图霍普菲尔德网络：基于能量的节点分类与关联记忆

Abinav Rao* **Alex Wa*** **Rishi Athavale***

*贡献相同

arXiv:2603.03464v1 [cs.LG] 3 Mar 2026

摘要

我们提出图霍普菲尔德网络（GHN），其能量函数将关联记忆检索与图拉普拉斯平滑结合，用于节点分类。对这一联合能量进行梯度下降，可得到一种迭代更新过程，交替执行霍普菲尔德检索与拉普拉斯传播。记忆检索带来的收益因应用场景而异：在稀疏引文网络上最高可提升 2.0 个百分点，在特征掩蔽下则可额外提高最高 5 个百分点的鲁棒性；迭代式能量下降架构本身也是一种强大的归纳偏置，所有变体（包括禁用记忆的 NoMem 消融）在亚马逊共同购买图上均优于标准基线。调整 $\lambda \leq 0$即可在不改变架构的情况下，为异配基准实现图锐化。

1 引言

现代霍普菲尔德网络（Krotov 与 Hopfield，2016年；Demircigil 等，2017年；Ramsauer 等，2021年）重新引发了人们对关联记忆这一计算基本原语的兴趣。它们与 Transformer 注意力的联系（Ramsauer 等，2021年）推动了其在计算机视觉、语言和科学领域的应用。然而，尽管相关理论认识不断加深（包括分层记忆组织（Krotov，2021年）、新型能量函数（Hoover 等，2025年）以及容量分析（Lucibello 与 M´ezard，2024年）），据我们所知，通过基于能量的形式将其应用于图上的节点分类仍未得到充分探索。Liang 等人（2022年）将现代霍普菲尔德网络应用于图级任务，但并未在联合能量中将记忆检索与图结构结合起来。

图神经网络（GNNs）通过聚合邻域信息来学习节点表示（Kipf & Welling, 2017年；Velickovic et al., 2018年；Hamilton et al., 2017年），但在边存在噪声或不完整时，性能会下降（Z¨ugneret al., 2018年）。联想记忆提供了另一种基于内容的信号：它不依赖局部结构，而是从特征内容中检索相关模式。

We p提出图霍普菲尔德网络 （GHN），将这两种机制结合在一个单一的图能量中：

$$E_{\mathrm{GH}}(\mathbf{X}) = \sum_{v \in \mathcal{V}} \left[-\mathrm{lse}(\beta, \mathbf{M}\mathbf{x}_v) + \tfrac{1}{2}\|\mathbf{x}_v\|^2 \right] + \lambda \,\mathrm{tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X}), \tag{1}$$

其中，$\mathrm{lse}(\beta, \mathbf{z}) := \beta^{-1} \log \sum_\mu \exp(\beta z_\mu)$ 是对数求和指数算子。第一项推动记忆检索趋向学习到的模式 $\mathbf{M}$，第二项则促进图拉普拉斯算子 $\mathbf{L}$上的光滑性。对该能量进行梯度下降，可得到一种交替执行霍普菲尔德检索与拉普拉斯传播的更新规则（第 2 节；推导见附录 D.2）。这种迭代式能量下降架构本身就具有很强的归纳偏置：所有 GHN 变体（包括禁用记忆消融）在亚马逊共同购买图上的表现都优于所有标准基线（第 3.2 节）。记忆可替代缺失的结构信号：在稀疏引文网络上最多提升 2.0 个百分点，在特征掩蔽下最多提升 5 个百分点的鲁棒性（第 3.3 节）；但在拉普拉斯算子已足够的稠密图上，记忆则是冗余的。调整 $\lambda \leq 0$ 可在不改变架构的情况下，为异配基准实现图锐化（第 3.4 节）。

2 图霍普菲尔德网络

2.1 能量函数与更新规则

给定一个图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其节点特征为 $\mathbf{X} \in \mathbb{R}^{N \times d}$，图的（对称归一化）拉普拉斯算子为 $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$，我们按照公式 1 定义图霍普菲尔德能量 。这两项编码了互补的目标：霍普菲尔德项推动每个节点的表示趋向相关的学习到的模式，而拉普拉斯项则促使相邻节点具有相似的表示。

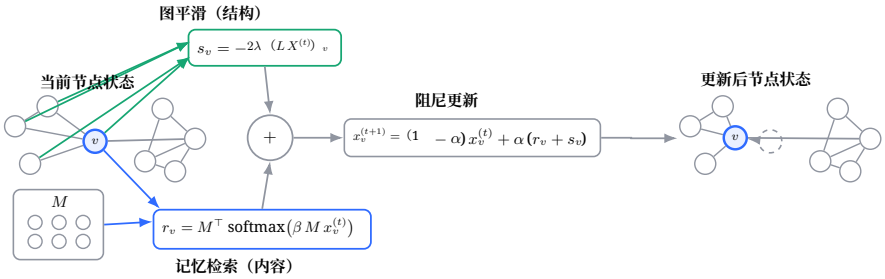

图 1：GHN 更新的一次迭代：节点特征 $\mathbf{x}_v$ 与以下两者进行融合：(i) 从模式库 $\mathbf{M}$中进行记忆检索；(ii) 在邻居上进行图拉普拉斯平滑。阻尼 α 控制步长。

对 $\mathbf{x}_v$ 求梯度可得到更新规则：

$$\mathbf{x}_v^{(t+1)} = (1-\alpha)\,\mathbf{x}_v^{(t)} + \alpha\left[\underbrace{\mathbf{M}^\top \mathrm{softmax}\Big(\beta \mathbf{M}\mathbf{x}_v^{(t)}\Big)}_{\text{memory retrieval}} - 2\lambda\ \underbrace{(\mathbf{L}\mathbf{X}^{(t)})_v}_{\text{graph smoothing}}\right] \tag{2}$$

其中，$\alpha \in (0,\ 1)$ 为阻尼系数。每次迭代自然地交替执行两个操作：从记忆库 $\mathbf{M}$中进行记忆检索，以及通过图拉普拉斯进行基于结构的平滑。

对于最小二乘检索（LSR；Hoover 等人（2025年）），softmax 被归一化的 Epanechnikov 核权重所替代：$w_\mu = \mathrm{ReLU}(1 - \frac{\beta}{2}\|\mathbf{x}_v - \mathbf{m}_\mu\|^2) / \sum_{\mu'} \mathrm{ReLU}(1 - \frac{\beta}{2}\|\mathbf{x}_v - \mathbf{m}_{\mu'}\|^2)$。在我们的实现中，$\beta$ 被视为可学习参数，因此检索锐度会在训练过程中进行调整。

基础 LSE 公式的收敛性在附录 D 中进行了分析；实践中，我们在所有实验中均观察到可在 $T{=}4$ 迭代次数内稳定收敛。

2.2 门控记忆检索

Me记忆检索在训练早期，或当查询远离任何存储时，可能产生较差的输出 d pattern；为防止其破坏表示，我们引入一个可学习的门控：

$$\mathbf{g}_v = \sigma(\mathbf{W}_g[\mathbf{x}_v \| \mathbf{r}_v] + \mathbf{b}_g)\,, \qquad \tilde{\mathbf{r}}_v = \mathbf{g}_v \odot \mathbf{r}_v + (1 - \mathbf{g}_v) \odot \mathbf{x}_v, \tag{3}$$

其中 $\mathbf{r}_v$ 是原始检索输出，σ 是 Sigmoid，$\|$ 表示拼接，$\odot$ 是逐元素乘法。我们将 $b_g = 2$ ($\mathbf{g}_v \approx 0.88$)初始化，以在训练早期鼓励使用记忆；门控输出 $\tilde{\mathbf{r}}_v$ 替换式 2 中的检索项。

2.3 分层记忆

当 K 相对于特征维度较大时，对 K 模式进行平坦注意力处理可能会不稳定（第 3.2 节）。受 Krotov （ 2021 年）的启发，我们将 K 模式划分为 G 组，并采用两阶段检索：（ i ）**路由**，根据由模式参数计算得到的质心，将查询软分配至各组；然后（ii）**检索**，在选定的各组内执行标准 Hopfield 检索。完整模型由一个编码器、 L 个堆叠的 GHN 层（每层包含 T 次迭代和门控）以及一个线性分类器组成；检索使用绑定键值。详情见附录 A。

3 实验

3.1 设置

数据集。我们在九个基准数据集上进行评估：三个同质性引文网络（Cora、CiteSeer、PubMed（Sen 等，2008））、两个亚马逊共同购买图（Photo 和 Computers（Shchur 等，2018））以及四个异配图（Texas、Wisconsin、康奈尔（Pei 等，2020）和 Actor（Tang 等，2009））。Planetoid 使用标准公开数据划分；亚马逊数据集采用随机 60/20/20 划分；异配数据集使用标准划分。

模型。我们将 GHN 变体（**LSE**（扁平式，softmax检索）、**LSR**（扁平式，Epanechnikov 检索）、**Hier**(G)（分层式，包含 G 个组）和 **NoMem**（禁用记忆消融，仅采用拉普拉斯平滑））与 GCN（Kipf & Welling，2017年）、图注意力网络（Velickovic 等，2018年）、Graph-SAGE（Hamilton 等，2017年）、APPNP（Klicpera 等，2019年）、图同构网络（Xu 等，2019年）、多层感知机和 GPR-GNN（Chien 等，2021年）（针对异配性设计；结果取自原论文，采用相同的评估协议）进行比较。

训练。在每个数据集上逐模型进行超参数调优；每个模型根据验证准确率选出最佳配置。所有结果均为 10 个随机种子上的均值 $\pm$ 标准差。完整的训练协议和超参数网格见附录 A。

3.2 节点分类

表1：经过逐模型超参数调优的节点分类准确率（%）。**粗体**：每列最佳。 †：双峰（标准差为 >10 %）。

	Cora	CiteSeer	PubMed	Photo	Computers
GHN 变体					
LSE	82.0±0.6	70.9±0.8	77.9±0.4	94.5±0.3	91.3±0.2
LSR	82.0±0.6	70.9±1.0	78.1±0.4	94.4±0.2	91.3±0.4
Hier(8)	82.3±0.4	70.3±1.1	77.6±0.8	94.2±0.2	91.3±0.2
NoMem	80.3±0.4	69.8±1.1	77.2±0.6	**94.7±0.2**	**91.5±0.2**
基线方法					
GCN	82.2±0.4	72.2±0.5	79.4±0.4	89.6±0.3	73.7±13.0†
GAT	**82.8±0.8**	72.3±0.6	78.2±0.4	92.2±0.3	87.9±0.6
GraphSAGE	80.2±0.6	70.5±0.8	77.9±0.4	93.9±0.2	87.4±0.4
APPNP	**82.8±0.7**	**72.4±0.5**	**80.3±0.2**	92.2±0.3	41.9±13.1†
GIN	77.5±1.2	63.8±2.3	77.6±0.7	93.0±0.5	88.9±0.5
MLP	60.0±0.8	60.6±0.6	73.4±0.5	88.5±0.3	75.4±11.9†

迭代架构是 Amazon 上的主要驱动因素。所有 GHN 变体（包括禁用记忆的 NoMem 消融）在两个 Amazon 数据集上的表现均优于最佳标准基线（Photo 上高出 +0.8 个百分点，Computers 上高出 +2.6 个百分点）。GCN、APPNP 和 MLP 在 Computers 上出现双峰式训练崩溃（标准差为 >10%），这是已知的失效模式，而迭代能量下降架构完全避免了这一问题。在 Amazon 上，启用记忆的变体与 NoMem 的差距不到 0.5 个百分点，这表明在稠密图上（Photo：238k 条边，Cora：10.6k 条边），拉普拉斯平滑已能捕捉相关信号，记忆带来的额外价值有限。

记忆检索在稀疏引文图上能带来额外价值。在全部三个 Planetoid 数据集上，表现最佳的启用记忆变体均优于 NoMem：Cora 上为 Hier(8)（高出 +2.0 个百分点，$\pm 1\sigma$ 不重叠），CiteSeer 上为 LSR（高出 +1.1 个百分点），PubMed 上也为 LSR（高出 +0.9 个百分点）；不过，后两项差距的标准差存在重叠。GHN 变体在 Planetoid 上具有竞争力，但未能超越图注意力网络或 APPNP。

在中等条件下，LSR 比 LSE 更稳定 K。A $\beta \times K$ 相图（图4）显示，在 Amazon Photo 上，LSE 需要 K=256 才能收敛，而 LSR 在中等 β 条件下于 K=64保持稳定（Hoover 等，2025）。

3.3 对图损坏的鲁棒性

我们在 Amazon Photo 上测试了三种扰动类型：边删除、特征掩蔽和加性特征噪声。

Tab表 2：Amazon Photo 上 50% 损坏时的准确率（%）。干净基线可能与 Tab表 1 不同，因为这里单独进行了针对鲁棒性优化的超参数搜索。**粗体**：每列最佳。

	干净	边丢弃	特征掩蔽	特征噪声
GHN 变体				
LSE	**94.4±0.3**	93.7±0.4	89.4±1.8	93.9±0.3
LSR	94.3±0.2	93.7±0.5	91.0±2.4	94.1±0.2
Hier(8)	94.2±0.2	93.7±0.5	**91.9±1.1**	93.9±0.1
NoMem	**94.4±0.3**	**94.0±0.5**	86.9±3.1	**94.2±0.3**
基线方法				
GCN	89.7±0.3	89.1±0.5	81.2±1.5	89.7±0.3
GAT	92.2±0.3	91.2±0.5	81.4±5.0	92.2±0.2
GraphSAGE	93.9±0.2	93.0±0.3	73.3±1.8	93.6±0.3
APPNP	92.2±0.3	91.7±0.3	65.0±1.8	91.8±0.3
GIN	92.6±0.4	89.6±0.8	91.1±0.9	92.6±0.5
MLP	88.5±0.3	88.5±0.3	70.7±0.6	86.6±0.2

所有启用记忆的 GHN 变体在每种扰动类型下均位列前四，而 NoMem 仅在特征掩蔽情况下未能与它们并列；在该情况下，GIN 的结构无关聚合（91.1%）优于 NoMem（86.9%）（表2）。在边删除和特征噪声情况下，NoMem 的表现与启用记忆的变体相当，表明迭代拉普拉斯传播是主要的鲁棒性机制。**在特征掩蔽情况下，记忆带来了显著收益**。在掩蔽率为 50% 时，Hier(8) 仍保持 91.9% 的准确率，而 NoMem 为 86.9%，相差 5.0 个百分点；在干净数据上则不存在这一差距。这是存储模式能够补偿节点特征退化的最明确证据，而仅基于结构的平滑无法实现这一机制。

在 Planetoid 上，GHN 的鲁棒性因数据集而异（LSE 在 PubMed 上表现最佳；详情见附录 C.1）；图 2 展示了边删除曲线。

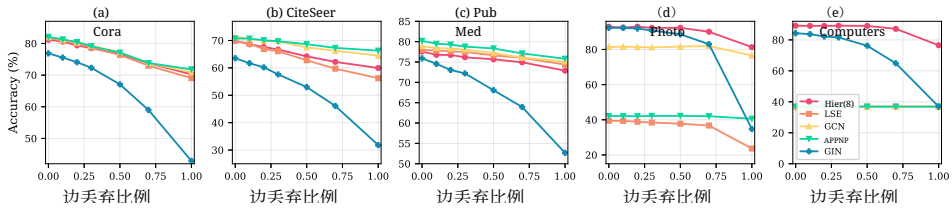

图 2：Amazon 和 Planetoid 数据集在逐步边删除下的准确率。GHN 变体在 Amazon 上的性能下降更为缓慢；Planetoid 的详情见附录 C.1。

3.4 异配图

在异配图上，拉普拉斯平滑（$\lambda > 0$）会将邻居推向相似的表示，从而可能损害分类。我们调节 $\lambda \in \{-0.1, -0.05, -0.01, 0\}$ 以实现图锐化（负 λ 会将邻居推开）。

表 3 显示，在调优负值 λ 后，GHN 变体的表现可与专为异配性设计的方法 GPR-GNN 相媲美，同时显著优于标准图神经网络（GCN、图注意力网络、APPNP），后者在图平滑作用下均会性能崩溃。我们报告了 GPR-GNN 在相同数据划分（†）上的已发表结果；我们没有重新调优或重新运行 GPR-GNN。GHN-LSE 在 Texas 上表现领先（ 81.4% 对 78.4% ），而 GPR-GNN 在 Wisconsin （ 82.9% 对 82.5% ）和康奈尔（80.3% 对 78.9%）上领先；所有差异都处于标准差重叠范围内。值得注意的是，GPR-GNN 的方差要高得多（例如，在康奈尔上为 ±8.1，而 GHN-LSR 为 ±1.7），这表明 GHN 在这些小型数据集的不同数据划分上更加稳定。其关键优势在于架构：GHN 能够同时处理

表 3：异配基准上的节点分类（%）。**粗体**：每列最佳结果。 †已发表结果（Chien 等人，2021年），采用相同的数据划分；未报告 Actor。

	Texas	Wisconsin	康奈尔	Actor
GHN 变体				
LSE	**81.4±1.5**	82.2±0.6	78.1±3.2	**38.0±0.5**
LSR	80.8±1.5	82.5±1.1	78.9±1.7	37.9±0.5
Hier(8)	80.5±3.3	82.2±1.7	78.1±2.0	37.8±0.6
NoMem	80.5±1.7	82.4±0.9	78.6±2.4	**38.0±0.7**
异配感知基线				
GPR-GNN†	78.4±4.4	**82.9±4.2**	**80.3±8.1**	–
标准基线				
MLP	80.3±1.8	82.2±0.6	73.0±0.0	37.6±0.4
GraphSAGE	78.4±1.8	80.0±0.8	76.2±1.7	37.0±0.5
GCN	70.3±1.3	55.7±1.0	40.3±0.9	28.8±0.4
GAT	68.6±2.3	58.0±1.9	43.8±4.0	29.4±0.6
APPNP	70.0±1.5	56.7±1.1	43.2±3.6	32.4±0.6
GIN	64.9±0.0	50.4±1.6	55.7±6.3	27.3±0.5

同质图和异配图，只需调优单个参数 λ；而 GPR-GNN 则需要学习针对图谱性质进行专门设计的 K 多项式滤波器系数。

所有最佳配置都选择 $\lambda \in \{-0.05, -0.01\}$；附录 C.2 展示了明显的相变（例如，康奈尔数据集上的 LSE 从 $\lambda{=}0$ 时的 73.5% 降至 $\lambda{=}0.3$ 时的 44.3%）。在 $\lambda \leq 0$ 时，启用记忆的变体与 NoMem 的表现相近，因为拉普拉斯项很小，且所有变体都简化为特征空间操作。

4 讨论与未来方向

本研究的核心发现是一种分解：对于图而言，迭代能量下降比下降所依据的对象更重要。NoMem 是一种移除 Hopfield 项的能量下降架构消融，在 Amazon 的两个数据集上均与启用记忆的变体表现相当或更优，在 Actor 上则打成平手；但在 Amazon 上，它仍优于所有标准基线。这表明，迭代式阻尼更新架构提供了强大的归纳偏置，能在前馈基线失效的图上稳定训练，并在边扰动下实现平稳退化。对于联想记忆，其优势取决于具体情形：记忆在稀疏图和特征掩蔽条件下能带来价值，而消融则阐明了架构与检索各自的作用。

记忆的作用更适合被理解为对图结构的替代，而非补充。在稠密图（Photo：23.8 万条边）上，拉普拉斯算子已能为分类提供足够的信号，因此记忆是冗余的。在稀疏图（Cora：1.06 万条边）或存在特征扰动时，结构信号不足，记忆便能弥补这一缺口：在干净的 Cora 上最多提升 2.0 个百分点，在 50% 特征掩蔽时最多提升 5.0 个百分点。这种替代视角解释了门控为何会学得近乎恒定的混合比例（附录 C.6）：记忆与结构之间的最佳平衡取决于图的情形，而非各个节点，因此静态门控便已足够。

局限性。 GHN 在干净的 Planetoid 图上不如图注意力网络/APPNP；在异配基准上，相较于多层感知机的提升有限，且通常在一个标准差范围内。平坦 LSE/LSR 需要 $K{=}256$ 才能避免在 Amazon 上发生坍塌；我们仅评估随机（而非对抗性）损坏；$O(NK)$ 检索会带来额外开销（跨数据集估算，每训练轮次的实际运行时间约为图注意力网络的 1.5–2× 倍）。收敛理论（附录 D）要求 $\beta\|\mathbf{M}\|^2 < 2$，但大多数数据集在训练运行点都不满足这一条件（表 11）；实践中，稳定性在 $T{=}4$ 次迭代内通过实验证实。值得探索的扩展包括按节点设置 λ、用固定的标量混合替代门控、对记忆模式进行对抗训练、采用稀疏检索以实现扩展，以及与图 Transformer（最直接的架构竞争者）进行比较。

参考文献

Eli Chien、Jianhao Peng、Pan Li 和 Olgica Milenkovic。自适应通用广义PageRank图神经网络。载于国际学习表征会议，2021年。Mete Demircigil、Judith Heusel、Matthias L¨owe、Sven Upgang 和 Franck Vermet。关于一种具有巨大存储容量的关联记忆模型。统计物理学杂志，168(2):288–299，2017年。doi: 10.1007/s10955-017-1816-y。URL https://arxiv.org/abs/1702.01929。Fangda Gu、Heng Chang、Weize Zhu、Somayeh Sojoudi 和 Laurent El Ghaoui。隐式图神经网络。载于神经信息处理系统进展，2020年。URL https: //arxiv.org/abs/2009.06211。

William L Hamilton、Rex Ying 和 Jure Leskovec。大型图上的归纳式表征学习。载于神经信息处理系统进展，第30卷，2017年。网址：https: //arxiv.org/abs/1706.02216。

Benjamin Hoover、Yuchen Liang、Bao Pham、Rameswar Panda、Hendrik Strobelt、Dmitry Krotov、Duen Horng Chau 和 Mohammed J Zaki。能量 Transformer。载于神经信息处理系统进展，2023年。网址：https://arxiv.org/abs/2302.07253。

Benjamin Hoover、Zhaoyang Shi、Krishnakumar Balasubramanian、Dmitry Krotov 和 Parikshit Ram。采用爱泼涅契科夫能量的稠密关联记忆。载于神经信息处理系统进展，2025年。网址：https://arxiv.org/abs/2506.10801。

Thomas N Kipf 和 Max Welling。《基于图卷积网络的半监督分类》。载于国际学习表征会议，2017年。网址：https:// arxiv.org/abs/1609.02907。

Johannes Klicpera、Aleksandar Bojchevski 和 Stephan G¨unnemann。《先预测后传播：图神经网络与个性化PageRank的结合》。载于国际学习表征会议，2019年。网址：https://arxiv.org/abs/1810.05997。

Dmitry Krotov。《层次化关联记忆》。*arXiv*预印本 *arXiv:2107.06446*，2021年。网址：https://arxiv.org/abs/2107.06446。

Dmitry Krotov 和 John J. Hopfield。《用于模式识别的稠密关联记忆》。载于神经信息处理系统进展，第29卷，2016年。网址：https://arxiv.org/ abs/1606.01164。

Qimai Li、Zhichao Han 和 Xiao-Ming Wu。深入探讨用于半监督学习的图卷积网络。发表于*AAAI*人工智能会议，2018年。网址：https: //arxiv.org/abs/1801.07606。

Yuchen Liang、Dmitry Krotov 和 Mohammed J Zaki。用于图嵌入的现代霍普菲尔德网络。*Frontiers inBig Data*，5，2022年。doi：10.3389/fdata.2022.1044709。网址：https: //arxiv.org/abs/2208.14376。

Carlo Lucibello 和 Marc M´ezard。稠密联想记忆的指数容量。物理评论快报，132(7):077301，2024年。doi：10.1103/PhysRevLett.132.077301。网址：https: //arxiv.org/abs/2304.14964。

Kenta Oono 和 Taiji Suzuki。图神经网络在节点分类中会指数级地丧失表达能力。载于国际学习表征会议，2020年。网址：https: //arxiv.org/abs/1905.10947。Hongbin Pei、Bingzhe Wei、Bairu Chang、Yunhe Lei 和 Bo Yang。Geom-GCN：几何图卷积网络。载于国际学习表征会议，2020年。网址：https://arxiv.org/abs/2002.05287。Ladislav Ramp´aˇsek、Mikhail Galkin、Vijay Prakash Dwivedi、Anh Tuan Luu、Guy Wolf 和 Do- minique Beaini。通用、强大且可扩展的图 Transformer 构建方案。载于神经信息处理系统进展，2022年。网址：https://arxiv.org/abs/2205.12454。

Hubert Ramsauer、Bernhard Sch¨afl、Johannes Lehner、Philipp Seidl、Michael Widrich、Thomas Adler、Lukas Gruber、Markus Holzleitner、Milena Pavlovi´c、Geir Kjetil Sandve 等。Hopfield 网络满足一切需求。载于国际学习表征会议，2021年。网址：https://arxiv.org/abs/2008.02217。Prithviraj Sen、Galileo Namata、Mustafa Bilgic、Lise Getoor、Brian Galligher 和 Tina Eliassi-Rad。网络数据中的集体分类。《*AI Magazine*》，29(3)：93–106，2008年。doi：10.1609/aimag. v29i3.2157。Oleksandr Shchur、Maximilian Mumme、Aleksandar Bojchevski 和 Stephan G¨unnemann。图神经网络评估中的陷阱。载于*NeurIPS*关系表征学习研讨会，2018年。网址：https://arxiv.org/abs/1811.05868。Jie Tang、Jimeng Sun、Chi Wang 和 Zi Yang。大规模网络中的社会影响力分析。载于*ACM SIGKDD*国际知识发现与数据挖掘会议，第807–816页，2009年。doi：10.1145/1557019.1557108。Jake Topping、Francesco Di Giovanni、Benjamin Paul Chamberlain、Xiaowen Dong 和 Michael M Bronstein。通过曲率理解图中的过度挤压与瓶颈。载于国际学习表征会议，2022年。网址：https://arxiv.org/abs/ 2111.14522。Petar Velickovic、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Li`o 和 Yoshua Bengio。图注意力网络。载于国际学习表征会议，2018年。网址：https://arxiv.org/abs/1710.10903。Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。图神经网络的能力有多强？载于国际学习表征会议，2019年。网址：https: //arxiv.org/abs/1810.00826。Chengxuan Ying、Tianle Cai、Shengjie Luo、Shuxin Zheng、Guolin Ke、Di He、Yanming Shen 和 Tie-Yan Liu。Transformer 在图表征上的表现真的很差吗？载于神经信息处理系统进展，2021年。网址：https://arxiv.org/abs/2106.05234。Daniel Z¨ugner、Amir Akbarnejad 和 Stephan G¨unnemann。针对图数据的神经网络对抗攻击。载于*ACM SIGKDD*国际知识发现与数据挖掘会议，第2847–2856页，2018年。网址：https://arxiv.org/abs/1805.07984。

A 实验细节

表4给出了完整的超参数网格。完整模型为编码器 $\rightarrow L$图 Hopfield 层（每层包含 T 次迭代和门控） $\rightarrow$分类器；检索使用绑定键值、跳跃连接（0.1）、丢弃和层归一化。模型特定的调优维度包括：GAT 注意力头 $\in \{4, 8\}$、APPNP $\alpha \in\{0。1、0.2\}$、GHN-LSE/LSR $K \in \{64, 256\}$。GHN 在同质性数据集上使用 λ=0.3 ，在异配数据集上使用 $\lambda \in\{-0.1, -0.05, -0.01, 0\}$ 。

表4：GHN 的超参数和训练协议。针对每个模型的超参数调优（表1）会根据所示网格范围，为每个数据集选择最佳配置。

参数	值/网格
GHN （所有变体）	
隐藏维度	$\{64, 128\}$
模式数量 K	$\{64, 256\}$ (LSE/LSR)
逆温度 β (初始值)	1.0（可学习）
拉普拉斯权重 λ	0.3
阻尼 α	0.3
每层迭代次数 T	4
层数 L	2
组数 G (Hier)	8
记忆头 H	$\{1, 2, 4, 8\}$ (消融；附录 C.5)
丢弃	$\{0.3, 0.5\}$
门偏置初始化	2.0
键/值	Tied
跳跃权重	0.1
训练（所有模型）	
学习率	$\{0.001, 0.005, 0.01\}$
权重衰减	$\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$
优化器	Adam
训练轮数	300
早停耐心值	50

Amazon Computers 上的基线模型坍塌。GCN、APPNP 和多层感知机在 Amazon Computers 上的双峰坍塌在整个超参数网格范围内均持续存在：我们遍历了学习率 $\{0.001, 0.005, 0.01\}$、权重衰减 $\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$ 和丢弃率 $\{0.3, 0.5\}$，没有任何配置能够消除这一失效模式。坍塌表现为部分随机种子正常收敛（$\sim$87%），而其他种子则发散至接近随机猜测的水平（$\sim$50%），从而产生了报告中的 $>$10% 标准差。这种不稳定性与已知现象相符：基于谱归一化的传播方法（GCN、APPNP）在具有异质度分布的图上十分敏感（Shchur 等，2018）；Amazon Computers 的度分布呈重尾（最大度数为 $>$400），这会加剧依赖初始化的收敛。我们没有采用标准实现之外的额外稳定化技术（例如批归一化、学习率预热），因为我们的目标是在统一的训练协议下比较不同架构，而不是单独优化每个基线模型的表现。

B 相关工作

Hopfield 网络与关联记忆。 现代（稠密）Hopfield 网络（Krotov & Hop- field，2016年；Demircigil 等，2017年；Ramsauer 等，2021年）能够存储指数级数量的模式，并且与 Transformer 注意力直接相关。其扩展包括分层记忆（Krotov，2021年）、能量 Transformer（Hoover 等，2023年）以及 Epanechnikov/LSR 能量（Hoover 等，2025年）。Liang 等人（2022年）将现代霍普菲尔德检索应用于图嵌入（图级分类），而我们关注的是通过联合能量进行节点 分类，该联合能量显式耦合检索与图平滑。

图 Transformer与内容–结构耦合。 图 Transformer（如 Graphormer（Ying 等，2021））及后续可扩展变体（Ramp´aˇsek 等，2022年）将基于内容的注意力与结构偏置相结合。我们的目标有所不同：不是替换消息

用完整的 Transformer 块进行传递，我们将节点更新表述为联合能量上的迭代下降步骤，从而使检索与平滑权衡明确且可分析。

隐式与迭代图模型。 隐式/平衡态图神经网络在图表示上求解不动点方程（Gu 等，2020）。我们的不动点视角与之相关，但算子源自受 Hopfield 启发的检索能量与拉普拉斯正则化，并为基础动力学得出了明确的平滑性/压缩条件（附录 D）。我们的迭代形式还与关于图传播中深度相关退化和表达能力损失的分析相关（Li 等，2018年；Oono 和 Suzuki，2020年）。

鲁棒图学习。 图神经网络鲁棒性通常是在对抗性结构扰动下进行研究的（Z¨ugner 等，2018年）。我们的研究设定与此互补：我们研究随机破坏，以剥离结构可靠性对模型准确率的影响，不作对抗性假设。我们的发现表明，聚合设计（求和、求均值或传送）在很大程度上决定了鲁棒性，这与对图神经网络信息流瓶颈的分析相关（Topping 等，2022年）。

C 附加结果

C.1 Planetoid 上的边删除

表 5 量化了在 Planetoid 数据集上删除边导致的相对准确率下降。GHN 的鲁棒性因数据集而异：在 PubMed 上，LSE 的表现最佳（下降 4.6%）；而在 CiteSeer 上，GHN 变体的鲁棒性较差（下降 14–20%），不及 GCN（10%）和 APPNP（6%）。GIN 在完全删边时尤其脆弱（损失 31–50%），这是因为它采用单射求和聚合（Xu 等，2019）；GCN（均值）和 APPNP（传送）则退化得更为平缓（5–14%）。在特征遮蔽下，GIN 的表现恰好相反：它较为鲁棒（遮蔽 50% 时准确率为 91.1%），但在边丢弃下很脆弱。这反映出单射聚合能最大程度地利用结构，而记忆检索则提供了一条不依赖结构的通道。

表 5：在 Planetoid 上删除边导致的相对准确率下降（%）。数值越小 = 鲁棒性越强。**粗体**：各数据集上最鲁棒的方法。

	50% 边丢弃			100% 边丢弃		
	Cora	CiteSeer	PubMed	Cora	CiteSeer	PubMed
GHN-LSE	−6.5	−10.2	**−1.8**	−15.5	−19.5	**−4.6**
GHN-Hier(8)	**−5.6**	−8.1	−2.5	−13.5	−14.2	−6.1
GCN	−5.9	−5.5	−2.2	−13.6	−9.7	−4.8
APPNP	−6.1	**−3.0**	−2.3	**−12.5**	**−6.3**	−5.5
GIN	−12.7	−16.5	−10.3	−44.1	−49.9	−30.5

跨数据集的特征掩蔽。图3展示了五个数据集上特征掩蔽造成的性能下降。与边删除不同，所有模型（包括 GHN ）在这些数据集上的性能都会随着特征被掩蔽而以相近的速率下降。不过，在 Amazon 数据集上，采用 $K{=}256$ 和调优后的超参数（表2）时，启用记忆的变体表现出比 NoMem 强得多的特征掩蔽鲁棒性（例如，在 Photo 上掩蔽 50% 特征时，Hier 的准确率为 91.9%，而 NoMem 为 86.9%）。这表明，更高的模式容量能够提供冗余表示，以弥补特征缺失；而在 $K{=}64$ 的 Planetoid 上未观察到这一优势。

LSE 与 LSR 相图。图4展示了 Amazon Photo 上的 $\beta \times K$ 相图。LSE（左）仅在 $K{=}256$时稳定；在 $K{\le}128$ 时，无论 β如何，都会出现双峰坍塌：部分随机种子能够收敛，其他则不能。LSR（右）在 $K{=}64$下针对 $\beta \in \{0.2, 0.5, 1.0, 5.0\}$可稳定达到 $\ge$93% 的准确率，但在极低的 β（< 0.2）以及 $\beta{=}2.0$时也会出现双峰现象。

记忆在 Amazon 上的贡献。在逐模型超参数调优下（表1），NoMem 在两个 Amazon 数据集上的表现都略优于所有启用记忆的变体（Photo 上为 94.7% 对 94.5%；Computers 上为 91.5% 对 91.3%）。这与记忆在 Planetoid 上始终带来提升的情况形成对比，并表明 Amazon 共同购买图具有更丰富的边结构（Photo：238k 条边；Cora：10.6k 条边），因此仅靠拉普拉斯平滑就已足够。GHN架构的

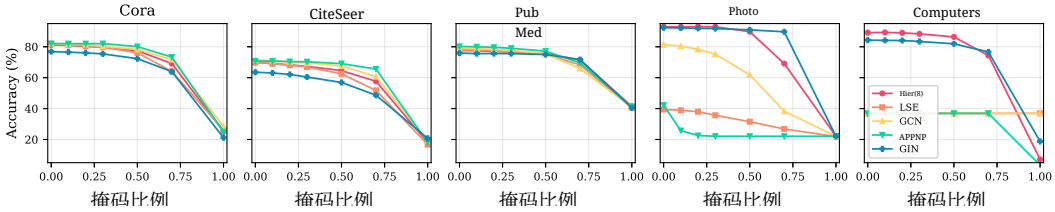

图3：五个数据集在特征掩蔽下的准确率。在 Plane-toid 上调优选择 K=64，在 Amazon 上选择 K=256。所有模型在 Planetoid 上的收敛速率相近；在 Amazon 上，启用记忆的变体表现出显著更强的鲁棒性（表 2）。

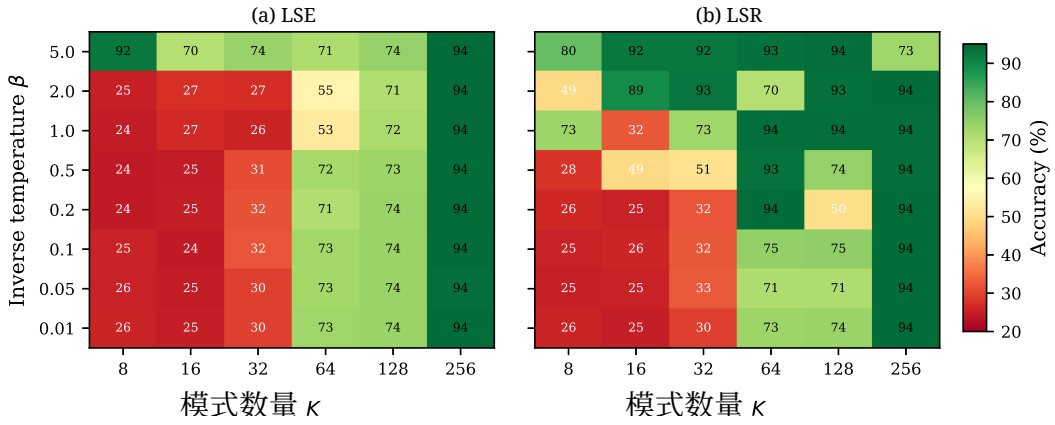

图4：Amazon Photo 上 LSE 与 LSR 能量的相图。每个单元格显示对 10 个随机种子计算的平均测试准确率。某些单元格中的高方差反映了双峰行为（随机种子要么收敛至 ∼94%，要么崩溃）。在 K=64 时，LSR 在中等程度的 β 下稳定，而 LSE 在所有测试的 β 下均呈双峰行为。

在 Amazon 上的优势来自其迭代拉普拉斯传播，而非记忆检索本身；所有 GHN 变体的表现都优于最佳标准基线 0.8–2.6 个百分点。

C.2 拉普拉斯权重 λ 消融实验

表 6 展示了 GHN-LSE 在所有七个数据集上随 $\lambda \in \{0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0\}$ 变化的准确率。在同质性数据集（Cora、CiteSeer、PubMed）上，准确率从 λ=0 开始单调上升，在 λ=0.2–0.3 处达到峰值，随后略有下降。在异配数据集（Texas、Wisconsin、康奈尔、Actor）上，随着 λ增加，准确率单调下降，这证实了当邻居标签不同时，拉普拉斯平滑会产生负面影响。

表 6：GHN-LSE 测试准确率（%）随 λ变化。**粗体**：每个数据集上的最佳 λ。

数据集	0	0.01	0.05	0.1	0.2	0.3	0.5	1.0
Cora	57.7	62.7	71.7	76.1	80.2	**81.8**	81.6	80.6
CiteSeer	55.6	59.1	65.1	67.6	**70.2**	69.6	68.9	69.4
PubMed	73.0	73.2	74.5	75.2	**78.2**	77.7	76.9	76.7
Texas	**74.1**	71.6	70.0	64.9	66.8	67.0	66.8	64.9
Wisconsin	79.8	**80.8**	72.2	67.1	60.0	55.1	45.3	44.9
康奈尔	**73.5**	72.4	63.8	46.2	41.1	44.3	42.2	41.4
Actor	**37.5**	37.3	36.5	35.5	30.8	28.1	25.5	25.0

其他 GHN 变体（LSR、Hier、NoMem）的结果也呈现出相同的模式。在 Cora 上，GHN-LSR 在 λ=0.3（81.5%）处达到峰值，GHN-Hier(8) 则在 λ=0.5（81.7%）处达到峰值。在异配数据集上，所有变体都在 λ=0 或 λ=0.01处取得最佳性能。

C.3 负 λ（图锐化）

设置 $\lambda < 0$会反转拉普拉斯正则化：它不再将邻居拉近，而是将它们推开（“图锐化”）。表 7 显示，较小的负值（$\lambda \in \{-0.05, -0.01\}$）能够在异配数据集上优于 λ=0。

表7：在异配数据集上使用负 λ 的GHN准确率（%）。每个数据集-模型的最佳结果以**粗体**标出。

		λ=−0.3	λ=−0.1	λ=−0.05	λ=−0.01
Texas	LSE	75.9	77.3	**78.6**	74.6
	LSR	75.9	77.3	**79.5**	78.6
Wisconsin	LSE	76.9	81.0	**81.8**	79.0
	LSR	78.8	**80.8**	80.4	79.0
康奈尔	LSE	61.9	66.5	74.6	**77.0**
	LSR	63.2	69.5	74.3	**77.8**
Actor	LSE	34.7	37.5	**38.0**	37.5
	LSR	34.9	37.2	37.3	**37.8**

最优负值 λ 取决于数据集： -0.05 在 Texas 和 Wisconsin 上效果最佳，而 Cornell 则更适合 -0.01 。与 λ=0 相比，性能提升幅度不大，但较为稳定（0.5–3.5 个百分点）。请注意，此处的准确率与表 3 中的结果不同，因为这项消融实验只改变了 λ，其他超参数（隐藏维度、丢弃、K）则固定为单一默认配置；而表 3 报告的是在完整网格上对各模型进行超参数调优的结果。非常负的取值（λ=−0.3）会降低性能，这表明过度锐化与过度平滑一样有害。

C.4 迭代次数消融

表8展示了每层的 Hopfield 更新迭代次数 T 如何影响一个同质性数据集（Cora）和两个异质性数据集（Texas、Wisconsin）上的准确率，使用的是 λ=0.3。

表8：使用 λ=0.3时，GHN-LSE 的准确率（%）与迭代次数 T 的关系。**最佳**：最佳 T。

	T=1	T=2	T=4	T=8	T=16
Cora	74.5±1.1	78.4±0.5	**81.8±1.4**	56.1±26.1	30.9±2.0
Texas	**73.5±2.8**	72.2±2.9	67.0±1.7	65.9±1.4	65.4±1.1
Wisconsin	**63.7±1.0**	58.6±2.3	55.1±5.9	44.1±1.9	46.3±1.4

在 Cora 上，LSE 在 T=4时达到峰值，并在 $T \geq 8$时崩溃（ T=8处较大的标准差表明收敛呈双峰分布）。在异配数据集上， T=1 表现最佳；使用 λ=0.3进行更多迭代会带来逐渐加剧的有害平滑。这与 λ 消融实验结果一致：在异配图上，尽可能减小拉普拉斯项的影响（无论是设置 λ=0 还是T=1）都至关重要。

C.5 记忆头 H 消融

表9展示了 GHN-LSE 在四个数据集上使用 $H \in \{1, 2, 4, 8\}$ 记忆头时的准确率。

表9：GHN-LSE 准确率（%）与记忆头数量 H的关系。

	H=1	H=2	H=4	H=8
Cora	81.5±0.9	81.6±1.0	81.8±1.4	80.3±0.4
CiteSeer	69.6±1.4	70.0±1.6	69.6±0.7	69.5±1.5
PubMed	77.6±0.5	77.6±0.6	77.7±0.9	77.9±1.0
Texas	64.9±1.8	65.9±2.9	67.0±1.7	67.6±0.0

记忆头的数量对准确率影响很小：在全部四个数据集中，最佳和最差 H 之间的差异在 Cora 上最多为 1.5 个百分点，在 Texas 上最多为 2.7 个百分点。这支持将 $H{=}1$ 作为默认设置，以简化操作且不牺牲性能。

C.6 特征损坏下的门控分析

我们在 Amazon Photo 上训练 GHN-LSE（300 个训练轮次），并提取特征掩蔽级别为 0%、10%、30%、50% 和 70% 时学到的门控值 $g_v = \sigma(W_g[\mathbf{x}_v \| \hat{\mathbf{x}}_v] + b_g)$。表10报告了各损坏级别下的平均门控值（对节点和迭代次数取平均）。

表10：特征掩蔽程度逐渐增加时，GHN-LSE 在 Amazon Photo 上的平均门控值 $\bar{g}$。数值越高 g，表示赋予记忆检索的权重越大。

掩码 %	0%	10%	30%	50%	70%
$\bar{g}$	$0.727_{\pm 0.039}$	$0.727_{\pm 0.039}$	$0.726_{\pm 0.038}$	$0.725_{\pm 0.038}$	$0.723_{\pm 0.037}$
准确率 (%)	$93.6_{\pm 0.5}$	$93.6_{\pm 0.5}$	$93.2_{\pm 0.5}$	$90.6_{\pm 1.3}$	$67.9_{\pm 12.8}$

即使在严重损坏下准确率大幅下降，门控仍几乎保持不变（$\Delta\bar{g} \leq 0.004$）。这表明门控不会在推理时动态调整记忆与特征的混合比例；相反，即便节点特征受到损坏，存储的原型模式仍能提供有用的类别级信息，从而带来鲁棒性。

D 基础 LSE 动力学的理论分析

对于单个查询，标准 Hopfield 能量在异步更新下保证递减（Ramsauer 等，2021）。我们的耦合方案对所有节点进行同步更新，并通过拉普拉斯项引入节点间耦合，因此这一保证并不直接适用。下面我们针对采用 LSE 检索的基础 GHN能量，以及在迭代更新期间固定的（$\mathbf{M}$ β, λ, $\mathbf{L}$），给出形式化结果：梯度下降和不动点更新的显式光滑性与压缩条件。这些结果并未涵盖包含门控、层次路由和联合学习参数的完整训练时系统；对于完整模型，我们依赖经验稳定性（所有实验中均在 $T{=}4$ 次迭代内稳定收敛）。理论假设 $\lambda \geq 0$；对于 $\lambda < 0$（在异配基准上进行图锐化），拉普拉斯项会变得不定，因此我们的凸性和压缩保证不再适用。在这些实验中，我们将 $|\lambda|$ 保持较小（$\lambda \in \{-0.1, -0.05, -0.01\}$），并依赖阻尼、早停和经验稳定性。表 11 报告了在训练运行点上、对 5 个随机种子取平均后的 $\beta\|\mathbf{M}\|_\sigma^2$（学习得到的 β 与$\mathbf{M}$的谱范数平方的乘积）。凸性条件 $\beta\|\mathbf{M}\|^2 < 2$ 仅在康奈尔（0.75）和 CiteSeer（1.88）上满足；其他所有数据集均超过阈值，其中 Amazon Photo 达到 35.96。这意味着在典型使用情况下，该系统远远超出命题 D.5 所述的强凸区域；尽管如此，所有实验中的梯度下降仍能在 $T{=}4$ 次迭代内可靠收敛，这表明更精细的分析（例如吸引子附近的局部凸性或 Łojasiewicz型论证）或许能弥合这一差距。

表 11：在训练运行点学习得到的 $\beta\|\mathbf{M}\|_\sigma^2$（5 个随机种子的均值）。凸性条件要求该乘积为 <2。

数据集	β	$\|\mathbf{M}\|_\sigma^2$	$\beta\|\mathbf{M}\|_\sigma^2$
Cora	1.11	1.94	2.15
CiteSeer	1.09	1.71	**1.88**
PubMed	1.36	5.24	6.54
Amazon Photo	0.99	36.37	35.96
Amazon Comp.	1.12	2.11	2.36
Texas	1.25	3.88	4.65
Wisconsin	1.05	2.48	2.60
康奈尔	1.07	0.68	**0.75**
Actor	1.00	4.61	4.72

D.1 设置

设 $\mathbf{X} \in \mathbb{R}^{N\times d}$ 为以 $\mathbf{x}_v$ 为行的节点表示矩阵，并设 $\mathbf{M} \in \mathbb{R}^{K\times d}$ 为记忆矩阵。对于固定的 $\beta > 0$ 和 $\lambda \geq 0$，定义

$$E_{\text{base}}(\mathbf{X}) = \sum_{v=1}^{N}\left[-\beta^{-1}\log\sum_{\mu=1}^{K}\exp\left(\beta\,\mathbf{m}_\mu^\top\mathbf{x}_v\right) + \frac{1}{2}\|\mathbf{x}_v\|^2\right] + \lambda\,\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}), \tag{4}$$

其中 $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ 是对称归一化拉普拉斯。

命题 D.1（归一化拉普拉斯二次型）。若 $\mathbf{A}$为对称矩阵，且 $\mathbf{D}_{vv} = d_v > 0$，则

$$\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) = \frac{1}{2}\sum_{u,v}A_{uv}\left\|\frac{\mathbf{x}_u}{\sqrt{d_u}} - \frac{\mathbf{x}_v}{\sqrt{d_v}}\right\|^2 \geq 0.$$

因此$\mathbf{L} \succeq 0$，拉普拉斯项是凸的。

证明。定义 $\mathbf{Y} := \mathbf{D}^{-1/2}\mathbf{X}$，因此 $\mathbf{y}_v = \mathbf{x}_v/\sqrt{d_v}$。利用 $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$，

$$\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) = \mathrm{tr}(\mathbf{X}^\top\mathbf{X}) - \mathrm{tr}(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}) = \sum_v\|\mathbf{x}_v\|^2 - \sum_{u,v}A_{uv}\,\mathbf{y}_u^\top\mathbf{y}_v.$$

现在展开

$$\frac{1}{2}\sum_{u,v}A_{uv}\|\mathbf{y}_u - \mathbf{y}_v\|^2 = \frac{1}{2}\sum_{u,v}A_{uv}\left(\|\mathbf{y}_u\|^2 + \|\mathbf{y}_v\|^2 - 2\mathbf{y}_u^\top\mathbf{y}_v\right) \tag{5}$$

$$= \sum_u\left(\sum_v A_{uv}\right)\|\mathbf{y}_u\|^2 - \sum_{u,v}A_{uv}\mathbf{y}_u^\top\mathbf{y}_v \tag{6}$$

$$= \sum_u d_u\left\|\frac{\mathbf{x}_u}{\sqrt{d_u}}\right\|^2 - \sum_{u,v}A_{uv}\frac{\mathbf{x}_u^\top\mathbf{x}_v}{\sqrt{d_u d_v}} \tag{7}$$

$$= \sum_u\|\mathbf{x}_u\|^2 - \sum_{u,v}A_{uv}\frac{\mathbf{x}_u^\top\mathbf{x}_v}{\sqrt{d_u d_v}} \tag{8}$$

$$= \mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}). \tag{9}$$

这就证明了该恒等式。右侧是由非负权重加权的平方范数之和，因此非负。

□

D.2 梯度与海森矩阵结构

命题 D.2（梯度公式）。对于每个节点 v，

$$\nabla_{\mathbf{x}_v}E_{\text{base}}(\mathbf{X}) = -\mathbf{M}^\top\mathrm{softmax}(\beta\,\mathbf{M}\mathbf{x}_v) + \mathbf{x}_v + 2\lambda(\mathbf{L}\mathbf{X})_v. \tag{10}$$

因此，式 *2* 中的阻尼更新恰好是在 E_{base} 上以步长 α 进行的梯度下降。

证明。设

$$E_{\text{base}}(\mathbf{X}) = \sum_{u=1}^{N}e_u(\mathbf{x}_u) + \lambda\,\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}), \quad e_u(\mathbf{x}) = -\beta^{-1}\log\sum_{\mu=1}^{K}e^{\beta\mathbf{m}_\mu^\top\mathbf{x}} + \frac{1}{2}\|\mathbf{x}\|^2.$$

只有 $e_v(\mathbf{x}_v)$ 和拉普拉斯项依赖于 $\mathbf{x}_v$。对于检索，设 $\mathbf{z} = \mathbf{M}\mathbf{x}_v$，并使用 $\nabla_{\mathbf{z}}\left[\beta^{-1}\log\sum_\mu e^{\beta z_\mu}\right] = \mathrm{softmax}(\beta\mathbf{z})$。根据链式法则可得 $\nabla_{\mathbf{x}_v}\mathrm{lse}(\beta,\mathbf{M}\mathbf{x}_v) = \mathbf{M}^\top\mathrm{softmax}(\beta\mathbf{M}\mathbf{x}_v)$。二次项贡献$\mathbf{x}_v$。对于拉普拉斯项，

$$\nabla_{\mathbf{X}}\mathrm{tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) = (\mathbf{L} + \mathbf{L}^\top)\mathbf{X} = 2\mathbf{L}\mathbf{X},$$

因为 $\mathbf{L}$ 是对称的。合并各项得到公式（10）。接着

$$\mathbf{x}_v^{(t+1)} = \mathbf{x}_v^{(t)} - \alpha\nabla_{\mathbf{x}_v}E_{\text{base}}(\mathbf{X}^{(t)}) = (1-\alpha)\mathbf{x}_v^{(t)} + \alpha\left[\mathbf{M}^\top\mathrm{softmax}(\beta\mathbf{M}\mathbf{x}_v^{(t)}) - 2\lambda(\mathbf{L}\mathbf{X}^{(t)})_v\right].$$

□

引理 D.3 (检索雅可比矩阵)。 设 $s(\mathbf{x}) = \mathbf{M}^\top \mathrm{softmax}(\beta \mathbf{M}\mathbf{x})$ 且 $\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\beta \mathbf{M}\mathbf{x})$。则

$$\nabla s(\mathbf{x}) = \beta\, \mathbf{M}^\top \Sigma(\mathbf{p}(\mathbf{x}))\mathbf{M}, \quad \Sigma(\mathbf{p}) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top.$$

证明。 对于 softmax $\mathbf{p}$，其分量雅可比为

$$\frac{\partial p_i}{\partial z_j} = p_i(\delta_{ij} - p_j),$$

因此写成矩阵形式为 $\nabla_{\mathbf{z}} \mathrm{softmax}(\mathbf{z}) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$。当 $\mathbf{z} = \beta \mathbf{M}\mathbf{x}$ 时，

$$\nabla s(\mathbf{x}) = \mathbf{M}^\top \left(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\right)(\beta \mathbf{M}) = \beta\, \mathbf{M}^\top \Sigma(\mathbf{p})\mathbf{M}.$$

□

引理 D.4 （Softmax 协方差界）。 对于任意概率向量 $\mathbf{p} \in \Delta^{K-1}$，定义 $\Sigma(\mathbf{p}) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$。则 $\Sigma(\mathbf{p}) \succeq 0$ 且 $\|\Sigma(\mathbf{p})\| \le \frac{1}{2}$。

证明。 $\Sigma(\mathbf{p})$ 是类别独热变量的协方差矩阵，因此为半正定。对于单位向量 $\mathbf{u}$，

$$\mathbf{u}^\top \Sigma(\mathbf{p})\mathbf{u} = \sum_i p_i u_i^2 - \left(\sum_i p_i u_i\right)^2 = \mathrm{Var}(u_I) \le \frac{(\max_i u_i - \min_i u_i)^2}{4} \le \frac{1}{2},$$

由波波维丘不等式以及 $\max_i u_i - \min_i u_i \le \sqrt{2}\|\mathbf{u}\| = \sqrt{2}$。对单位向量 $\mathbf{u}$ 取上确界可得 $\|\Sigma(\mathbf{p})\| \le \frac{1}{2}$。 □

命题 D.5 (凸性与强凸性区域)。 设 $\mu := 1 - \beta \|\mathbf{M}_2\|^2$。那么海森矩阵满足

$$\nabla^2 E_{\mathrm{base}}(\mathbf{X}) \succeq \mu\, \mathbf{I}_{Nd} + 2\lambda(\mathbf{L} \otimes \mathbf{I}_d).$$

因此：

1. 若 $\beta\|\mathbf{M}\|^2 \le 2$，$E_{\mathrm{base}}$ 是凸的；*2.* 若 $\beta\|\mathbf{M}\|^2 < 2$，$E_{\mathrm{base}}$ 是 μ-强凸的。

证明。 对于每个节点，

$$\nabla^2_{\mathbf{x}_v} \mathrm{lse}(\beta, \mathbf{M}\mathbf{x}_v) = \beta\, \mathbf{M}^\top \Sigma(\mathbf{p}_v)\mathbf{M}, \quad \mathbf{p}_v := \mathrm{softmax}(\beta \mathbf{M}\mathbf{x}_v).$$

根据引理 D.4，

$$\mathbf{0} \preceq \beta \mathbf{M}^\top \Sigma(\mathbf{p}_v)\mathbf{M} \preceq \frac{\beta}{2}\|\mathbf{M}\|^2 \mathbf{I}_d.$$

因此

$$\nabla^2_{\mathbf{x}_v}\left[-\mathrm{lse}(\beta, \mathbf{M}\mathbf{x}_v) + \frac{1}{2}\|\mathbf{x}_v\|^2\right] = \mathbf{I}_d - \beta \mathbf{M}^\top \Sigma(\mathbf{p}_v)\mathbf{M} \succeq \left(1 - \frac{\beta\|\mathbf{M}\|^2}{2}\right)\mathbf{I}_d.$$

将所有节点堆叠起来，并加上拉普拉斯海森矩阵的贡献 $2\lambda(\mathbf{L} \otimes \mathbf{I}_d) \succeq 0$，得到

$$\nabla^2 E_{\mathrm{base}}(\mathbf{X}) \succeq \left(1 - \frac{\beta\|\mathbf{M}\|^2}{2}\right)\mathbf{I}_{Nd} + 2\lambda(\mathbf{L} \otimes \mathbf{I}_d).$$

这表明当 $\beta\|\mathbf{M}\|^2 \le 2$ 时具有凸性，而当 $\beta\|\mathbf{M}\|^2 < 2$ 时具有 μ-强凸性。 □

D.3 光滑性与极小化子的存在性

命题 D.6(L-光滑性)。 ∇E_{base} 在弗罗贝尼乌斯范数下是利普希茨的:

$$\|\nabla E_{\text{base}}(\mathbf{X}) - \nabla E_{\text{base}}(\mathbf{Y})\|_F \le L_{\text{lip}} \|\mathbf{X} - \mathbf{Y}\|_F,$$

with

$$L_{\text{lip}} = \frac{\beta}{2}\|\mathbf{M}\|^2 + 1 + 2\lambda\|\mathbf{L}\|.$$

证明。 定义逐行检索 $s(\mathbf{x}) = \mathbf{M}^\top \text{softmax}(\beta \mathbf{M}\mathbf{x})$ 和 $[S(\mathbf{X})]_v = s(\mathbf{x}_v)$。根据命题 D.2,

$$\nabla E_{\text{base}}(\mathbf{X}) = \mathbf{X} - S(\mathbf{X}) + 2\lambda \mathbf{L}\mathbf{X}.$$

因此

$$\|\nabla E(\mathbf{X}) - \nabla E(\mathbf{Y})\|_F \le \|\mathbf{X} - \mathbf{Y}\|_F + \|S(\mathbf{X}) - S(\mathbf{Y})\|_F + 2\lambda\|\mathbf{L}\| \, \|\mathbf{X} - \mathbf{Y}\|_F \tag{11}$$

$$= (1 + 2\lambda\|\mathbf{L}\|) \, \|\mathbf{X} - \mathbf{Y}\|_F + \|S(\mathbf{X}) - S(\mathbf{Y})\|_F. \tag{12}$$

由引理 D.3 和引理 D.4,

$$\|\nabla s(\mathbf{x})\| = \|\beta \mathbf{M}^\top \Sigma(\mathbf{p}) \mathbf{M}\| \le \beta \|\mathbf{M}\|^2 \|\Sigma(\mathbf{p})\| \le \frac{\beta}{2}\|\mathbf{M}\|^2.$$

因此，根据中值定理， s 是 $\frac{\beta}{2}\|\mathbf{M}\|^2$-利普希茨的:

$$\|s(\mathbf{x}) - s(\mathbf{y})\| \le \frac{\beta}{2}\|\mathbf{M}\|^2 \|\mathbf{x} - \mathbf{y}\|.$$

逐行应用:

$$\|S(\mathbf{X}) - S(\mathbf{Y})\|_F^2 \le \left(\frac{\beta}{2}\|\mathbf{M}\|^2\right)^2 \|\mathbf{X} - \mathbf{Y}\|_F^2,$$

so

$$\|S(\mathbf{X}) - S(\mathbf{Y})\|_F \le \frac{\beta}{2}\|\mathbf{M}\|^2 \|\mathbf{X} - \mathbf{Y}\|_F.$$

代入式 12，即可得到所述常数。 □

推论 D.7 (步长稳定性条件)。 对于梯度下降$\mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} - \eta \nabla E_{\text{base}}(\mathbf{X}^{(t)})$，任意 $\eta \in (0, 2/L_{\text{lip}})$ 都能保证单调下降。由于归一化拉普拉斯算子的 $\|\mathbf{L}\| \le 2$，一个方便的充分界为

$$\eta < \frac{2}{\frac{\beta}{2}\|\mathbf{M}\|^2 + 1 + 4\lambda}.$$

证明。 第一项结论是标准的光滑梯度步长条件。第二项由命题 D.6 和 $\|\mathbf{L}\| \le 2$ 得出。

□

命题 D.8 (强制性与极小化者的存在性)。 $E_{\text{base}}(\mathbf{X}) \to +\infty$ 当 $\|\mathbf{X}\|_F \to \infty$ 时。因此 E_{base} 至少能取得一个全局极小化者。

证明。 对于每个节点，令 $\mathbf{z} = \mathbf{M}\mathbf{x}_v$,

$$\text{lse}(\beta, \mathbf{M}\mathbf{x}_v) = \beta^{-1} \log \sum_{\mu=1}^{K} e^{\beta z_\mu} \le \beta^{-1} \left(\log K + \beta \max_\mu z_\mu\right) \le \beta^{-1} \log K + \|\mathbf{M}\| \, \|\mathbf{x}_v\|.$$

因此

$$E_{\text{base}}(\mathbf{X}) \ge \frac{1}{2}\|\mathbf{X}\|_F^2 - \|\mathbf{M}\| \sum \|\mathbf{x}_v\| - N\beta^{-1} \log K + \lambda \text{tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X}).$$

利用 $\sum_v \|\mathbf{x}_v\| \le \sqrt{N}\|\mathbf{X}\|_F$ 以及来自命题 D.1 的 $\text{tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X}) \ge 0$,

$$E_{\text{base}}(\mathbf{X}) \ge \frac{1}{2}\|\mathbf{X}\|_F^2 - \|\mathbf{M}\|\sqrt{N} \, \|\mathbf{X}\|_F - N\beta^{-1} \log K.$$

右侧是关于 $\|\mathbf{X}\|_F$ 的强制性二次下界，因此当 $\|\mathbf{X}\|_F \to \infty$ 趋于 $+\infty$ as 时，它会趋于无穷大。因此，E_{base} 是强制的。由连续性可知，它能达到一个全局极小化者。 □

D.4 梯度下降收敛

定理 D.9（单调下降与收敛至平稳性）。 假设 $\eta \in (0,\ 2/L_{\text{lip}})$ 并进行迭代

$$\mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} - \eta \nabla E_{\text{base}}(\mathbf{X}^{(t)}).$$

那么：

1. $E_{\text{base}}(\mathbf{X}^{(t)})$是非增的，且

$$E_{\text{base}}(\mathbf{X}^{(t+1)}) \le E_{\text{base}}(\mathbf{X}^{(t)}) - \eta \left(1 - \frac{\eta L_{\text{lip}}}{2}\right) \|\nabla E_{\text{base}}(\mathbf{X}^{(t)})\|_F^2;$$

2. $\sum_{t=0}^{\infty} \|\nabla E_{\text{base}}(\mathbf{X}^{(t)})\|_F^2 < \infty$，因此 $\|\nabla E_{\text{base}}(\mathbf{X}^{(t)})\|_F \to 0$；

3. 每个聚点都是驻点。

证明。 根据命题 D.6， E_{base} 是 L_{lip}-光滑的，因此对于任意$\mathbf{X}$,$\mathbf{Y}$,

$$E(\mathbf{Y}) \le E(\mathbf{X}) + \langle \nabla E(\mathbf{X}), \mathbf{Y} - \mathbf{X} \rangle + \frac{L_{\text{lip}}}{2} \|\mathbf{Y} - \mathbf{X}\|_F^2. \tag{13}$$

令$\mathbf{Y} = \mathbf{X} - \eta \nabla E(\mathbf{X})$:

$$E(\mathbf{X} - \eta \nabla E(\mathbf{X})) \le E(\mathbf{X}) - \eta \|\nabla E(\mathbf{X})\|_F^2 + \frac{L_{\text{lip}} \eta^2}{2} \|\nabla E(\mathbf{X})\|_F^2 \tag{14}$$

$$= E(\mathbf{X}) - \eta \left(1 - \frac{\eta L_{\text{lip}}}{2}\right) \|\nabla E(\mathbf{X})\|_F^2. \tag{15}$$

在$\mathbf{X} = \mathbf{X}^{(t)}$ 处应用此式即可证明第 (1) 项。令 $c := \eta(1 - \eta L_{\text{lip}}/2) > 0$。从 $t = 0$ 到$T - 1$ 求和可得

$$E(\mathbf{X}^{(0)}) - E(\mathbf{X}^{(T)}) \ge c \sum_{t=0}^{T-1} \|\nabla E(\mathbf{X}^{(t)})\|_F^2.$$

利用下有界性（命题 D.8）并令 $T \to \infty$，可得$\sum_{t=0}^{\infty} \|\nabla E(\mathbf{X}^{(t)})\|_F^2 < \infty$，因此范数趋于零：这就证明了第 (2) 项。对于第 (3) 项，若$\mathbf{X}^{(t_k)} \to \mathbf{X}^\star$，则由 ∇E的连续性，

$$\nabla E(\mathbf{X}^\star) = \lim_{k \to \infty} \nabla E(\mathbf{X}^{(t_k)}) = \mathbf{0}.$$

□

定理 D.10（收敛速率）。 在定理 *D.9* 的假设下，令 $E_{\text{inf}} := \inf_{\mathbf{X}} E_{\text{base}}(\mathbf{X})$ 且$c := \eta(1 - \eta L_{\text{lip}}/2)$。那么对于任意 $T \ge 1$,

$$\min_{0 \le t \le T-1} \|\nabla E_{\text{base}}(\mathbf{X}^{(t)})\|_F^2 \le \frac{E_{\text{base}}(\mathbf{X}^{(0)}) - E_{\text{inf}}}{cT}.$$

若 $\beta\|\mathbf{M}\|^2 < 2$（因此 E_{base} 是 μ-强凸的，且 $\mu = 1 - \beta\|\mathbf{M}\|^2/2$），那么对于 $\eta \in (0,\ 1/L_{\text{lip}}]$ 以及极小化者$\mathbf{X}^\star$,

$$\|\mathbf{X}^{(t)} - \mathbf{X}^\star\|_F \le (1 - \eta\mu)^t \|\mathbf{X}^{(0)} - \mathbf{X}^\star\|_F,$$

and

$$E_{\text{base}}(\mathbf{X}^{(t)}) - E_{\text{base}}(\mathbf{X}^\star) \le (1 - \eta\mu)^t \left(E_{\text{base}}(\mathbf{X}^{(0)}) - E_{\text{base}}(\mathbf{X}^\star)\right).$$

证明。 对于第一个论断，定理 D.9 给出

$$E(\mathbf{X}^{(t+1)}) \le E(\mathbf{X}^{(t)}) - c\|\nabla E(\mathbf{X}^{(t)})\|_F^2.$$

对 $t = 0$至 $T - 1$求和，并使用 $E(\mathbf{X}^{(T)}) \ge E_{\text{inf}}$,

$$c \sum_{t=0}^{T-1} \|\nabla E(\mathbf{X}^{(t)})\|_F^2 \le E(\mathbf{X}^{(0)}) - E_{\text{inf}}.$$

除以 T，并用最小值对平均值进行下界估计。

对于强凸性，令 $\mathbf{X}^\star \in \arg\min_{\mathbf{X}} E(\mathbf{X})$（唯一）。根据光滑性和强凸性，对于每个 t，存在对称的 $\mathbf{A}^{(t)}$，满足 $\mu\mathbf{I} \preceq \mathbf{A}^{(t)} \preceq L_{\text{lip}}\mathbf{I}$，使得

$$\nabla E(\mathbf{X}^{(t)}) - \nabla E(\mathbf{X}^\star) = \mathbf{A}^{(t)}(\mathbf{X}^{(t)} - \mathbf{X}^\star).$$

由于 $\nabla E(\mathbf{X}^\star) = 0$,

$$\mathbf{X}^{(t+1)} - \mathbf{X}^\star = \left(\mathbf{I} - \eta\mathbf{A}^{(t)}\right)(\mathbf{X}^{(t)} - \mathbf{X}^\star).$$

因此

$$\|\mathbf{X}^{(t+1)} - \mathbf{X}^\star\|_F \le \max_{\lambda\in[\mu,L_{\text{lip}}]} |1-\eta\lambda| \, \|\mathbf{X}^{(t)} - \mathbf{X}^\star\|_F.$$

对于 $\eta \le 1/L_{\text{lip}}$，最大值为 $1-\eta\mu$，由此得到迭代速率。对于函数值，combin e

$$E(\mathbf{X}^{(t+1)}) - E(\mathbf{X}^\star) \le E(\mathbf{X}^{(t)}) - E(\mathbf{X}^\star) - \eta\left(1 - \frac{\eta L_{\text{lip}}}{2}\right)\|\nabla E(\mathbf{X}^{(t)})\|_F^2,$$

对于 μ-强凸函数，具有梯度支配,

$$\|\nabla E(\mathbf{X}^{(t)})\|_F^2 \ge 2\mu\left(E(\mathbf{X}^{(t)}) - E(\mathbf{X}^\star)\right),$$

并 $\eta \le 1/L_{\text{lip}}$ 得出结论

$$E(\mathbf{X}^{(t+1)}) - E(\mathbf{X}^\star) \le (1-\eta\mu)\left(E(\mathbf{X}^{(t)}) - E(\mathbf{X}^\star)\right).$$

□

D.5 不动点压缩与阻尼

定义

$$[T(\mathbf{X})]_v = \mathbf{M}^\top \text{softmax}(\beta\mathbf{M}\mathbf{x}_v) - 2\lambda(\mathbf{L}\mathbf{X})_v.$$

注D.11（梯度下降与不动点迭代）。区分以下两种方案很有用：

1. **梯度下降**：$\mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} - \eta\nabla E_{\text{base}}(\mathbf{X}^{(t)})$。在 L_{lip}-光滑性和 $\eta \in (0, 2/L_{\text{lip}})$ 条件下，可保证能量单调下降。2. **不动点迭代**：$\mathbf{X}^{(t+1)} = T(\mathbf{X}^{(t)})$。这并不自动意味着能量下降；只有当 T 是压缩映射时，才能保证全局收敛。

定理 D.12（不动点压缩条件）。映射 T 是全局利普希茨连续的，且满足

$$\text{Lip}(T) \le \rho := \frac{\beta}{2}\|\mathbf{M}\|^2 + 2\lambda\|\mathbf{L}\|.$$

如果 $\rho < 1$，那么 T 是压缩映射：它有唯一的不动点$\mathbf{X}^\star$，且$\mathbf{X}^{(t+1)} = T(\mathbf{X}^{(t)})$ 从任意初始化出发都会几何收敛到$\mathbf{X}^\star$。

证明。设 $T(\mathbf{X}) = S(\mathbf{X}) - 2\lambda\mathbf{L}\mathbf{X}$。根据命题 D.6 的证明，$S$是 $\frac{\beta}{2}\|\mathbf{M}\|^2$-利普希茨的，且 $\|\mathbf{L}(\mathbf{X}-\mathbf{Y})\|_F \le \|\mathbf{L}\|\,\|\mathbf{X}-\mathbf{Y}\|_F$。因此

$$\|T(\mathbf{X}) - T(\mathbf{Y})\|_F \le \left(\frac{\beta}{2}\|\mathbf{M}\|^2 + 2\lambda\|\mathbf{L}\|\right)\|\mathbf{X}-\mathbf{Y}\|_F.$$

Hence $\text{Lip}(T) \le \rho$。若 $\rho < 1$，巴拿赫不动点定理保证$\mathbf{X}$的存在性和唯一性 * and

$$\|\mathbf{X}^{(t)} - \mathbf{X}^\star\|_F \le \rho^t\|\mathbf{X}^{(0)} - \mathbf{X}^\star\|_F.$$

□

推论 D.13（阻尼映射）。对于 $T_\alpha(\mathbf{X}) = (1-\alpha)\mathbf{X} + \alpha T(\mathbf{X})$ 满足 $\alpha \in (0,1]$,

$$\text{Lip}(T_\alpha) \le (1-\alpha) + \alpha\rho.$$

因此，如果 $\rho < 1$，每个 $\alpha \in (0,1]$ 都是一个压缩映射。

证明。 对于任意 $\mathbf{X}$，$\mathbf{Y}$，

$$\|T_\alpha(\mathbf{X}) - T_\alpha(\mathbf{Y})\|_F \le (1-\alpha)\|\mathbf{X}-\mathbf{Y}\|_F + \alpha\|T(\mathbf{X}) - T(\mathbf{Y})\|_F,$$

然后应用定理 D.12。 □

Cor推论 D.14 (归一化拉普拉斯矩阵的便捷收缩条件)。 由于 $\|\mathbf{L}\| \le 2$ *fo r nor*归一化$\mathbf{L}$，无阻尼不动点收缩的一个充分条件是

$$\frac{\beta}{2}\|\mathbf{M}\|^2 + 4\lambda < 1.$$

证明。 由定理 D.12 和 $\|\mathbf{L}\| \le 2$ 可直接得到。 □

D.6 临界点与初始化

命题 D.15 (临界点)。 $\mathbf{X}^\star$ 是一阶临界点，当且仅当 $\nabla E_{\text{base}}(\mathbf{X}^\star) = 0$。等价的定点形式：

$$\mathbf{X}^\star = T(\mathbf{X}^\star).$$

若 $\nabla^2 E_{\text{base}}(\mathbf{X}^\star) \succ 0$，则 $\mathbf{X}^\star$ 是严格局部极小点；若 $\nabla^2 E_{\text{base}}(\mathbf{X}^\star)$ 有负特征值，则 $\mathbf{X}^\star$ 是严格鞍点。

证明。 一阶条件：根据定义，$\mathbf{X}^\star$ 是临界点，当且仅当 $\nabla E_{\text{base}}(\mathbf{X}^\star) = 0$。根据命题 D.2，这等价于 $\mathbf{X}^\star = T(\mathbf{X}^\star)$。
对于二阶分类，在$\mathbf{X}^\star$附近作泰勒展开：

$$E(\mathbf{X}^\star + \Delta) = E(\mathbf{X}^\star) + \langle \nabla E(\mathbf{X}^\star), \Delta\rangle + \frac{1}{2}\langle \Delta, \nabla^2 E(\mathbf{X}^\star)\Delta\rangle + o(\|\Delta\|_F^2).$$

在临界点处，线性项为零。若 $\nabla^2 E(\mathbf{X}^\star) \succ 0$，则对于所有充分小的非零 Δ，二次项均为正，因此 $\mathbf{X}^\star$ 是严格局部极小点。若 $\nabla^2 E(\mathbf{X}^\star)$ 有负特征值，则取沿对应特征向量方向的 $\Delta = \epsilon\mathbf{v}$，并令 $\epsilon > 0$ 足够小，使二次项为负，从而得到严格鞍点。 □

命题 D.16 （初始化无关）。 若 $\beta\|\mathbf{M}\|^2 < 2$，则梯度下降从任意初始化出发都会收敛到唯一的全局极小化者。若满足定理 *D.12* 中的 $\rho < 1$，则不动点迭代从任意初始化出发都会收敛到唯一的不动点。

证明。 第一项结论是强凸区域中的定理 D.10。第二项结论由巴拿赫不动点定理得出。

□

备注 D.17（初始化可能产生影响）。 在强凸区域／压缩区域之外，E_{base} 可能是非凸的，并存在多个驻点。此时，吸引域几何和算法选择（步长、阻尼、停止条件）可能影响最终到达哪个驻点。