb{E}_{v\in\mathcal{D}}\langle\, y_v \cdot \mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle + \mathbb{E}_{v\in\mathcal{D}}\langle\, y_v \cdot \mathcal{M}_z^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle \\
&:= c_\eta \cdot (I_3 + I_4).
\end{aligned} \tag{101}$$

的界 I_4。按照（80）的类似推导，我们可以得到

$$|I_4| \lesssim \sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \tag{102}$$

的界 I_3 当 $\boldsymbol{p}=\boldsymbol{p}_+$时。由（101），我们有

$$\begin{aligned}
I_3 &= \langle\, \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}_+\,\rangle \\
&= \langle\, \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}_+\,\rangle \\
&\geq 0.
\end{aligned} \tag{103}$$

的界 I_3 当 $\boldsymbol{p}\in\mathcal{P}_N$时。为了界定任意固定的 $\boldsymbol{p}\in\mathcal{P}_N$ 对应的 I_4，我们维持子集 $\mathcal{S}_k(t) \subseteq \mathcal{D}_-$，使得 $\mathcal{S}_k(t) = \{v\in\mathcal{D}_- \mid \mathcal{M}_p^{(t)}(v;\ \boldsymbol{w}_k^{(t)}) = \boldsymbol{p}_-\}$。
首先，当 $\mathcal{S}_k(t) = \emptyset$时，很容易验证

$$I_3 = 0. \tag{104}$$

因此，对于所有 $t'\geq t$，$\mathcal{S}_k(t+1) = \emptyset$ 且 $\mathcal{S}_k(t') = \emptyset$。
其次，当 $\mathcal{S}_k(t) \neq \emptyset$ 时，我们有

$$I_3 = -\frac{|\mathcal{S}_k(t)|}{|\mathcal{D}_-|}\|\boldsymbol{p}_-\|^2. \tag{105}$$

注意，如果 $\langle\, \boldsymbol{w}_k^{(t)}\,,\, \boldsymbol{p}_-\,\rangle < -\sigma$，那么 $\mathcal{S}_k(t)$ 必须是空集合。因此，经过至多 $t_0 = \frac{\|\boldsymbol{w}_k^{(0)}\|_2+\sigma}{c_\eta}$ 次迭代后，$\mathcal{S}_k(t_0) = \emptyset$，且 $\mathcal{S}_k(t) = \emptyset$，其中 $t\geq t_0$ 来自 (104)。

接下来，对于某个足够大的 t_0 和任意 $v\in\mathcal{V}$，我们定义一个映射 $\mathcal{F}\colon\ \mathcal{V}\longrightarrow\mathbb{R}^{rd}$，使得

$$\mathcal{F}(v) = \begin{bmatrix}\boldsymbol{o}_{v_1}^\top & \boldsymbol{o}_{v_2}^\top & \cdots & \boldsymbol{o}_{v_r}^\top\end{bmatrix}^\top. \tag{106}$$

由 (101)，我们有

$$I_4 = \mathbb{E}_{v\in\mathcal{V}}\langle\, y_v\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\, \boldsymbol{p}\,\rangle, \tag{107}$$

其中 $\{v_1,\cdots,v_r\} = \mathcal{N}^{(t)}(v)$，并且

$$\begin{cases}\boldsymbol{o}_{v_i} = \boldsymbol{x}_{v_i}, & \text{if } \boldsymbol{x}_{v_i}\neq\boldsymbol{p}_+ \text{ or } \boldsymbol{p}_- \\ \boldsymbol{o}_{v_i} = \boldsymbol{0}, & \text{if } \boldsymbol{x}_{v_i}=\boldsymbol{p}_+ \text{ or } \boldsymbol{p}_-\end{cases}. \tag{108}$$

当 $y=1$时，我们有

$$\langle\, \mathcal{M}_p^{(t)}(\boldsymbol{x})\,,\, \boldsymbol{p}\,\rangle = \begin{cases}\langle\, \mathcal{M}_p^{(t)}(\mathcal{F}(v))\,,\, \boldsymbol{p}\,\rangle, & \text{if} \quad \mathcal{M}_p^{(t)}(\mathcal{F}(v))\neq\boldsymbol{p}_+ \\ 0\leq\langle\, \mathcal{M}_p^{(t)}(\mathcal{F}(v))\,,\, \boldsymbol{p}\,\rangle, & \text{if} \quad \mathcal{M}_p^{(t)}(\mathcal{F}(v))=\boldsymbol{p}_+\end{cases}. \tag{109}$$

当 $y=-1$ 且 $t \geq t_0$时，回顾一下 $\mathcal{S}_k(t) = \mathcal{D}_-$，那么我们有

$$\langle \mathcal{M}_p^{(t)}(\boldsymbol{x}) , \boldsymbol{p} \rangle = \langle \mathcal{M}_p^{(t)}(\mathcal{F}(\boldsymbol{x})) , \boldsymbol{p} \rangle. \tag{110}$$

结合 (109) 和 (110)，我们得到

$$I_3 \leq \mathbb{E}_{v\in\mathcal{V}}\big[y \cdot \langle \mathcal{M}_p^{(t)}(\mathcal{F}(v)) , \boldsymbol{p} \rangle\big]. \tag{111}$$

由假设 **(A2)** 和 (17) 可知，$\boldsymbol{p}_+$ 和 $\boldsymbol{p}_-$ 的分布相同，且任意类别无关模式 $\boldsymbol{p} \in \mathcal{P}_N$ 的分布与 y 独立。因此，容易验证 y 和 $\mathcal{F}$（v）彼此独立，于是我们得到

$$I_3 \leq \mathbb{E}_{v\in\mathcal{D}} y_v \cdot \mathbb{E}_v \langle \mathcal{M}(\mathcal{F}(v)) , \boldsymbol{p} \rangle. \tag{112}$$

由 (112) 和 (78)，我们得到

$$I_3 \leq \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \tag{113}$$

对任意 $\boldsymbol{p} \in \mathcal{P}_N$。

陈述 1 的证明。由 (102) 和 (103)，我们得到

$$\begin{aligned} \langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle &= c_\eta \cdot (I_3 + I_4) \\ &\geq -c_\eta |I_4| \\ &\geq -c_\eta \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \end{aligned} \tag{114}$$

因此，我们得到

$$\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p} \rangle - \langle \boldsymbol{w}_k^{(0)} , \boldsymbol{p} \rangle \geq -c_\eta \sigma \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \cdot t. \tag{115}$$

命题 2 的证明。当 $\boldsymbol{p} = \boldsymbol{p}_-$ 时，根据 (101) 中 I_3 的定义，我们知道

$$\begin{aligned} I_3 &= \langle \mathbb{E}_{v\in\mathcal{D}_+} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-} \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p}_- \rangle \\ &= -\langle \mathbb{E}_{v\in\mathcal{D}_-} \mathcal{M}_p^{(t)}(v; \mathcal{D}_-) , \boldsymbol{p}_- \rangle \\ &\leq 0, \end{aligned} \tag{116}$$

$\boldsymbol{w}_k$ 沿 $\boldsymbol{p}_-$ 方向的更新有如下上界：

$$\begin{aligned} \langle \boldsymbol{w}_k^{(t+1)} , \boldsymbol{p}_- \rangle - \langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p}_- \rangle &\leq c_\eta (I_3 + I_4) \\ &\leq c_\eta I_4 \\ &\leq c_\eta \cdot \sigma \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}. \end{aligned} \tag{117}$$

因此，我们有

$$\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p}_- \rangle - \langle \boldsymbol{w}_k^{(0)} , \boldsymbol{p} \rangle \leq c_\eta \cdot \sigma \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}} \cdot t. \tag{118}$$

为推导下界，我们通过数学归纳法证明以下等式：

$$\langle \boldsymbol{w}^{(t)} , \boldsymbol{p}_- \rangle \geq -c_\eta \left(1 + \sigma + \sigma t \cdot \sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right). \tag{119}$$

显然，当 $t=0$ 时，(119) 成立。假设 (119) 对 t 成立。

当 $\langle \boldsymbol{w}_k^{(t)} , \boldsymbol{p}_- \rangle \leq -\sigma$时，我们有

$$\mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) \neq \boldsymbol{p}_-, \quad \text{and} \quad \langle \mathcal{M}_p^{(t)}(v; \boldsymbol{w}_k^{(t)}) , \boldsymbol{p}_- \rangle = 0 \quad \text{for any} \quad v \in \mathcal{D}. \tag{120}$$

因此，我们有

$$I_3 = \langle\, \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\,\boldsymbol{p}_-\,\rangle = 0, \tag{121}$$

and

$$\begin{aligned}\langle\, \boldsymbol{w}_k^{(t+1)}\,,\,\boldsymbol{p}_-\,\rangle &= \langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle + c_\eta\cdot(I_3+I_4)\\ &= \langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle + c_\eta\cdot I_4\\ &\geq -c_\eta\cdot\left(1+\frac{\sigma}{c_\eta}+\sigma t\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right) - c_\eta\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\\ &= -c_\eta\cdot\left(1+\frac{\sigma}{c_\eta}+\sigma(t+1)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right).\end{aligned} \tag{122}$$

当 $\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle > -\sigma$时，我们有

$$\begin{aligned}&\langle\, \boldsymbol{w}_k^{(t+1)}\,,\,\boldsymbol{p}_-\,\rangle\\ =&\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle + c_\eta(I_3+I_4)\\ =&\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle + c_\eta\cdot\langle\, \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)}) - \mathbb{E}_{v\in\mathcal{D}_-}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\,\boldsymbol{p}_-\,\rangle + c_\eta I_4\\ =&\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_-\,\rangle - c_\eta\cdot\langle\, \mathbb{E}_{v\in\mathcal{D}_+}\mathcal{M}_p^{(t)}(v;\boldsymbol{w}_k^{(t)})\,,\,\boldsymbol{p}_-\,\rangle + c_\eta I_4\\ \geq& -\sigma - c_\eta - c_\eta\sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.\end{aligned} \tag{123}$$

因此，由（122）和（123）可知，（119）对 $t+1$成立。

命题 3 的证明。由（102）和（113），我们有

$$\begin{aligned}|\langle\, \boldsymbol{w}_k^{(t+1)}\,,\,\boldsymbol{p}\,\rangle - \langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}\,\rangle| &= c_\eta\cdot|I_3+I_4|\\ &\leq c_\eta\cdot(1+\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}.\end{aligned} \tag{124}$$

因此，我们有

$$|\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}\,\rangle - \langle\, \boldsymbol{w}_k^{(0)}\,,\,\boldsymbol{p}\,\rangle| \leq c_\eta\cdot(1+\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\cdot t. \tag{125}$$

□

H.4 引理 6 的证明

主要思路是证明，幸运神经元的权重会沿着类别相关模式的方向持续增加。因此，幸运神经元会在聚合函数中持续选择类别相关模式。

引理 *6* 的证明。对于任意 $k\in\mathcal{U}(0)$，我们有

$$\langle\, \boldsymbol{w}_k^{(0)}\,,\,(1-\sigma)\cdot\boldsymbol{p}_+\,\rangle \geq \langle\, \boldsymbol{w}_k^{(0)}\,,\,(1+\sigma)\cdot\boldsymbol{p}\,\rangle \tag{126}$$

对于任意 $\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_+$，根据式 (24) 中幸运神经元的定义。

接下来，由引理 2 可得

$$\begin{aligned}&\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}_+\,\rangle \geq c_\eta\left(\alpha-\sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right)\cdot t,\\ \text{and}\quad &\langle\, \boldsymbol{w}_k^{(t)}\,,\,\boldsymbol{p}\,\rangle \leq c_\eta(1+\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\cdot t.\end{aligned} \tag{127}$$

结合式 (126) 和 (127)，可得

$$\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ \rangle - \frac{1+\sigma}{1-\sigma}\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p} \rangle \gtrsim c_\eta \cdot \left(\alpha - (1+4\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right)\cdot t. \tag{128}$$

当 $|\mathcal{D}| \gtrsim \alpha^{-2}\cdot(1+r^2)\log q$ 时，我们有 $\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p}_+ \rangle \geq \frac{1+\sigma}{1-\sigma}\langle \boldsymbol{w}_k^{(t)}, \boldsymbol{p} \rangle$。

因此，我们有 $k \in \mathcal{W}(t)$，并且

$$\mathcal{W}(0) \subseteq \mathcal{W}(t). \tag{129}$$

按照上述类似步骤，可以推导出 $\mathcal{U}(t)$的证明。 □

H.5 引理 7 的证明

证明基于引理 2 和 4 中关于幸运神经元和不幸神经元的陈述。幸运神经元在 $\boldsymbol{p}_+$ 方向上的幅度为 α 量级，而不幸神经元的幅度至多为 $1/\sqrt{|\mathcal{D}|}$。给定足够大的 $|\mathcal{D}|$，幸运神经元的幅度总是大于不幸神经元的幅度。

引理 7 的证明。 根据引理 6，我们知道：(1) 如果 $\boldsymbol{w}_{k_1}^{(0)}$ 是幸运神经元，那么对于任意 $t' \geq 0$，$\boldsymbol{w}_{k_1}^{(0)}$ 仍是幸运神经元；(2) 如果 $\boldsymbol{w}^{(t')}{}_{k_1}$ 是不幸神经元，那么对于任意 $t'' \leq t'$，$\boldsymbol{w}^{(t'')}{}_{k_1}$ 仍是不幸神经元。对于任意 $k_1 \in \mathcal{W}(0)$，根据引理 2，我们有

$$\begin{aligned}\langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{w}_{k_1}^{(t')} \rangle^{\frac{1}{2}} \geq& \langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p}_+ \rangle \\ \geq& c_\eta \cdot \left(\alpha - \sigma\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\right)t'.\end{aligned} \tag{130}$$

对于任意 $k_2 \in \mathcal{W}^c(t')$，根据引理 2，我们有

$$\begin{aligned}\langle \boldsymbol{w}_{k_2}^{(t')}, \boldsymbol{w}_{k_2}^{(t')} \rangle^{\frac{1}{2}} =& \sum_{\boldsymbol{p}\in\mathcal{P}}\langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p} \rangle \\ =& \sum_{\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_+}\langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p} \rangle + \langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p}_+ \rangle \\ \leq& \sum_{\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_+}\langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p} \rangle + \max_{\boldsymbol{p}\in\mathcal{P}/\boldsymbol{p}_+}\langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{p} \rangle \\ \leq& L\cdot c_\eta\cdot(1+\sigma)\cdot\sqrt{\frac{(1+r^2)\log q}{|\mathcal{D}|}}\cdot t'.\end{aligned} \tag{131}$$

因此，$\langle \boldsymbol{w}^{(t')}{}_{k_2}, \boldsymbol{w}^{(t')}{}_{k_2} \rangle^{\frac{1}{2}} < \langle \boldsymbol{w}_{k_1}^{(t')}, \boldsymbol{w}^{(t')}{}_{k_1} \rangle^{\frac{1}{2}}$ 若 $|\mathcal{D}|$ 大于 $\alpha^{-2}(1+r^2)L^2\log q$。 □

H.6 引理 8 的证明

引理 8 的证明。 根据 Janson（2004）中的定义，存在一个由$\{(\mathcal{X}_j, w_j)\}_j$构成的族，其中 $\mathcal{X}_j \subseteq \mathcal{X}$和 $w_j \in [0,1]$，使得 $\sum_j w_j \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j} = \sum_{n=1}^N x_n$，并且由 Janson（2004）中的公式 (2.1) 和 (2.2) 可得$\sum_j w_j \leq d_{\mathcal{X}}$。然后，令 p_j 为满足 $\sum_j p_j = 1$的任意正数。根据 Jensen 不等式，对于任意 $s \in \mathbb{R}$，我们有

$$e^{s\sum_{n=1}^N x_n} = e^{\sum_j p_j \frac{s w_j}{p_j} X_j} \leq \sum_j p_j e^{\frac{s w_j}{p_j} X_j}, \tag{132}$$

其中 $X_j = \sum_{x_{n_j}\in\mathcal{X}_j} x_{n_j}$。

那么，我们有

$$\begin{aligned}\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^{N} x_n} \leq & \mathbb{E}_{\mathcal{X}} \sum_j p_j e^{\frac{sw_j}{p_j} X_j} = \sum_j p_j \prod_{\mathcal{X}_j} \mathbb{E}_{\mathcal{X}} e^{\frac{sw_j}{p_j} x_{n_j}} \\ \leq & \sum_j p_j \prod_{\mathcal{X}_j} e^{\frac{Cw_j^2}{p_j^2} s^2} \\ \leq & \sum_j p_j e^{\frac{C|\mathcal{X}_j| w_j^2}{p_j^2} s^2}.\end{aligned} \tag{133}$$

令 $p_j = \frac{w_j |\mathcal{X}_j|^{1/2}}{\sum_j w_j |\mathcal{X}_j|^{1/2}}$，则有

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^{N} x_n} \leq \sum_j p_j e^{C\left(\sum_j w_j |\mathcal{X}_j|^{1/2}\right)^2 s^2} = e^{C\left(\sum_j w_j |\mathcal{X}_j|^{1/2}\right)^2 s^2}. \tag{134}$$

根据柯西-施瓦茨不等式，有

$$\left(\sum_j w_j |\mathcal{X}_j|^{1/2}\right)^2 \leq \sum_j w_j \sum_j w_j |\mathcal{X}_j| \leq d_{\mathcal{X}} N. \tag{135}$$

因此，有

$$\mathbb{E}_{\mathcal{X}} e^{s\sum_{n=1}^{N} x_n} \leq e^{Cd_{\mathcal{X}} N s^2}. \tag{136}$$

□

I 扩展至多类别分类

考虑一个包含四个类别的分类问题，我们使用标签 $\boldsymbol{y} \in \{+1, -1\}^2$ 表示相应的类别。与第 2 节中的设定类似，这里有四个正交的类别相关模式，即 $\boldsymbol{p}_1$，$\boldsymbol{p}_2$，$\boldsymbol{p}_3$，$\boldsymbol{p}_4$。在第一层（隐藏层）中，我们有 K 个神经元，其权重为 $\boldsymbol{w}_k \in \mathbb{R}^d$。在第二层（线性层）中，第 k 个神经元的权重记为 $\boldsymbol{b}_k \in \mathbb{R}^2$。令 $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ 和 $\boldsymbol{B} \in \mathbb{R}^{2\times K}$ 分别为 $\boldsymbol{w}_k$ 和 $\boldsymbol{b}_k$ 的集合。于是，图神经网络的输出记为 $\boldsymbol{g} \in \mathbb{R}^2$，计算如下

$$\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{B}; \boldsymbol{X}_{\mathcal{N}(v)}) = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{b}_k \cdot \mathrm{AGG}(\boldsymbol{X}_{\mathcal{N}(v)}, \boldsymbol{w}_k). \tag{137}$$

然后，图神经网络生成的标签写作

$$\boldsymbol{y}_{\text{est}} = \mathrm{sign}(\boldsymbol{g}). \tag{138}$$

给定数据集合 $\mathcal{D}$，我们将其分为四组，如下所示

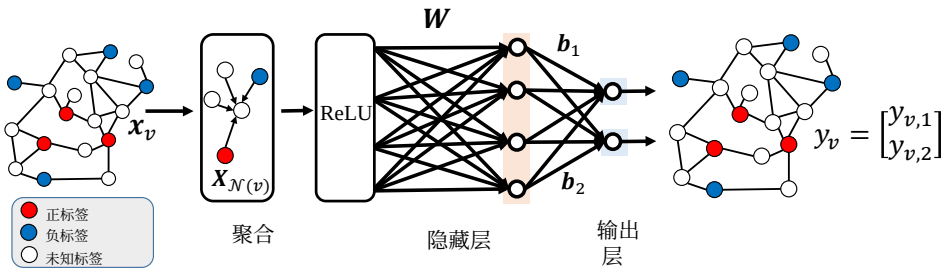


图28：图神经网络进行多类分类学习的示意图

$$\begin{aligned}\mathcal{D}_1 &= \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} = (1, 1)\}, \\ \mathcal{D}_2 &= \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} = (1, -1)\}, \\ \mathcal{D}_3 &= \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} = (-1, 1)\}, \\ \text{and} \quad \mathcal{D}_4 &= \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} = (-1, -1)\}.\end{aligned} \tag{139}$$

式（2）中对应的损失函数可改写为

$$\hat{f}_{\mathcal{D}}(\boldsymbol{W},\boldsymbol{U}) = -\frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}} \boldsymbol{y}_v^\top \boldsymbol{g}\big(\boldsymbol{W},\boldsymbol{b};\boldsymbol{X}_{\mathcal{N}(v)}\big) \tag{140}$$

然后，我们将 $\boldsymbol{b}_k$ 的权重初始化为 $(1,\ 1)$，$(1,\ -1)$，$(-1,\ 1)$，以及 $(-1,\ -1)$，且各组权重相等。接下来，我们根据 $\boldsymbol{b}_k$ 的值将权重 $\boldsymbol{w}_k$ 分为四组，使得

$$\begin{gathered}\mathcal{W}_1 = \{k \mid \boldsymbol{b}_k = (1,1)\},\\ \mathcal{W}_2 = \{k \mid \boldsymbol{b}_k = (1,-1)\},\\ \mathcal{W}_3 = \{k \mid \boldsymbol{b}_k = (-1,1)\},\\ \text{and}\quad \mathcal{W}_4 = \{k \mid \boldsymbol{b}_k = (-1,-1)\}.\end{gathered} \tag{141}$$

因此，对于任意 $k\in\mathcal{W}_1$，我们有

$$\frac{\partial f}{\partial \boldsymbol{w}} = -\frac{1}{|\mathcal{D}|}\sum_{v\in\mathcal{D}}\left(y_{v,1}\cdot\frac{\partial g_1}{\partial \boldsymbol{w}} - y_{v,2}\cdot\frac{\partial g_2}{\partial \boldsymbol{w}}\right). \tag{142}$$

当 $x\in\mathcal{D}_1$ 时，我们有

$$y_{v,1}\cdot\frac{\partial g_1}{\partial \boldsymbol{w}} - y_{v,2}\cdot\frac{\partial g_2}{\partial \boldsymbol{w}} = -2\phi'(\boldsymbol{w}^\top\boldsymbol{x})\cdot\boldsymbol{x}. \tag{143}$$

当 $x\in\mathcal{D}_2$ 时，我们有

$$y_{v,1}\cdot\frac{\partial g_1}{\partial \boldsymbol{w}} - y_{v,2}\cdot\frac{\partial g_2}{\partial \boldsymbol{w}} = 0. \tag{144}$$

当 $x\in\mathcal{D}_3$ 时，我们有

$$y_{v,1}\cdot\frac{\partial g_1}{\partial \boldsymbol{w}} - y_{v,2}\cdot\frac{\partial g_2}{\partial \boldsymbol{w}} = 0. \tag{145}$$

当 $x\in\mathcal{D}_4$ 时，我们有

$$y_{v,1}\cdot\frac{\partial g_1}{\partial \boldsymbol{w}} - y_{v,2}\cdot\frac{\partial g_2}{\partial \boldsymbol{w}} = 2\phi'(\boldsymbol{w}^\top\boldsymbol{x})\cdot\boldsymbol{x}. \tag{146}$$

那么，对于任意 $k\in\mathcal{W}_1$，我们有

$$\boldsymbol{w}_k^{(t+1)} = \boldsymbol{w}_k^{(t)} + c_\eta\cdot\mathbb{E}_{v\in\mathcal{D}_1}\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}) - c_\eta\cdot\mathbb{E}_{v\in\mathcal{D}_4}\mathcal{M}^{(t)}(v;\boldsymbol{w}_k^{(t)}), \tag{147}$$

比较（147）和（29）可知，权重 $\boldsymbol{w}_k^{(t)}$ 将主要沿着 $\boldsymbol{p}_1$ 的方向更新，但在其他方向上的幅值有界。按照（142）至（147）的步骤，对于满足 $1\le i\le 4$ 的任意 $k\in\mathcal{W}_i$，也可以推导出类似的结果。