arXiv:2301.02151v1 [cs.LG] 5 Jan 2023

超越谱隙（扩展版）：拓扑在去中心化学习中的作用

Thijs Vogels*
thijs.vogels@epfl.ch
Hadrien Hendrikx*
hadrien.hendrikx@epfl.ch
Martin Jaggi
martin.jaggi@epfl.ch
洛桑联邦理工学院机器学习与优化实验室，瑞士洛桑

摘要

在机器学习模型的数据并行优化中，工作节点通过协作来改进对模型的估计：更准确的梯度让它们能够使用更大的学习率，从而更快地完成优化。在去中心化设置中，工作节点通过稀疏图进行通信，现有理论未能反映现实行为的重要特征。首先，通信图的“谱隙”无法预测其在（深度）学习中的经验性能。其次，现有理论没有解释协作为何能让工作节点使用比单独训练时更大的学习率。事实上，现有理论规定使用更小的学习率，而且随着图规模增大，学习率还会进一步降低，因此无法解释无限图中的收敛动力学。本文旨在准确描绘稀疏连接的分布式优化。我们量化了图拓扑如何影响二次玩具问题中的收敛，并为一般的光滑凸目标函数及（强）凸目标函数提供理论结果。我们的理论与深度学习中的经验观察相符，并准确描述了不同图拓扑的相对优劣。本文是 Vogels 等人会议论文（2022）的扩展版本。代码：github.com/epfml/topology-in-decentralized-learning。

关键词：去中心化学习，凸优化，随机梯度下降，Gossip 算法，谱隙

1. 引言

分布式数据并行优化算法帮助我们应对机器学习模型日益增长的复杂性，以及用于训练这些模型的数据日益增长的复杂性。我们可以将这些训练算法分为集中式 或去中心化两类，并且通常认为，与“单独”训练相比，这两种设置各有不同的优势。在集中式 设置下，工作节点在独立的小批量数据上计算梯度，并在所有工作节点之间对这些梯度求平均。由此带来的更新方差降低，使得采用更大的学习率并加快训练成为可能。在去中心化 设置下，工作节点只与图中一组稀疏的“邻居”交换并平均各自的模型，而非彼此全部相连；它们还可能拥有从不同分布中采样得到的私有数据集。谈到去中心化学习的优势，我们通常只关注其对其他工作节点数据集的（间接）访问，而不关注训练速度的提升。

同质（独立同分布）设定。 尽管去中心化学习通常研究工作节点间数据集存在异质性的情形，但当工作节点的数据服从相同分布（独立同分布）时，它们之间的稀疏（去中心化）平均同样有用（Lu 和 Sa，2021年）。例如，

©2022 Thijs Vogels、Hadrien Hendrikx 和 Martin Jaggi。*: 贡献相同。预印本。审稿中。许可协议：知识共享署名 4.0，参见 https://creativecommons。org/licenses/by/4。0/。

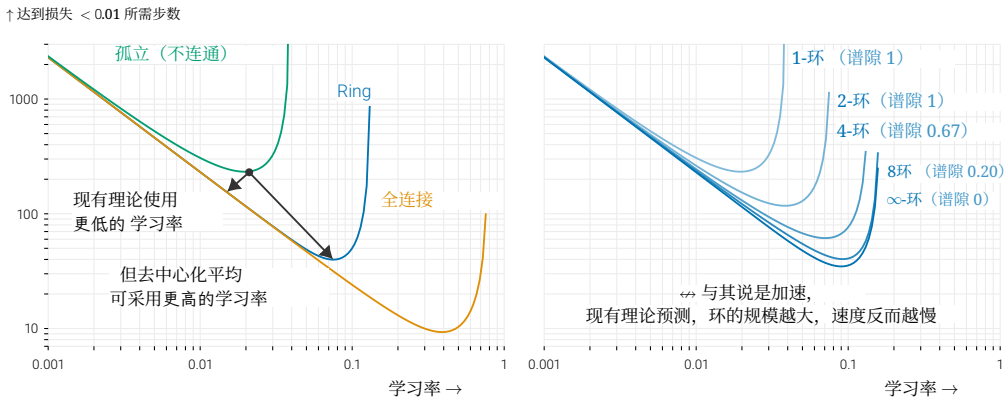


图 1：在独立同分布各向同性二次数据集（Lian等人，2017年）上采用恒定学习率时，去中心化随机梯度下降（第3.1节）达到目标所需的时间。噪声在最优点处消失。与单独进行优化相比，环中的 32 个工作节点（左图）在任何学习率下都更快，但最大的提升来自能够使用较大的学习率。现有理论未能捕捉到这一优势；相反，它建议使用比单独训练更小的学习率。在右图中可以看到，随着环的规模增大，能够使用更大的学习率，优化速度也更快。由于环的谱隙会随环规模增大而趋于零，这无法用现有理论解释。

数据中心使用稀疏平均来缓解通信瓶颈（Assran等人，2019年）。当数据独立同分布（或异质性较弱）时，稀疏平均的目标是更快地优化，就像在集中式（全连接）图中一样。然而，现有去中心化学习理论无法很好地解释这种加速。分析通常表明，对于 足够小的 学习率，采用稀疏平均进行训练的效果与采用全连接平均相同（Lian等人，2017年；Koloskova等人，2020年），因此，与单独训练相比，它能将梯度方差降低工作节点数倍，前提是使用相同的小学习率。然而在实践中，这种小学习率根本不会被采用。事实上，方差降低理应让我们能够使用比单独训练时更大的学习率，而不是被迫使用更小的学习率。与现有理论相反，我们证明，（稀疏）平均在训练的所有阶段（包括初始阶段和渐近阶段）都会降低方差，使我们能够采用更高的学习率，从而直接加快收敛。我们刻画了采用各种通信图进行平均时方差降低的幅度，并表明使用最优大学习率时，并不总能达到集中式性能（方差除以工作节点数）。我们在图 1中展示了所解释的这一现象。

异质（非独立同分布）设置。 在标准分析中，异质性以一种极端悲观的方式影响收敛。直观而言，标准保证对应于这样一种悲观情形：距离最远的工作节点，其函数差异最大。当工作节点拥有从同一分布中独立同分布抽取的不同有限数据集，或者每个工作节点的近邻具有很高的多样性时，这些保证通常较为宽松。在这项工作中，我们刻画了异质性与通信图共同产生的影响，

从而即使对于异质性模式非对抗性的无限图，也能给出非平凡的保证。

谱隙。 在同质和异质设置中，图拓扑都会通过其平均（Gossip）矩阵的谱隙 影响现有的收敛率。谱隙给出了一个保守的下界，衡量每一步平均能让所有工作节点的模型彼此接近多少。谱隙越大越好。如果谱隙很小，就需要使用显著更小的学习率，才能使算法的表现接近采用相同学习率进行全连接平均的随机梯度下降。不幸的是，我们通过实验观察到，在深度学习和凸优化中，经过调优学习率后，通信图的谱隙并不能预测 其性能。

谱隙这一量的问题，在一个简单的例子中得到了清楚的体现。设通信图为一个大小可变的环。随着环的规模趋于无穷，其谱隙趋于零，因为所有工作节点之间越来越难以达成共识。由此，现有理论预测的优化进展也会趋于零。在某些情况下，如果工作节点的目标函数具有对抗性异质性，且需要工作节点从所有其他节点获取信息，那么确实会出现这种情况。然而，在典型情况下，这种观点过于悲观。具体而言，这种观点与独立同分布数据下的实证表现并不相符。对于独立同分布数据，随着环的规模增大，收敛率实际上会提高(图 1)，直到达到一个取决于问题的饱和点。

在这项工作中，我们旨在从理论和实践两方面准确描述采用稀疏平均的分布式学习算法的行为。我们不仅研究此前在本文会议版本中探讨过的高学习率情形Vogels 等人(2022)，也研究小学习率情形，并刻画拓扑结构与数据异质性以及随机噪声之间的相互作用。

- 我们在一个旨在模拟深度学习初始阶段的二次玩具问题中，量化图的作用（第3.1节），并表明平均能够支持更大的学习率。

- 基于这些洞见，我们提出图中一种与问题无关的“有效邻居数”概念。该概念适用于时变拓扑结构和无限图，并能预测图在凸优化和深度学习中的经验性能。

- 我们为（强）凸目标函数给出不依赖于图的谱隙的收敛性证明（第4节），并转而考察更精细的谱量。我们的收敛速率将同质与异质设置区分开来，并指出，当迭代点远离最优点时，所有问题的表现都如同处于同质设置中。

从本质上说，我们的分析并不要求全局共识，而只要求图中彼此相近的工作节点之间达成共识。我们的理论表明，稀疏平均能够以严格的理论保证支持更大的学习率，从而加快优化。这些洞见对于深度学习也很有意义：我们能够准确描述各种拓扑结构的性能，而谱隙无法做到这一点（第5节）。

2. 相关工作

去中心化随机梯度下降。 本文研究去中心化随机梯度下降。 Koloskova等人 (2020年) 在工作节点优化异质目标函数的一般设定下，为该算法给出了最紧的界。他们表明，八卦平均能够降低算法所受的渐近方差，但代价是初始线性收敛项会退化（退化程度取决于八卦矩阵的谱隙）。这一关键项不会因协作而改善，因此导致学习率更小，低于单独训练时的学习率。此外，正如上文所述，这意味着即使不存在异质性，在图规模趋于无穷大时也无法进行优化：例如，无限环的谱隙为零，因此学习率也将为零。

这些速率表明，去中心化平均能够加快训练的最后阶段（该阶段主要受方差影响），代价是减慢初始阶段（线性收敛阶段）。除Koloskova等人（2020年）的工作外，许多论文都致力于研究相较于单独优化的线性加速（方差阶段），并在各种设定下证明了类似结果（Lian等人，2017年；Tang等人，2018年；Lian等人，2018年）。所有这些结果都基于以下认识：线性加速只有在学习率较小时才能实现，而随机梯度下降最终无论如何都需要采用这样小的学习率（例如，由于随机噪声或非光滑性）。这一观察促使这些研究者认为“拓扑无关紧要”。确实如此，但这仅适用于非常小的学习率，如图 1所示。此外，尽管非常小的学习率确实可能实现线性加速，但无论学习率多大，平均都应能带来一定程度的方差缩减。实践中，平均既能加快训练的初始阶段，也能加快最后阶段，而且加速方式可能是非线性的。这正是我们在这项工作中从理论和实践两方面展示的内容。

另一类研究在本地数据满足统计假设的条件下，研究去中心化随机梯度下降。具体而言，Richards 和 Rebeschini(2020年) 展示了去中心化随机梯度下降在依赖于图的隐式正则化下的良好性质，并达到最优统计速率。他们提出的学习率确实取决于通信网络的谱隙，并且会随着谱隙缩小而趋于零。Richards 和 Rebeschini (2019年) 还表明，在去中心化梯度下降中可以使用更大的（常数）学习率，但他们的分析聚焦于去中心化核回归。他们的分析依赖于局部目标函数的统计集中，而本文的分析则依赖于局部邻域这一概念。

无限图上的八卦传播。我们的结果有一个重要特点：它们不依赖谱隙，因此适用于任何规模的图。相反，我们的结果依赖于一些新的量，这些量综合考虑了图拓扑和异质性模式。极端情况下，这些量可能取决于谱隙，但一般而言表现要好得多。Berthier 等人(2020年) 研究了无限图上八卦平均的加速，并得出了与我们相同的结论：尽管谱隙有助于描述渐近行为（信息在整个图中传播需要多长时间），却无法准确描述八卦平均的暂态阶段，即在最初几个八卦轮次中，信息在局部邻域内传播得有多快。这对优化尤其构成限制（相较于单纯的平均），因为每一步都需要对新的局部更新进行平均。对最新梯度更新的平均始终从暂态阶段开始，这意味着八卦平均的暂态阶段会深刻影响去中心化随机梯度下降的渐近行为。在这项工作中，我们借助Berthier 等人(2020年的工具来证明

我们引入的一个关键量——有效邻居数——与图的谱维数之间的关系。

图拓扑的影响。Lian等人(2017年)认为，图的拓扑并不重要。但这只适用于特定设置下的渐近速率，如图 1所示。Neglia等人(2020年)研究了图对去中心化优化的影响，并对这一观点提出了反驳。与我们类似，他们表明，图在训练的早期阶段会产生影响。在异质设置下，他们的分析取决于梯度异质性在拉普拉斯矩阵特征子空间中的分布。然而，他们的假设与我们的不同，而且结果仍不可避免地依赖于图的谱隙。我们的结果性质不同，表明了平均的益处，并通过选择较大的学习率揭示图的影响；对于给定的学习率，我们的结果对噪声和异质性的依赖性更优。Even等人(2021年)也考察了图对去中心化学习的影响。他们关注对异构延迟的非最坏情形依赖性，并且仍然得到了类似谱隙的量，不过这些量是针对重加权 gossip 矩阵的。

另一类研究探讨拓扑与特定数据异质性模式之间的相互作用（Le Bars等人，2022；Dandi等人，2022），以及如何在考虑这种异质性的前提下优化图。我们的分析通过不同于这些研究的一个量凸显了异质性的作用，我们认为该量是紧的。此外，这两项研究要么试图在整个轨迹中降低这种异质性，要么同时优化图的谱隙和异质性项。相反，我们表明，异质性会改变算法的不动点，但不会改变全局动态。

时变拓扑。 由于具有很强的混合能力，时变拓扑在数据中心的去中心化深度学习中广受欢迎（Assran等人，2019年；Wang等人，2019年）。通过标准理论难以解释随时间变化通信拓扑所带来的益处，因此需要专门的分析（Ying等人，2021年）。虽然我们的证明仅适用于静态拓扑，但分析中出现的量也可以用于时变方案。借助这些量，我们可以在同一框架下对静态和时变方案进行实证研究。

会议版本。 本文是对 Vogels 等人（2022）的扩展。该研究聚焦于所有工作节点共享相同全局最优点的齐次设定。在此扩展中，我们提出了一种更简单的分析方法，它严格改进并推广了先前的方法，并将结果扩展到重要的异质设置。在会议版本中，较大的学习率是否只能得益于同质性，这一点仍不清楚。我们还将引入的量与图的谱维数联系起来，并利用这种联系，根据谱维数推导出最优学习率的显式公式。这使我们能够与先前的界限（例如Koloskova等人（2020年）的结果进行准确比较，并表明我们在所有设置下都优于它们。

3. 衡量去中心化学习中的协作

本文对一般凸目标函数的去中心化随机梯度下降的分析，以及深度学习实验，都围绕一个我们将在第3.2节介绍的“有效邻居数”概念展开。本节旨在基于

一个简单的玩具模型来阐释这一量，该模型的收敛情况可以被精确刻画（第3.1节）。随后，我们将在第3.3节中把这一量与谱隙、谱维数等典型图指标联系起来。

3.1 一个玩具问题：在各向同性随机二次函数上进行去中心化随机梯度下降

本节旨在帮助读者建立直观认识，同时避免一般性分析的复杂性。为使本节内容简洁，我们省略所有推导。（Vogels 等人，2022）的附录给出了本节的扩展版本，其中包含推导和证明。

我们考虑 n 个工作节点，它们共同优化一个各向同性二次函数 $\mathbb{E}_{\mathbf{d}\sim\mathcal{N}^d(0,\,1)}\frac{1}{2}(\mathbf{d}^\top\mathbf{x})^2=\frac{1}{2}\|\mathbf{x}\|^2$，其全局最小值唯一，为 $\mathbf{x}^\star=\mathbf{0}$。工作节点通过以下形式的随机梯度访问该二次函数：$\mathbf{g}(\mathbf{x})=\mathbf{d}\mathbf{d}^\top\mathbf{x}$，其中 $\mathbf{d}\sim\mathcal{N}^d(0,\,1)$。这对应于一个拥有无限数据的线性模型，且模型能够完美拟合数据，因此接近最优点时，随机噪声会趋于零。我们的实证结果表明，这个简单模型能够有效近似（过度参数化）深度学习的初始阶段（第5节）。该模型的一个优点是，我们可以计算出精确的收敛速率。这些速率展示了我们在第4节的理论中更一般地刻画的行为。

这个玩具问题中的随机性可以用噪声水平量化

$$\zeta=\sup_{\mathbf{x}\in\mathbb{R}^d}\frac{\mathbb{E}_{\mathbf{d}}\|\mathbf{g}(\mathbf{x})\|^2}{\|\mathbf{x}\|^2}=\sup_{\mathbf{x}\in\mathbb{R}^d}\frac{\mathbb{E}_{\mathbf{d}}\|\mathbf{d}\mathbf{d}^\top\mathbf{x}\|^2}{\|\mathbf{x}\|^2},\tag{1}$$

由于$\mathbf{d}$服从随机正态分布，该噪声水平等于 $\zeta=d+2$。

工作节点运行去中心化随机梯度下降算法（Lian 等人，2017）。每个工作节点 i 都有自己的模型副本 $\mathbf{x}_i\in\mathbb{R}^d$，并交替进行本地模型更新 $\mathbf{x}_i\leftarrow\mathbf{x}_i-\eta\mathbf{g}(\mathbf{x}_i)$，以及与其他工作节点平均其模型：$\mathbf{x}_i\leftarrow\sum_{j=1}^n w_{ij}\mathbf{x}_j$。平均权重 w_{ij} 由八卦矩阵 $\mathbf{W}\in\mathbb{R}^{n\times n}$汇总。非零权重 w_{ij} 表示 i 和j 直接相连。下文中，我们假设 $\mathbf{W}$ 是对称且双随机的：$\sum_{j=1}^n w_{ij}=1\;\forall i$。

对于我们的目标函数，去中心化随机梯度下降要么收敛，要么线性发散。只要它收敛，即学习率足够小时，就存在一个收敛率 r，使得

$$\mathbb{E}\|\mathbf{x}_i^{(t)}\|^2\le(1-r)\|\mathbf{x}_i^{(t-1)}\|^2,$$

等式成立时为 $t\to\infty$。当工作节点单独训练（$\mathbf{W}=\mathbf{I}$）时，给定学习率 η 的收敛率为：

$$r_{\text{alone}}=1-(1-\eta)^2-(\zeta-1)\eta^2.\tag{2}$$

最优学习率 $\eta^\star=\frac{1}{\zeta}$ 平衡了优化项 $(1-\eta)^2$ 和随机项 $(\zeta-1)\eta^2$。在集中式（全连接）设置下（$w_{ij}=\frac{1}{n}\;\forall i,j$，收敛率同样简单：

$$r_{\text{centralized}}=1-(1-\eta)^2-\frac{(\zeta-1)\eta^2}{n}.\tag{3}$$

工作节点之间的平均可减轻梯度噪声的影响，最优学习率增至 $\eta^\star=\frac{n}{n+\zeta-1}$。我们发现，采用一般八卦矩阵 $\mathbf{W}$ 的去中心化随机梯度下降可在这些结果之间进行插值。

3.2 有效邻居数

为了一般性地量化 $(\zeta-1)\eta^2$ 项的减少，我们引入与问题无关的概念：有效邻居数 $n_{\mathbf{W}}(\gamma)$，八卦矩阵 $\mathbf{W}$ 的衰减参数 γ。

定义 1（有效邻居数）有效邻居数 $n_{\mathbf{W}}(\gamma) = \lim_{t\to\infty} \frac{\sum_{i=1}^{n}\mathrm{Var}[\mathbf{y}_i^{(t)}]}{\sum_{i=1}^{n}\mathrm{Var}[\mathbf{z}_i^{(t)}]}$ 衡量这些过程的渐近方差之比

$$\mathbf{y}^{(t+1)} = \sqrt{\gamma}\cdot\mathbf{y}^{(t)} + \boldsymbol{\xi}^{(t)}, \quad where\ \mathbf{y}^{(t)} \in \mathbb{R}^n\ and\ \boldsymbol{\xi}^{(t)} \sim \mathcal{N}^n(0,1) \tag{4}$$

and

$$\mathbf{z}^{(t+1)} = \mathbf{W}(\sqrt{\gamma}\cdot\mathbf{z}^{(t)} + \boldsymbol{\xi}^{(t)}), \quad where\ \mathbf{z}^{(t)} \in \mathbb{R}^n\ and\ \boldsymbol{\xi}^{(t)} \sim \mathcal{N}^n(0,1). \tag{5}$$

我们将$\mathbf{y}$ 和$\mathbf{z}$ 称为随机游走，因为工作节点会反复向其状态中添加噪声，这有点像随机梯度下降的参数更新。不要将其与图中节点上的“随机游走”混淆。

由于通过 $\mathbf{W}$ 进行平均，最多只能使随机游走的方差降低 n，因此有效邻居数介于 1 和 n之间。衰减 γ 会调节对通信延迟的敏感度。如果 $\gamma=0$，工作节点只能从与直接邻居进行平均中获益。随着 γ 增大，多跳连接会发挥越来越重要的作用。当 γ 趋近于 **1** 时，延迟和无延迟的噪声贡献权重相等，并且对于任何连通拓扑，降幅趋于 n 。

命题2 对于正则双随机对称闲聊矩阵$\mathbf{W}$ 其特征值为$\lambda_1,\ldots,\lambda_n$, $n_{\mathbf{W}}(\gamma)$ 具有闭式表达式

$$n_{\mathbf{W}}(\gamma) = \frac{\frac{1}{1-\gamma}}{\frac{1}{n}\sum_{i=1}^{n}\frac{{\lambda_i}^2}{1-\lambda_i^2\gamma}}. \tag{6}$$

这可通过展开 $\mathbf{y}$ 和 $\mathbf{z}$的递推关系，利用 $\mathbf{W}$的特征分解，以及极限 $\lim\, t\to\infty \sum_{k=1}^{t} x^k = \frac{x}{1-x}$ 得出。

虽然这个闭式表达式仅适用于一类有限的闲聊矩阵，但随机游走中的方差缩减概念自然也可推广到无限拓扑结构或时变平均方案。 图2 展示了 $n_{\mathbf{W}}$ 在各种拓扑结构下的情况。

在对各向同性二次玩具问题中去中心化随机梯度下降的收敛进行精确刻画时，我们发现，在 n 的全连接收敛率中，有效邻居数取代了工作节点数 方程3。收敛率 r 是以下方程的唯一解

$$r = 1-(1-\eta)^2 - \frac{(\zeta-1)\eta^2}{n_{\mathbf{W}}\left(\frac{(1-\eta)^2}{1-r}\right)}. \tag{7}$$

对于全连接和不连通的 $\mathbf{W}$， $n_{\mathbf{W}}(\gamma)=n$ 分别为 **1**，无论 γ如何，方程 7都能还原方程 2 和 3。对于其他图，有效工作节点数取决于学习率。现有理论只考虑 $n_{\mathbf{W}}\approx n$的情况，但是

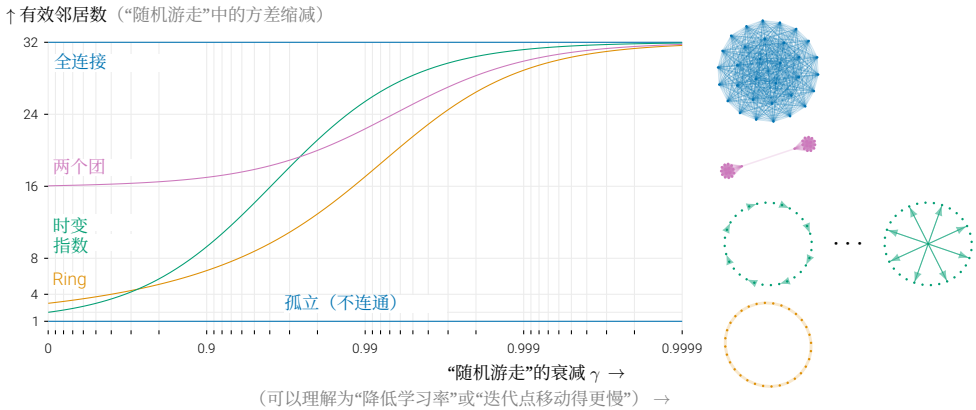


图2：通过方差缩减衡量的若干拓扑结构的有效邻居数（见5）。相关的 γ 轴上的 x点取决于学习率和任务。哪种拓扑结构“最佳”因问题而异。对于较大的衰减率 γ （对应较小的学习率），所有连通拓扑结构实现的方差缩减都接近全连接图。对于较小的衰减率（较大的学习率），工作节点只能从其直接邻居中受益（例如，环形结构中的3个邻居）。对于恒定拓扑结构，这些曲线可以显式计算；对于时变指数方案，则可以高效模拟（Assran等人，2019年）。

这要求学习率足够小，而这可能会使该项 $(1-\eta)^2$ 过大，从而违背协作的初衷。

除了这个简单示例，我们还发现，所提出的有效邻居数概念对于分析一般目标函数（第4节）和深度学习（第5节）同样有意义。

3.3 有效邻居数与其他图数量之间的联系

一般而言，有效邻居数函数 $n_{\mathbf{W}}(\gamma)$ 无法用单个标量概括。图2表明，该函数的行为因图而异。不过，我们可以利用已知的图数量（例如其谱隙或谱维数）来界定有效邻居数。

我们的目标是为有限图和无限图都建立界限。为此，我们将命题2 推广为对八卦矩阵的谱测度$\mathbf{d}\sigma$ 进行积分，而不是对其特征值求和：

$$n_{\mathbf{W}}(\gamma)^{-1} = (1-\gamma)\int_0^1 \frac{\lambda^2}{1-\gamma\lambda^2}\mathbf{d}\sigma(\lambda). \tag{8}$$

对于有限图，$\mathbf{d}\sigma$ 是质量为 $\frac{1}{n}$ 的狄拉克 delta 函数之和，$_$分别位于矩阵$\mathbf{W}$的各个特征值处，从而恢复方程（6）。

3.3.1 上界和下界

我们可以利用所有特征值 λ 均为 ≤ 1 这一事实，得到：

$$n_{\mathbf{W}}(\gamma)^{-1} \leq (1-\gamma)\int_0^1 \frac{1}{1-\gamma}\mathbf{d}\sigma(\lambda) = 1, \tag{9}$$

这个下界 对应于非连通图的“有效邻居数”。

另一方面，对于有限图，我们可以利用这样一个事实： $\sigma(\lambda)$包含一系列 n狄拉克点质量。对应于完全平均状态的 $\lambda = 1$处峰值为 $\frac{1}{n}$，而其他峰值为 ≥ 0。利用这个界，我们得到

$$n_{\mathbf{W}}(\gamma)^{-1} \geq \frac{1-\gamma}{1-\gamma}\frac{1}{n} = \frac{1}{n}. \tag{10}$$

这个上界 对于全连通图，有效邻居数的界是紧的。

3.3.2 根据谱隙给出界

如果图存在谱隙 α，这意味着 $\sigma(\lambda)$ 在 $\lambda = 1$ 处包含一个质量为 $\frac{1}{n}$ 的狄拉克 delta 函数，对应于完全平均状态。 $\sigma(\lambda)$ 的其余部分质量为 $\frac{n-1}{n}$，且包含在子域 $\lambda \in [0, 1-\alpha]$ 中。在这种情况下，我们得到

$$n_{\mathbf{W}}(\gamma)^{-1} \leq \frac{1}{n} + \frac{n-1}{n}\frac{(1-\gamma)(1-\alpha)^2}{1-\gamma(1-\alpha)^2}. \tag{11}$$

这个有效邻居数的下界通常较为保守，但对于有限次八卦矩阵 $\mathbf{W} = (1-\alpha)\mathbf{I} + \frac{\alpha}{n}\mathbb{1}\mathbb{1}^\top$，该下界是紧的。

3.3.3 根据谱维数给出界

接下来，我们将把有效邻居数的概念与图的谱维数 d_s联系起来（Berthier，2021年，例如定义 1.9），谱维数控制着特征值在 1 附近的衰减。这个概念通常与图的拉普拉斯算子的谱测度相关。不过，为避免引入过多与图相关的量，我们定义相对于八卦矩阵 $\mathbf{W}$ 的谱维数。采用拉普拉斯矩阵$\mathbf{L}_{\mathbf{W}} = \mathbf{I} - \mathbf{W}$ 的标准定义与此等价。在本文余下部分中，“图”始终指由 $\mathbf{W}$ （其拉普拉斯矩阵为 $\mathbf{L}_{\mathbf{W}}$）隐式诱导的通信图。

定义 3（谱维数）如果存在 $c_s > 0$，使得对于所有 $\lambda \in [0, 1]$，其特征值密度满足以下界，则该八卦矩阵的谱维数至少为 d_s

$$\sigma((\lambda, 1)) \leq c_s^{-1}(1-\lambda)^{\frac{d_s}{2}}. \tag{12}$$

此处的记号 $\sigma((\lambda, 1))$ 表示积分 $\int_\lambda^1 \sigma(l)\mathbf{d}l$。图的谱维数具有自然的几何解释。例如，线（或环）的谱维数为 $d_s = 1$，而二维网格的谱维数为 2。更一般地， d 维环面的谱维数为 d。此外，谱维数描述宏观拓扑特征，且对微观变化具有鲁棒性。例如，随机几何图的谱维数为 2。

注意，由于有限图具有谱隙， $\sigma((\lambda_2(\mathbf{W}),1)) = 0$ 因此有限图对任意谱维数 d_s都满足 (12)。然而，谱维数这一概念对有限图仍然有意义，因为当 c_s 大于拓扑相似的无限图的实际谱维数时，常数 d_s 会发散。另一种做法是在 (12) 中明确考虑谱隙，例如 Berthier 等人 (2020年，第 6 节)。

现在我们继续利用谱维数来给 $n_{\mathbf{W}}(\gamma)$ 设定上界。由于 $\lambda \mapsto \lambda^2(1-\gamma\lambda^2)^{-1}$ 是 $[0,1]$, 上的非负非递减函数，我们可以使用 Berthier 等人 (2020年，引理 C.1）得到:

$$n_{\mathbf{W}}(\gamma)^{-1} \leq \frac{1}{n} + c_s^{-1}(1-\gamma)\int_0^1 \frac{\lambda^2}{1-\gamma\lambda^2}(1-\lambda)^{\frac{d_s}{2}-1}\mathbf{d}\lambda. \tag{13}$$

项 $\frac{1}{n}$ 源于这样一个事实：对于有限图，密度 $\mathbf{d}\sigma$ 包含一个位于特征值 1 处、质量为 $\frac{1}{n}$ 的狄拉克 delta-函数。谱维数不会影响这个狄拉克 delta 函数；为保证一致性，它也是必需的，因为它确保对任意有限图都有 $n_{\mathbf{W}}(\gamma) \leq n$ 。接下来，为了计算积分，我们区分三种情况。

情况 $d_s > 2$。 由于 $\gamma\lambda < 1$，因此 $1-\lambda \leq 1-\gamma\lambda^2$。特别地，我们使用分部积分得到:

$$\begin{aligned} n_{\mathbf{W}}(\gamma)^{-1} - n^{-1} &\leq c_s^{-1}(1-\gamma)\int_0^1 \lambda^2(1-\gamma\lambda^2)^{\frac{d_s}{2}-2}\mathbf{d}\lambda \\ &\leq -\frac{(1-\gamma)c_s^{-1}}{2\gamma(d_s/2-1)}\int_0^1 -2\gamma\lambda(d_s/2-1)(1-\gamma\lambda^2)^{\frac{d_s}{2}-2}\mathbf{d}\lambda \\ &= \frac{(1-\gamma)c_s^{-1}}{\gamma(d_s-2)}\left[1-(1-\gamma)^{\frac{d_s}{2}-1}\right]. \end{aligned}$$

这使得其按以下方式缩放:

$$n_{\mathbf{W}}(\gamma) \geq \left(\frac{1}{n} + \frac{(1-\gamma)}{\gamma(d_s-2)c_s}\right)^{-1}. \tag{14}$$

当 n 足够大时，我们得到与上一节相同的 $(1-\gamma)^{-1}$ 缩放，这表明对于连接足够良好的网络（$d_s > 2$），谱维数只影响常数，而不影响 γ 中的缩放。

Case $d_s = 2$. 当 $d_s = 2$ 时，只有被积函数的原函数发生变化，导致 o:

$$n_{\mathbf{W}}(\gamma) \geq \left(\frac{1}{n} - \frac{(1-\gamma)\ln(1-\gamma)}{2\gamma c_s}\right)^{-1} \tag{15}$$

情形 $d_s < 2$。 在这种情形下，我们首先将积分拆分为:

$$(1-\gamma)\int_0^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda = (1-\gamma)\int_0^\gamma \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda + (1-\gamma)\int_\gamma^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda$$

对于第一项，注意到 $\gamma\lambda \le 1$，因此 $(1-\gamma\lambda^2)^{-1} \le (1-\lambda)^{-1}$，从而得到：

$$(1-\gamma)\int_0^\gamma \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda \le (1-\gamma)\int_0^\gamma (1-\lambda)^{\frac{d_s}{2}-2}\mathbf{d}\lambda$$
$$= \frac{2(1-\gamma)}{2-d_s}\left[(1-\gamma)^{\frac{d_s}{2}-1}-1\right] \le \frac{2}{2-d_s}(1-\gamma)^{\frac{d_s}{2}}.$$

对于第二项，注意到 $\lambda^2 \le 1$，因此 $(1-\gamma\lambda^2)^{-1} \le (1-\gamma)^{-1}$，从而得到：

$$(1-\gamma)\int_\gamma^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda \le \int_\gamma^1 (1-\lambda)^{\frac{d_s}{2}-1}\mathbf{d}\lambda = \frac{2}{d_s}(1-\gamma)^{\frac{d_s}{2}}. \tag{16}$$

最终，我们得到 $n_{\mathbf{W}}(\gamma)^{-1} - \frac{1}{n} \le \frac{2}{c_s}\left[\frac{1}{2-d_s}+\frac{1}{d_s}\right](1-\gamma)^{\frac{d_s}{2}}$，因此：

$$n_{\mathbf{W}}(\gamma) \ge \left(\frac{1}{n}+\frac{4(1-\gamma)^{\frac{d_s}{2}}}{d_s(2-d_s)c_s}\right)^{-1}. \tag{17}$$

在这种情况下，γ 中的缩放会受到谱维数的影响。连接性更强的图从更高的 γ 中获益更多。

4. 收敛性分析

4.1 符号与定义

上一节中，我们针对一个特定函数推导出了精确的收敛率。现在，我们给出一般（强）凸函数的收敛率，这些结果与上一节中的观察一致。我们得到的收敛率取决于噪声水平、目标函数的难度以及图的拓扑。更形式化地说，我们假设要解决以下问题：

$$\min_{\theta\in\mathbb{R}^d}\sum_{i=1}^n f_i(\theta) = \min_{\mathbf{x}\in\mathbb{R}^{nd},\mathbf{x}_i=\mathbf{x}_j}\sum_{i=1}^n f_i(\mathbf{x}_i). \tag{18}$$

在这种情况下，$\mathbf{x}_i \in \mathbb{R}^d$ 表示节点 i的局部变量，而$\mathbf{x} \in \mathbb{R}^{nd}$ 表示所有节点的变量堆叠。我们假设去中心化随机梯度下降采用以下迭代：

$$\text{(D-SGD):} \qquad \mathbf{x}_i^{(t+1)} = \sum_{j=1}^n w_{ij}\mathbf{x}_j^{(t)} - \eta\nabla f_{\xi_i^{(t)}}(\mathbf{x}_i^{(t)}) \tag{19}$$

其中，$f_{\xi_i^{(t)}}$ 表示采样数据点，而对等通信权重w_{ij} 是$\mathbf{W}$的元素。记$\mathbf{L_W} = \mathbf{I}-\mathbf{W}$，我们将该表达式改写为矩阵形式：

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \left[\eta\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)}) + \mathbf{L_W}\mathbf{x}^{(t)}\right], \tag{20}$$

其中 $(\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)}))_i = \nabla f_{\xi_i^{(t)}}(\mathbf{x}_i^{(t)})$。这里对记号稍作滥用：$\mathbf{W} \in \mathbb{R}^{nd\times nd}$ 现在表示标准 $n\times n$八卦矩阵与 $d\times d$单位矩阵的克罗内克积。

此定义与本工作的会议版本（Vogels 等人，2022）略有不同；会议版本随机交替执行八卦步骤和梯度更新，而不是依次执行。在随机化设定下仍然可以进行分析，但当目标函数存在异质性 $\mathbf{x}_i \neq \sum_{j=1}^n w_{ij}\mathbf{x}j$ 时，即使对于去中心化随机梯度下降 (19) 的不动点也是如此；而随机化更新会引入不必要的方差。同样，也可以分析流行的变体 $\mathbf{x}^{(t+1)} = \mathbf{W}[\mathbf{x}^{(t)} - \eta\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)})],$，该变体会先在本地对随机梯度求平均，再应用这些梯度。然而，(19) 中的去中心化随机梯度下降算法允许并行执行通信和计算，并使分析更简单。我们在以下假设下分析此模型，其中 $D_f(x,y) = f(x) - f(y) - \nabla f(y)^\top(x-y)$ 表示 f 两点 x 和 y之间的 Bregman散度。

假设 4 随机梯度满足以下条件：*(*I*)* 采样数据点 $\xi_i^{(t)}$ 和$\xi_j^{(\ell)}$ 在不同时间 t、ℓ 以及不同节点 $i \neq j$之间相互独立。*(*II*)* 随机梯度局部无偏：$\mathbb{E}\left[f_{\xi_i^{(t)}}\right] = f_i$ 对所有 t, i *(*III*)* 对所有 $t, i,$ 目标函数 $f_{\xi_i^{(t)}}$ 均为凸且 ζ_ξ-光滑，并且 $\mathbb{E}\left[\zeta_\xi D_{f_\xi}(x,y)\right] \leq \zeta D_f(x,y)$ 对所有 x, y成立。*(* IV*)* 对于 $\mu \geq 0$，所有局部目标函数 f_i均为 μ-强凸且 L-光滑。

大学习率。 光滑常数 ζ 随机函数的 f_ξ 定义了问题在暂态阶段的噪声水平（越低越好）。比率ζ/L 比较了在达到“方差区域”之前，使用随机梯度与使用真实全局梯度进行优化的难度；去中心化随机梯度下降在恒定学习率下的迭代点几乎必然位于该区域内，$t \to \infty$。因此，在插值情形下该比率尤为重要：此时所有 $f_{\xi_i^{(t)}}$ 具有相同的最小值，因而“方差区域”会缩减为最优点 $\mathbf{x}^\star$。假设全局平均目标函数的平滑性优于局部函数，是证明工作节点之间进行平均能够采用更大学习率的关键。若不进行通信，则当学习率满足 $\eta \leq 1/\zeta$时，可确保收敛至“方差区域”。如果 $\zeta \approx L$，噪声很小，协作只能降低最终方差，并使结果更接近全局最小值（而非仅接近自身的最小值）。然而，在高噪声情形下（$\zeta \gg L$），例如第3.1节中 $\zeta = d + 2 \gg 1 = L$，平均能够支持更大的学习率，最高可达 $\min(1/L, n/\zeta)$，从而大幅加快初始训练阶段。这正是我们将在定理 6中证明的内容。

如果工作节点始终彼此接近（$\mathbf{x}_i \approx \frac{1}{n}(\mathbf{x}_1 + \ldots + \mathbf{x}_n)$ $\forall i$，或者等价地，$\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{x} \approx \mathbf{x}$），那么去中心化随机梯度下降的表现与对平均参数 $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ 进行随机梯度下降相同，而学习率取决于 $\max(\zeta/n, L)$，这表明方差缩减了 n。然而，要维持 “$\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{x} \approx \mathbf{x}$”，就需要较小的学习率。这是分析去中心化随机梯度下降时常见的出发点，尤其是 Koloskova 等人（2020）的证明。反过来，如果我们不假设工作节点彼此接近，则“$\mathbf{I}\mathbf{x} \approx \mathbf{x}$”始终成立。在这种情况下，没有方差缩减，但也不需要较小的学习率。在第3.1节中，我们发现，在最优学习率下，工作节点并非与所有其他工作节点都接近，但与图中距离不太远的其他节点接近。

我们通过定义邻域矩阵$\mathbf{M} \in \mathbb{R}^{n\times n}$来刻画“局部接近性”这一概念。它使我们能够在直接邻居之外进行半局部平均，而无需对整个图进行完全平均。我们确保“$\mathbf{M}\mathbf{x} \approx \mathbf{x}$”，从而使平滑性得到提升，效果介于 ζ （单独实现时）与 ζ/n （保持全局共识时实现）之间

。每个邻域矩阵$\mathbf{M}$ 都意味着对学习率的要求，同时也能提升平滑性。

虽然我们可以使用任意$\mathbf{M}$进行分析，但在学习率要求与平滑性提升之间取得良好平衡的矩阵最为值得关注。基于第3.1节，因此我们聚焦于一种特定的矩阵构造：我们选择$\mathbf{M}$作为图上衰减型γ “随机游走过程”的协方差，如（5）所示，即

$$\mathbf{M} = (1-\gamma)\sum_{k=1}^{\infty}\gamma^{k-1}\mathbf{W}^{2k} = (1-\gamma)\mathbf{W}^2(\mathbf{I}-\gamma\mathbf{W}^2)^{-1}. \tag{21}$$

改变 γ 会产生从$\mathbf{M} = \mathbf{W}^2$ $(\gamma = 0)$到$\mathbf{M} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ $(\gamma = 1)$的一系列平均邻域。 γ 还意味着有效邻居数 $n_{\mathbf{W}}(\gamma)$： γ越大， $n_{\mathbf{W}}(\gamma)$越大。我们对邻域矩阵$\mathbf{M}$作出如下假设：

假设5 邻域矩阵$\mathbf{M}$ 具有（21）的形式，且所有对角元素具有相同的值，即$\mathbf{M}_{ii} = \mathbf{M}jj$ 对于所有 i， j。

假设5意味着$\mathbf{M}_{ii}{}^{-1} = n_{\mathbf{W}}(\gamma)$：式（6）中定义的有效邻居数等于$\mathbf{M}$的自权重的倒数。这是因为$\mathbf{M}$的迹等于其特征值之和。否则，所有需要假设5的结果均可通过将 $n_{\mathbf{W}}(\gamma)$替换为$\min_i \mathbf{M}_{ii}{}^{-1}$而成立。除了有效邻居数 $n_{\mathbf{W}}(\gamma)$这一有趣关系之外，我们还将关注$\mathbf{M}$的另一个谱性质，即常数 $\beta(\gamma)$（它仅通过$\mathbf{M}$依赖于 γ，但我们将这种依赖关系明确表示出来），其满足：

$$\mathbf{L}_{\mathbf{M}} \preccurlyeq \beta(\gamma)^{-1}\mathbf{L}_{\mathbf{W}}\mathbf{W} \tag{22}$$

这个常数可以解释为由$\mathbf{L}_{\mathbf{W}}\mathbf{W}$定义的半范数相对于由$\mathbf{L}_{\mathbf{M}}$定义的半范数的强凸性。由于$\mathbf{M}$的形式，我们有 $1-\lambda_2(\mathbf{W}) \le \beta(\gamma) \le 1$，且对于 $\gamma \to 1$， 下界是紧的。不过， $\mathbf{M}$ 的特定形式（涉及由$\mathbf{W}$定义的邻域）以及 $\gamma < 1$ 的使用，通常会确保一个大得多的常数 $\beta(\gamma)$。

D-(S)GD 的不动点。 在 Vogels 等人 (2022) 中，我们考虑一个齐次设定，其中 $\mathbb{E} f_{\xi_i^{(t)}} = f$ 对所有 i均成立。现在，我们将分析拓展到更一般的情形，考虑局部函数 f_i 可能各不相同的设定。在这种情况下，采用恒定学习率的去中心化梯度下降（去中心化随机梯度下降的确定性版本）不会收敛到平均函数的最小化解，而是收敛到另一个解。现在考虑这个不动点 $\mathbf{x}_\eta^\star$， 它满足：

$$\eta\nabla F(\mathbf{x}_\eta^\star) + \mathbf{L}_{\mathbf{W}}\mathbf{x}_\eta^\star = 0. \tag{23}$$

请注意， $\mathbf{x}_\eta^\star$ 很大程度上取决于学习率 η （我们在符号表示中特别强调了这一点），并且通常不处于共识状态（$\mathbf{L}_{\mathbf{W}}\mathbf{x}_\eta^\star \neq 0$）。在存在随机噪声时，去中心化随机梯度下降会在该不动点 $\mathbf{x}_\eta^\star$附近（范围与梯度方差成正比）振荡，因此从现在起，我们将$\mathbf{x}_\eta^\star$ 称为去中心化随机梯度下降的不动点。

在本节余下部分，我们证明，只要将全局最小值点Vogels 等人（2022）$\mathbf{x}^\star$ （问题（18）的解）替换为这个固定的

点$\mathbf{x}_\eta^\star$。更具体地说，我们通过确保以下李雅普诺夫函数的下降来衡量收敛：

$$\mathcal{L}_t = \|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 + \omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}}^2 = (1-\omega)\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 + \omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|^2, \tag{24}$$

对于某个参数 $\omega \in [0,1]$，且 $\mathbf{L_M} = \mathbf{I} - \mathbf{M}$。接下来，我们将说明这些结果如何推导出收敛到 $\mathbf{x}_\eta^\star$ 的一个邻域，并且该邻域会随着学习率 η 的减小而缩小。具体而言，本节安排如下：

1. 定理 6 首先证明收敛到$\mathbf{x}_\eta^\star$的一般收敛结果，其中 $\mathbf{x}$ 是 D-(S)GD 的不动点。
2. 定理 9 随后界定一般学习率下与真实最优解的距离。
3. 推论 10 最后给出采用优化学习率时的完整收敛结果。想要快速比较我们的结果与当前最先进结果的读者，可以直接跳到这一结果。

4.2 一般收敛结果

定理 6 给出了任意选择参数 γ 来确定邻域矩阵 $\mathbf{M}$以及任意李雅普诺夫参数 ω时的收敛率。特定的 γ 和 ω 能够取得最佳收敛率，它们在平均带来的益处与其对邻居间接近程度施加的限制之间取得平衡。下一节将更深入地讨论这些选择。

定理 6 若假设 *4* 和 *5* 成立，且若 η 满足

$$\eta \leq \min\left(\frac{\beta(\gamma)\omega}{L}, \frac{1}{4\left(\left[n_{\mathbf{W}}(\gamma)^{-1} + \omega\right]\zeta + L\right)}\right), \tag{25}$$

则由 (24)定义的李雅普诺夫函数满足以下结论：

$$\mathcal{L}^{(t+1)} \leq (1-\eta\mu)\mathcal{L}^{(t)} + \eta^2\sigma_{\mathbf{M}}^2,$$

其中 $\sigma_{\mathbf{M}}^2 = 2[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1} + \omega]\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right]$。

该定理表明，无论“真正的”最小化解 $\mathbf{x}^\star$为何，去中心化随机梯度下降都会收敛至不动点 $\mathbf{x}_\eta^\star$（收敛至一个方差区域）。根据具体应用场景，收敛至 $\mathbf{x}_\eta^\star$ 或许并不理想（但请记住，随着 η 缩小，$\mathbf{x}_\eta^\star \to \mathbf{x}^\star$），但这正是去中心化随机梯度下降的行为，因此我们认为，首先明确说明这一点很重要。齐次情形不存在这个问题，因为对于所有 η，只要 η 满足收敛条件，就有 $\mathbf{x}_\eta^\star = \mathbf{x}^\star$。

参数 $\omega \in [0,1]$ 是自由的，通常方便将其选为 $\omega = \eta L/\beta(\gamma)$，以消除对 η 的第一个条件。不过，我们以自由参数 ω 的形式给出该结果，因为正如本节后续内容将会展示的，设定 $\omega = n_{\mathbf{W}}(\gamma)^{-1}$ 可以得到简单的推论。

证明 现在我们详细阐述该证明，它对以下文献中的定理 IV 进行了简化和推广：Vogels 等人 (2022)。

1 - 一般分解 我们首先分析李雅普诺夫函数中的第一项（24），并利用（23）中的不动点条件，写为：

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}_\eta^\star\|_{\mathbf{M}}^2\right] &= \|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{M}}^2+\|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{\mathbf{M}}^2 \\ &\quad -2\eta\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top\mathbf{M}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)-2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}\mathbf{M}}^2.\end{aligned} \tag{26}$$

第二项与将$\mathbf{M}$ 替换为$\mathbf{I}$时相同。

2 - 误差项 我们首先对误差项进行界定，并利用最优性条件得到：

$$\begin{aligned}&\mathbb{E}\left[\|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{\mathbf{M}}^2\right] \\ &=\mathbb{E}\left[\|\eta\left[\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]+\mathbf{L}_{\mathbf{W}}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right] \\ &=\mathbb{E}\left[\|\eta\left(\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\right)+\left[\eta\left(\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\right)+\mathbf{L}_{\mathbf{W}}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\right]\|_{\mathbf{M}}^2\right] \\ &\le 2\eta^2\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right]+2\eta^2\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right]+2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}\mathbf{M}\mathbf{L}_{\mathbf{W}}}^2,\end{aligned}$$

最后一个不等式来自偏差-方差分解。第二项对应方差，而第一项和最后一项将由下降项抵消。

随机梯度噪声。 为了界定第一项，我们关键地利用了不同节点的随机噪声相互独立这一性质，因此特别有：

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right] &= n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|^2\right] \\ &\quad+\|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}-n_{\mathbf{W}}(\gamma)^{-1}\mathbf{I}}^2 \\ &\le 2n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\zeta_{\xi_t}D_{F_{\xi_t}}(\mathbf{x}_\eta^\star,\mathbf{x}^{(t)})\right]+\|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|^2 \\ &\le 2\left[n_{\mathbf{W}}(\gamma)^{-1}\zeta+L\right]D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star),\end{aligned}$$

这里我们用到了 $\mathbf{M}\preccurlyeq\mathbf{I}$、 L-余强制性 F以及噪声假设，即$\mathbb{E}\left[\zeta_{\xi_t}D_{F_{\xi_t}}\right]\le\zeta D_F$。有效邻居数 $n_{\mathbf{W}}(\gamma)$ 在此处起作用，因为假设 5 意味着 $\mathbf{M}$ 的对角线等于 $n_{\mathbf{W}}(\gamma)^{-1}\mathbf{I}$。再次利用独立性，我们得到：

$$\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right]=n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|^2\right] \tag{27}$$

对 $\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|^2\right]$ 项进行相同的计算，并加上共识误差，可得：

$$\begin{aligned}\mathbb{E}\left[\|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{(1-\omega)\mathbf{M}+\omega\mathbf{I}}^2\right] &\le 4\left[\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega\right]\zeta+(1-\omega)L\right]D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star) \\ &\quad+2\eta^2((1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega)\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|^2\right] \\ &\quad+2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}[\mathbf{M}+\omega\mathbf{L}_{\mathbf{M}}]\mathbf{L}_{\mathbf{W}}}^2\end{aligned} \tag{28}$$

在这里，第一项将通过梯度项带来的下降来控制，第二项则通过通信项来控制。

3 - 下降项

梯度项 我们首先分析所有梯度项的影响。具体而言，我们利用 $(1-\omega)\mathbf{M}+\omega\mathbf{I} = \mathbf{I} - (1-\omega)\mathbf{L_M}$。然后，我们利用

$$\left[\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\right]^\top (\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star) = D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) + D_F(\mathbf{x}_\eta^\star, \mathbf{x}^{(t)}),$$

and:

$$\begin{aligned} 2\left[\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\right]^\top \mathbf{L_M}(\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star) &\leq 2\|\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\|\|\mathbf{L_M}(\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star)\| \\ &\leq \frac{1}{2L}\|\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\|^2 + 2L\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2} \\ &\leq D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) + 2L\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2}. \end{aligned}$$

总体而言，梯度项之和为：

$$\begin{aligned} &-2\left[\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\right]^\top (\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star) + 2(1-\omega)\left[\nabla F(\mathbf{x}^{(t)}) - \nabla F(\mathbf{x}_\eta^\star)\right]^\top \mathbf{L_M}(\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star) \\ &\leq -2D_F(\mathbf{x}_\eta^\star, \mathbf{x}^{(t)}) - (1+\omega)D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) + 2(1-\omega)L\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2} \\ &\leq -\mu\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|^2 - D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) + 2L\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2} \\ &\leq -(1-\omega)\mu\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 - \omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|^2 - D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) + 2\beta(\gamma)^{-1}L\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M L_W W}}, \end{aligned} \tag{29}$$

其中我们利用了 $\mathbf{L_M} \preccurlyeq \beta(\gamma)^{-1}\mathbf{L_W W}$。

Gossip terms. 我们在此回顾用于下降分析的 Gossip 项，其表达式为：

$$-2\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_W M}}^2 - 2\omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_W L_M}}^2. \tag{30}$$

4 - 汇总所有内容。现在，我们将所有的下降项和误差项相加。具体而言，利用方程（28）、（29）和（30），我们得到：

$$\begin{aligned} \mathcal{L}^{(t+1)} \leq\ & (1-\eta\mu)\mathcal{L}^{(t)} \\ & - 2\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_W M(I-L_W)}}^2 \\ & - 2\omega\left[1 - \eta L/(\omega\beta(\gamma))\right]\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_W L_M W}}^2 \\ & - \eta\left(1 - 4\eta\left[\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1} + \omega\right]\zeta + (1-\omega)L\right]\right)D_F(\mathbf{x}^{(t)}, \mathbf{x}_\eta^\star) \\ & + 2\eta^2\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1} + \omega\right]\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right]. \end{aligned}$$

定理中的条件经过选择，使第 3 行和第 4 行中的项为正（第 2 行自动满足这一点），并利用 $1-\omega \leq 1$（因为 ω 无论如何都很小）。

■

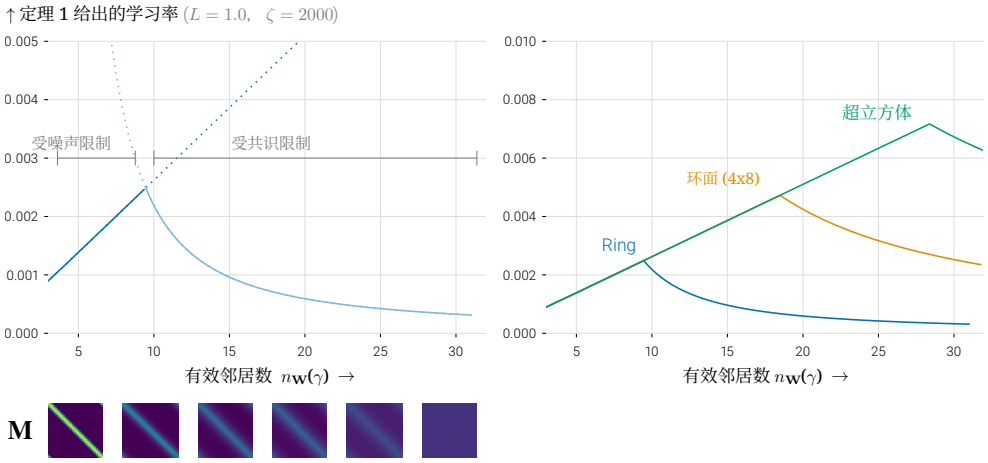


图 3：定理 6 所规定的最大学习率，展示参数 γ 的变化，该参数隐含着有效邻域大小（x 轴）以及一个平均矩阵 $\mathbf{M}$ （以热图形式绘制）。在左侧，我们展示了一个由 32 个工作节点组成的环形拓扑的详细情况；在右侧，我们将其与连接更紧密的拓扑结构进行比较。增大 γ （以及随之增大的 $n_{\mathbf{W}}$（γ））最初会因噪声降低而带来更高的学习率。在最优点处，共识的代价超过了进一步降低噪声所带来的收益。

4.3 主要推论

4.3.1 较大学习率：加快大误差下的收敛

现在，我们来考察噪声 σ^2 和异质性$\|\nabla F(\mathbf{x}^\star)\|^2$ $\mathbf{L}_{\mathbf{W}}{}^\dagger$ 都较小（相较于 $\mathcal{L}^{(0)}$）时定理 6 的情形，因此我们希望采用尽可能高的学习率，以确保目标函数快速下降（这与图1一致）。利用（25），我们可以针对每个控制局部邻域大小的参数 γ 求得一个速率（请记住，$\beta(\gamma)$ 取决于 γ）。接下来要做的是找到能带来最佳收敛保证（即最大学习率）的 γ 参数。正如前文所述，绝不应为了接近他人而降低学习率，因为在这种情形下（我们不受方差和异质性的影响），协作的目标是提高学习率。

我们在图 3中对此进行了说明，该图通过选择 $\omega = n_{\mathbf{W}}(\gamma)^{-1}$并针对不同的 γ取值计算（25）中的两项得到。曲线的线性部分（在共识占主导之前）的表达式见推论 7。

推论 7 设假设*4* 和*5* 成立，则最大的（忽略常数因子）学习率为：

$$\eta = (8\zeta/n_{\mathbf{W}}(\gamma) + 4L)^{-1},\ \textit{for } \gamma \textit{ such that } 4n_{\mathbf{W}}(\gamma)^{-1}\beta(\gamma)(2n_{\mathbf{W}}(\gamma)^{-1}\zeta + L) \leq L \qquad (31)$$

我们可以看到，在这种情况下，学习率与有效邻居的数量呈线性关系（这相当于采用大小与 $n_{\mathbf{W}}(\gamma)$ 成线性关系的小批量），直至达到一定数量

的邻居（右侧条件），或达到集中式性能（$\zeta = n_{\mathbf{W}}(\gamma)L$）。条件 γ 始终有解，因为当 $\gamma \approx 0$时，$\beta(\gamma)$和 $n_{\mathbf{W}}(\gamma)^{-1}$ 都接近 1，并且都会随着 γ 的增大而减小。该推论可直接由定理 $\omega = n_{\mathbf{W}}(\gamma)^{-1}$ 取 6得出。注意，若令 $\omega = \eta\beta(\gamma)/L$，则可得到一个稍紧一些的选择。

考察 $\beta(\gamma)$。现在我们评估 $\beta(\gamma)$ 以获得更精确的界。具体而言，选择$\mathbf{M}$ 如（21）所示，$\mathbf{L_M}$ 的特征值为：

$$\lambda_i^{\mathbf{L_M}} = \frac{1-\lambda_i^2}{1-\gamma\lambda_i^2}, \tag{32}$$

其中 λ_i 是$\mathbf{W}$的特征值。特别地，$\beta(\gamma)\mathbf{L_M} \preccurlyeq \mathbf{W}\mathbf{L_W}$ 意味着，对于所有满足 $\lambda_i \neq 1$（在此情况下自动成立）的 i，我们希望对所有 i 都有：

$$\beta(\gamma) \leq \frac{1-\gamma\lambda_i^2}{1-\lambda_i^2}(1-\lambda_i)\lambda_i = \frac{\lambda_i(1-\gamma\lambda_i^2)}{1+\lambda_i}. \tag{33}$$

现在我们作出一个简化假设，即 $\lambda_{\min}(\mathbf{W}) \geq \frac{1}{2}$（我们总可以通过取$\mathbf{W}' = (\mathbf{I}+\mathbf{W})/2$来保证这一点），但请注意，该理论无论如何都成立。作出这一简化假设的理由是：对于任意小的谱隙，（33）右侧的值总会在 $\lambda_2(\mathbf{W})$ 处取得最小值，γ 足够大时，因此 $\lambda_{\min}(\mathbf{W}) < 1$ 的实际值并不重要。特别是在这种情况下，忽略谱隙的影响，我们可以直接取：

$$\beta(\gamma) = \frac{1-\gamma\lambda_2(\mathbf{W})}{4} \geq \frac{1-\gamma}{4}, \tag{34}$$

注意，$\beta(\gamma)$ 允许在谱隙 $1-\lambda_2(\mathbf{W})$ 较大时采用较大的 γ，但只要满足 $\gamma < 1$，即使 $\lambda_2(\mathbf{W}) = 1$（无限图），我们也允许使用非平凡的学习率 $\eta > 0$。

最优选择 $n_{\mathbf{W}}(\gamma)$。利用第 3.1 节关于谱维数的结果，我们得到以下推论：

推论 8 在假设 *4* 和 *5* 下，并假设 $\lambda_{\min}(\mathbf{W}) \geq \frac{1}{2}$，通信图的谱维数为 $d_s > 2$，且 $\zeta \gg L$，则最高可能的学习率为

$$\eta = \frac{1}{8}\left(\frac{c_s(d_s-2)}{\zeta^2 L}\right)^{\frac{1}{3}}, \ \textit{obtained for } n_{\mathbf{W}}(\gamma) = \left(c_s(d_s-2)\frac{\zeta}{L}\right)^{\frac{1}{3}} \tag{35}$$

该结果由推论 7 得出；当 $\zeta \gg L$ 时，可写为：

$$\frac{L}{\zeta} \geq 8 n_{\mathbf{W}}(\gamma)^{-2}\beta(\gamma) = n_{\mathbf{W}}(\gamma)^{-3} c_s(d_s-2), \tag{36}$$

其中右侧部分是将 $\beta(\gamma)$ 在（34）中的表达式代入$n_{\mathbf{W}}(\gamma)^{-1} \leq {}^{2}{}_{(1-\gamma)\, c_s(d_s-2)}$（14）中得到的（假设 $\gamma \geq 1/2$）。然后，可以解出 $1-\gamma$。除假设 4 之外的假设使得在这一特定情况下可以得到一个简单结果，但也可以很容易地得到适用于 $d_s \leq 2$ 和 $\zeta < L n_{\mathbf{W}}(\gamma)$的类似表达式。

4.3.2 小学习率：任意接近最优点

定理 6 给出了收敛到 $\mathbf{x}_\eta^\star$（即去中心化随机梯度下降的不动点）的结果。上一节中，我们研究了大学习率下去中心化随机梯度下降的行为。在定理 9中，我们关注误差较小的情况，此时 方差 和 异质性 项占主导地位，因此我们希望采用小学习率 η。在这种情况下，我们界定当前迭代点与真实最小化点 $\mathbf{x}^\star$ 之间的距离，而非与 $\mathbf{x}_\eta^\star$之间的距离。我们还给出了一个结果，消除了对 $\mathbf{x}_\eta^\star$的所有依赖，并且只显式依赖于学习率 η。

定理 9 在与定理 *6* 和推论 *8*相同的假设以及学习率条件下，我们有：

$$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}} \leq 2(1-\eta\mu)^t \mathcal{L}^{(0)} + \frac{2\eta\sigma_{\mathbf{M}}^2}{\mu} + 2\eta^2(1+\kappa)\|\mathbf{L}_{\mathbf{W}}{}^\dagger \nabla F(\mathbf{x}_\eta^\star)\|^2 \tag{37}$$

我们还可以从界中消去$\mathbf{x}_\eta^\star$，得到：

$$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}} \leq 2(1-\eta\mu)^t \mathcal{L}^{(0)} + \frac{6\eta\sigma_{\mathbf{M},\star}^2}{\mu} + 6\eta^2\kappa p^{-1}\Delta_{\mathbf{W}}^2,$$

其中 $\sigma_{\mathbf{M},\star}^2 = (n_{\mathbf{W}}(\gamma)^{-1} + \omega)\,\mathbb{E}\left[\|\nabla F_\xi(x^\star) - \nabla F(x^\star)\|^2\right]$ 且 $p^{-1} = max_\eta \frac{\|L_{\mathbf{W}}{}^\dagger \nabla F(x_\eta^\star)\|^2}{\|\nabla F(x_\eta^\star)\|^2_{\dagger}}$ $(x)\,(x)\,LW$，因此 $p \geq 1 - \lambda_2(W)$，且 $\Delta_{\mathbf{W}}^2 = \|\nabla F(x^\star)\|^2 L_{\mathbf{W}}{}^\dagger$

范数 $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}}^2$ 考察局部平均邻域的收敛，但 $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}}^2 \geq \|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$ 因为 $\mathbb{1}$ 是 $\mathbf{M}$ 对应特征值 1的特征向量。现在，我们简要讨论该推论中的各项，然后给出证明。

异质性项。 异质性所导致的项仅取决于真实最优解 $\mathbf{x}^\star$ 与不动点 $\mathbf{x}_\eta^\star$之间的距离，随后我们将其转化为关于 $\|\nabla F(\mathbf{x}^\star)\|^2$ $\mathbf{L}_{\mathbf{W}}{}^\dagger$ 的条件。具体而言，它不受 $\mathbf{M}$ 的选择影响（因此也不受 γ 的影响）。

常数 p。 我们引入常数 p 以消除对 $\mathbf{x}_\eta^\star$的显式依赖。p^{-1} 直观而言，它表示 $\mathbf{L}_{\mathbf{W}}{}^\dagger$ 在 $\nabla F(\mathbf{x}_\eta^\star)$ 方向上的大小。例如，若 $\nabla F(\mathbf{x}_\eta^\star)$ 是 $\mathbf{W}$ 对应特征值 λ的特征向量，则有 $p = 1-\lambda$。最坏情况下，$p = 1 - \lambda_2(\mathbf{W})$，但通常 p可以好得多：当异质性均匀分布时，相较于相距较远的节点上各自具有差异很大的函数，情况会更好。

方差项。 在这种情况下，通过尽可能减小 n，可获得最大的方差缩减（量级为 ω 和 $n_{\mathbf{W}}(\gamma)^{-1}$）。对于大到无法满足$n_{\mathbf{W}}(\gamma)^{-1} \approx n^{-1}$的学习率，降低学习率会通过两种方式减小方差项：(I) 直接通过 η 项实现；(II) 间接地使得可以取更小的 $n_{\mathbf{W}}(\gamma)^{-1}$。

对于非常大的（无限）图，我们可以取 $\omega = n_{\mathbf{W}}(\gamma)^{-1}$，在这种情况下，定理 6给出的最小 $n_{\mathbf{W}}(\gamma)^{-1}$ 为 $n_{\mathbf{W}}(\gamma)^{-1}\beta(\gamma) = \eta L$。利用谱维数结果（例如取 $d_s > 2$），我们可以得到（类似于推论 8）我们可以取

$\beta(\gamma) = n_{\mathbf{W}}(\gamma)^{-1} c_s(d_s-2)/8$，因此：

$$n_{\mathbf{W}}(\gamma)^{-1} = \sqrt{\frac{8\eta L}{c_s(d_s-2)}}, \tag{38}$$

因此，对于这一选择，$n_{\mathbf{W}}(\gamma)^{-1}$ 的残差方差项的阶为：

$$\mathcal{O}\left(\frac{\eta^{\frac{3}{2}}}{\mu}\sqrt{\frac{L}{c_s(d_s-2)}}\,\mathbb{E}\left[\|\nabla F_\xi(\mathbf{x}^\star)-\nabla F(\mathbf{x}^\star)\|^2\right]\right) \tag{39}$$

特别是，降低学习率 η 时，由于还能获得更多有效邻居，我们可以实现超线性扩展。再次注意，情况 $d_s \leq 2$ 也可以用相同的方法处理。

证明 [定理 9] 我们先写出：

$$\|\mathbf{x}^{(t)}-\mathbf{x}^\star\|^2_{\mathbf{M}} \leq 2\|\mathbf{x}^{(t)}-\mathbf{x}^\star_\eta\|^2_{\mathbf{M}} + 2\|\mathbf{x}^\star_\eta-\mathbf{x}^\star\|^2_{\mathbf{M}} \leq 2\mathcal{L}^{(t)} + 2\|\mathbf{x}^\star_\eta-\mathbf{x}^\star\|^2. \tag{40}$$

定理 6 保证 $\mathcal{L}^{(t)}$ 变得很小，因此我们只需界定 $\mathbf{x}^\star_\eta$ 与$\mathbf{x}^\star$之间的距离。

1 - 到全局最小值点的距离。 我们定义 $\mathbf{x}^\star_\eta = \mathbb{1}\mathbb{1}^\top \mathbf{x}^\star_\eta/n$。利用 $\mathbf{x}^\star_\eta$ 和 $\mathbf{x}^\star$ 都处于共识状态这一事实，以及 $\mathbb{1}^\top \nabla F(\mathbf{x}^\star_\eta) = 0$ (可直接由 (23) 得出)，我们写出：

$$\begin{aligned} D_F(\mathbf{x}^\star, \mathbf{x}^\star_\eta) &= F(\mathbf{x}^\star) - F(\mathbf{x}^\star_\eta) - \nabla F(\mathbf{x}^\star_\eta)^\top(\mathbf{x}^\star - \mathbf{x}^\star_\eta) \\ &= F(\overline{\mathbf{x}^\star_\eta}) - F(\mathbf{x}^\star_\eta) - \nabla F(\mathbf{x}^\star_\eta)^\top(\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star_\eta) + F(\mathbf{x}^\star) - F(\overline{\mathbf{x}^\star_\eta}) \\ &\leq D_F(\overline{\mathbf{x}^\star_\eta}, \mathbf{x}^\star_\eta), \end{aligned} \tag{41}$$

其中，最后一行是因为$\mathbf{x}^\star$ 是共识空间上 F 的最小化解。因此：

$$\begin{aligned} \|\mathbf{x}^\star_\eta - \mathbf{x}^\star\|^2 &= \|\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star\|^2 + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \frac{1}{\mu} D_F(\mathbf{x}^\star, \mathbf{x}^\star_\eta) + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \frac{1}{\mu} D_F(\overline{\mathbf{x}^\star_\eta}, \mathbf{x}^\star_\eta) + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \left(1+\frac{L}{\mu}\right)\|\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star_\eta\|^2 = \eta^2\left(1+\frac{L}{\mu}\right)\|\mathbf{L}_{\mathbf{W}}{}^\dagger \nabla F(\mathbf{x}^\star_\eta)\|^2. \end{aligned}$$

请注意，该结果取决于不动点处梯度的异质性模式，即使$\mathbf{W}$ 没有谱隙，该结果也可能有界（甚至很小）。然而，在最坏情况下，该量与谱隙倒数的平方成正比。

2 - 单调性（关于 η）。现在我们证明，随着 η 增大，$\|\nabla F(\mathbf{x}^\star_\eta)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}$ 会减小，因此在 $\eta = 0$ 时达到最大值，对应于 $\mathbf{x}^\star_\eta = \mathbf{x}^\star$。具体而言：

$$\frac{\mathbf{d}\|\nabla F(\mathbf{x}^\star_\eta)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}}{\mathbf{d}\eta} = \frac{\mathbf{d}\left[\eta^{-2}\|\mathbf{x}^\star_\eta\|^2_{\mathbf{L}_{\mathbf{W}}}\right]}{\mathbf{d}\eta} = -\frac{2\|\mathbf{x}^\star_\eta\|^2_{\mathbf{L}_{\mathbf{W}}}}{\eta^3} + 2\eta^{-2}(\mathbf{x}^\star_\eta)^\top \mathbf{L}_{\mathbf{W}} \frac{\mathbf{d}\mathbf{x}^\star_\eta}{\mathbf{d}\eta}$$

对不动点条件求导，可得

$$\eta\nabla^2 F(\mathbf{x}_\eta^\star)\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} + \nabla F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} = 0, \tag{42}$$

因此：

$$\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} = -\left(\eta\nabla^2 F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\right)^{-1}\nabla F(\mathbf{x}_\eta^\star) = \eta^{-1}\left(\eta\nabla^2 F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\right)^{-1}\mathbf{L}_\mathbf{W}\mathbf{x}_\eta^\star. \tag{43}$$

将其代入前一个表达式，并利用 $\nabla^2 F(\mathbf{x}_\eta^\star)$ 是半正定的，可得：

$$\begin{aligned}\frac{\mathbf{d}\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}}{\mathbf{d}\eta} &= -\frac{2}{\eta^3}(\mathbf{x}_\eta^\star)^\top\left[\mathbf{L}_\mathbf{W} - \mathbf{L}_\mathbf{W}\left(\mathbf{L}_\mathbf{W} + \eta\nabla^2 F(\mathbf{x}_\eta^\star)\right)^{-1}\mathbf{L}_\mathbf{W}\right]\mathbf{x}_\eta^\star \\ &\le -\frac{2}{\eta^3}(\mathbf{x}_\eta^\star)^\top\left[\mathbf{L}_\mathbf{W} - \mathbf{L}_\mathbf{W}\mathbf{L}_\mathbf{W}^\dagger\mathbf{L}_\mathbf{W}\right]\mathbf{x}_\eta^\star = 0.\end{aligned}$$

3 - 消去 $\mathbf{x}_{\eta^\star}^\star$。 根据 p 的定义，我们可以写成：

$$\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2 \le p^{-1}\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger} \le p^{-1}\|\nabla F(\mathbf{x}^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}. \tag{44}$$

请注意，为了利用 η 中 $\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}$ 的单调性，我们必须对这个常数 p 加以界定，因为当 $\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2$ 时，这一结论并不成立。对于方差，我们可以写成：

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right] &\le 3\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F_{\xi_t}(\mathbf{x}^\star)\|^2\right] \\ &\quad + 3\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^\star) - \nabla F(\mathbf{x}^\star)\|^2\right] + 3\|\nabla F(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}^\star)\|^2 \\ &\le 3\sigma^2_{\mathbf{M},\star} + 3\left(\zeta + L\right)D_F(\mathbf{x}^\star, \mathbf{x}_\eta^\star).\end{aligned}$$

从这里开始，我们使用公式（41），并得到：

$$\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right] \le 3\sigma^2_{\mathbf{M},\star} + 3L\left(\zeta + L\right)\eta^2\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2. \tag{45}$$

为获得最终结果，我们利用 $\eta(n_\mathbf{W}(\gamma)^{-1} + \omega)(\zeta + L) \le 1/4$，这得益于学习率所满足的条件。

■

4.3.3 与现有工作的比较。

按照 Koloskova 等人（2020）的形式表达，我们可以通过选取 η 作为可能的最大常数（如~~推论 8 中所示），或选取 $\eta = O(1/(\mu T))$，将前面的推论总结为以下结果。这里，O 表示忽略对数因子后的不等式；请回忆，$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2_\mathbf{M} \ge \|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$。我们回顾一下：$L$ 是全局目标函数 f的平滑性，ζ 是随机函数 f_ξ的平滑性，μ 是强凸参数，d_s 是八卦矩阵 W 的谱维数（并且我们假设 $d_s > 2$），而c_s 是相应的常数。

推论 10（最终结果。）在与推论8相同的假设下，存在一种学习率（等价地，衰减参数 γ^*_{large} 和 γ^*_{small}）的选取方式，使得经过 T 步去中心化随机梯度下降 $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2$后，到全局最优点的期望平方距离的阶为：

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 T n_{\mathbf{W}}(\gamma^*_{\text{small}})} + \frac{L\Delta^2_{\mathbf{W}}}{\mu^3 p T^2} + \exp\left[-n_{\mathbf{W}}(\gamma^*_{\text{large}})\frac{\mu}{\zeta}T\right]\right), \tag{46}$$

其中 $\Delta^2_{\mathbf{W}}$ 和 p在定理 9中定义，且 $\mathbf{x}^{(t)}$ 为平均参数。在较小和较大学习率设置下，最优有效邻居数分别为：

$$n_{\mathbf{W}}(\gamma^*_{\text{small}}) = \min\left(\sqrt{\frac{d_s T}{L c_s}}, n\right) \text{ and } n_{\mathbf{W}}(\gamma^*_{\text{large}}) = \min\left(\left(\frac{c_s d_s \zeta}{L}\right)^{\frac{1}{3}}, n\right). \tag{47}$$

这一结果可与 Koloskova等人 (2020年)给出的结果进行对比：

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 T}\left[\frac{1}{n} + \frac{L}{\mu(1-\lambda_2(\mathbf{W}))T}\right] + \frac{L\Delta^2}{\mu^3(1-\lambda_2(\mathbf{W}))^2 T^2} + \exp\left[-\frac{\mu}{(1-\lambda_2(\mathbf{W}))\zeta}T\right]\right), \tag{48}$$

现在我们可以得出以下观察结果。

学习率调度。 此处，学习率可以选择为 $n_{\mathbf{W}}\ \gamma^*_{\text{large}}\ /\zeta$ $\eta_{\text{small}} = \tilde{O}\ \ \mu T\ ^{-1}$ $\eta_{\text{large}}\ =$ ( )，也可以选择为 (( ))。实际应用中，可以先采用较大学习率；当训练不再改善时（此时异质性 / 方差项占主导），再切换到 η 。

指数下降项。 我们首先展示指数下降项的显著改进。事实上， $n_{\mathbf{W}}(\gamma^*_{\text{large}})/(1-\lambda_2(\mathbf{W}))$，即我们的分析所允许的最大学习率与现有分析所允许的最大学习率之比，始终较大，因为 $n_{\mathbf{W}}(\gamma^*_{\text{large}}) \geq 1$ 和$1-\lambda_2(\mathbf{W}) \leq 1$。此外，在我们的分析中，指数下降项不再受谱隙影响；谱隙只会影响 $n_{\mathbf{W}}(\gamma)$的大小。即使在 $\zeta = L$ （此时只需 $n_{\mathbf{W}}(\gamma) = 1$ ）的情况下，这一改进仍然成立。这是因为异质性只影响低阶项，因此，当协作没有带来任何收益时，也不会损害收敛。

异质性的影响。异质情形下的改进并不依赖于某个 γ，而是依靠以非最坏情况的方式界定异质性。事实上， $\zeta_{\mathbf{W}}$ 和p 刻画了异质性在节点间的分布方式与图的实际拓扑之间的相互作用。请注意，这与Koloskova等人(2020年)给出的下界并不矛盾，因为 $\Delta^2_{\mathbf{W}}/p = \Delta^2/(1-\lambda_2(\mathbf{W}))^2$ 对应的是最坏情况。在最坏情况下， $\nabla F(\mathbf{x}^\star)$ 的异质性模式与$\mathbf{L}_{\mathbf{W}}$的最小特征值相对应；也就是说，相距很远的节点具有截然不同的目标函数。然而， p通常能给出更精细的界，具体取决于实际的异质性模式。

方差项。 两种分析的一个关键区别在于涉及 σ^2 的方差项。两种分析都取决于单个节点的方差 $\sigma^2/(\mu T)$，它

随后乘以一个“方差缩减”项。在两种情况下，该项的形式都是$n_{\mathbf{W}}(\gamma)^{-1}+\eta L\beta(\gamma)^{-1}$。不过，标准分析隐式地使用了 $\gamma=1$，因此 $n_{\mathbf{W}}(\gamma)=n$,~以及 $\beta(\gamma)=1-\lambda_2(\mathbf{W})$。接着，对 $\eta=O(1/(\mu T))$ 取值即可得到式 (48)。而我们的分析则依赖于调节 γ，使得 $n_{\mathbf{W}}(\gamma)^{-1}+\eta L\beta(\gamma)^{-1}$ 尽可能小，因此严格优于仅考虑 $\gamma=1$。假设图的谱维数 $d_s>2$ 已知，可得式 (46)；但任何能精确关联 $n_{\mathbf{W}}(\gamma)$ 与 γ 的假设，都能得到类似结果。

尽管 Koloskova 等人（2020）方差项中的 $\tilde{O}(T^{-2})$ 似乎优于我们的$\tilde{O}(T^{-3/2})$项，但这具有误导性，因为在这种情况下，常数非常重要。我们的收敛速率通过对 γ 进行优化得出，这考虑到了这样一种情况：如果 $\tilde{O}(T^{-2})$项占主导地位，那么缩小邻域会更好。在这种情况下，无论如何我们都无法从 n^{-1} 方差缩减中获益。相反，我们的结果对式（48）中的两个方差项进行了最优平衡。得益于这种平衡，我们发现，对于谱维数为$d_s>2$的图，方差会以 $\tilde{O}(T^{-\frac{3}{2}})$的速度下降，学习率为 $\tilde{O}(T^{-1})$；这是学习率降低与增加有效邻居共同作用的结果。对于有限图，这种效应会达到上限 $n_{\mathbf{W}}(\gamma)=n$。

最后请注意，我们的分析与Koloskova等人(2020年的分析允许对标准框架进行不同的推广：我们的分析适用于任意大的（无限）图，而Koloskova等人(2020年则能够处理仅要求较弱（多轮）连通性假设的时变图。

5. 深度学习中的实证意义

虽然本文的理论结果针对的是凸函数，但这项工作的最初动机源于深度学习中的观察。首先，在深度学习中，训练初期使用较大的学习率至关重要（Li 等人，2019年）。与现有理论的建议相反，在去中心化优化中，我们并不会像独自训练时那样降低学习率（即使数据具有异质性）。其次，我们发现，拓扑的谱隙无法预测该拓扑在深度学习实验中的表现。

在本节中，我们在 CIFAR-10（Krizhevsky 等人）上使用 VGG-11 模型（Simonyan 和 Zisserman，2015年）对多种 32 工作节点拓扑结构进行实验。与近期其他研究（Lin 等人，2021年；Vogels 等人，2021年）一样，我们选择了这一较早的模型，因为它不包含批归一化（Ioffe 和 Szegedy，2015年），而批归一化给去中心化随机梯度下降带来了一个独立的挑战。实验设置的完整细节请参阅（Vogels 等人，2022）的附录 E。我们的拓扑集合包括环和环面等正则图，也包括二叉树（Vogels 等人，2021年）和社交网络（Davis 等人 （1930年）等非正则图，以及时变指数方案（Assran等人，2019年）。我们关注训练的初始阶段，在我们的实验中为 25k 步；在这一阶段，训练损失和测试损失都近乎线性地收敛。研究发现，在此阶段使用较大的学习率对获得良好的泛化能力十分重要（Li 等人，2019年）。

图 4 展示了所有拓扑结构在前 2.5k 个随机梯度下降步数后达到的损失，以及在密集学习率网格下的结果。这些曲线的整体结构与各向同性二次函数的结果相同 图 1：（稀疏）平均能在小学习率下略微加快速度，但相较于单独训练，显著的提升来自能够提高学习

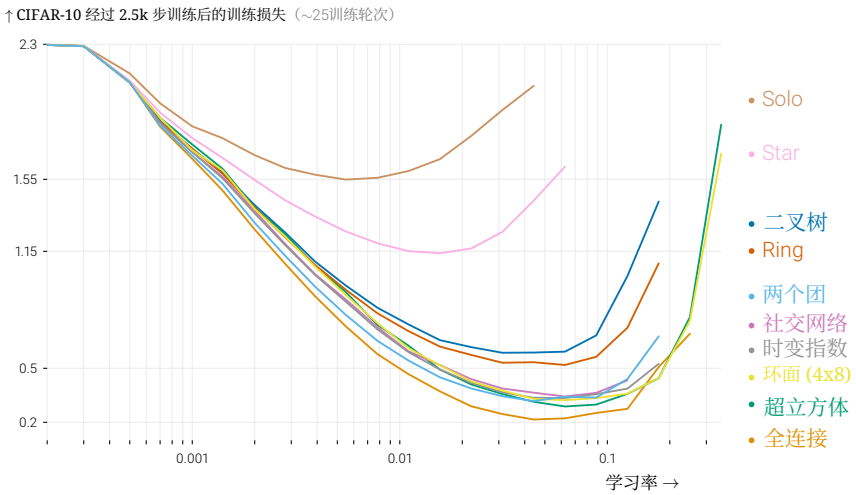


图 4：采用多种图拓扑时，经过 2.5k 个随机梯度下降步数后达到的训练损失。在所有情况下，小学习率下的平均都能略微提升速度，但相比单独训练，能够提高学习率带来了显著收益。尽管星形的谱隙（0.031）大于环（0.013），其表现却更差，也不允许使用大学习率。作为参考，不同规模的全连接图对应的类似曲线见Vogels 等人（2022）的附录。

率。最佳方案支持的学习率几乎与 32 个全连接工作节点相同，性能也十分接近。

我们还发现，第3.1节 中介绍的随机游走可作为深度学习中工作节点间方差的良好模型。图 5 展示了经过 100 个随机梯度下降步数后工作节点之间的经验协方差。与各向同性二次函数的情况一样，对于某个衰减率，随机游走过程中的协方差能够准确地模拟该协方差 γ。

最后，我们观察到，通过随机游走中的方差缩减（第3.1节）计算出的有效邻居数，能够准确描述在我们的任务中，针对图拓扑调整学习率后的相对性能，其中也包括不规则拓扑和时变拓扑。这与拓扑的谱隙形成鲜明对比，因为我们发现谱隙无法预测性能。我们拟合了一个衰减率 $\gamma = 0.951$，它似乎捕捉到了我们问题的具体特征，并在图 6中展示了相关性。

(Vogels 等人，2022）的附录 F 在不同设置下重复了相同的实验。在该设置中，我们使用了更大的图（包含 64 个工作节点）、不同的模型和数据集（在 Fashion MNIST 上训练的多层感知机 Xiao 等人 (2017年)），且不使用动量或权重衰减。该设置下的结果在定性上与上文所述结果相当。

6. 结论

我们证明，去中心化学习中的稀疏平均使得采用更大的学习率成为可能，并能加快训练速度。使用最优的大学习率时，无法保证工作节点的模型始终接近其全局平均值。强制实现全局共识通常没有必要，而且它所要求的小学习率可能适得其反。事实上，

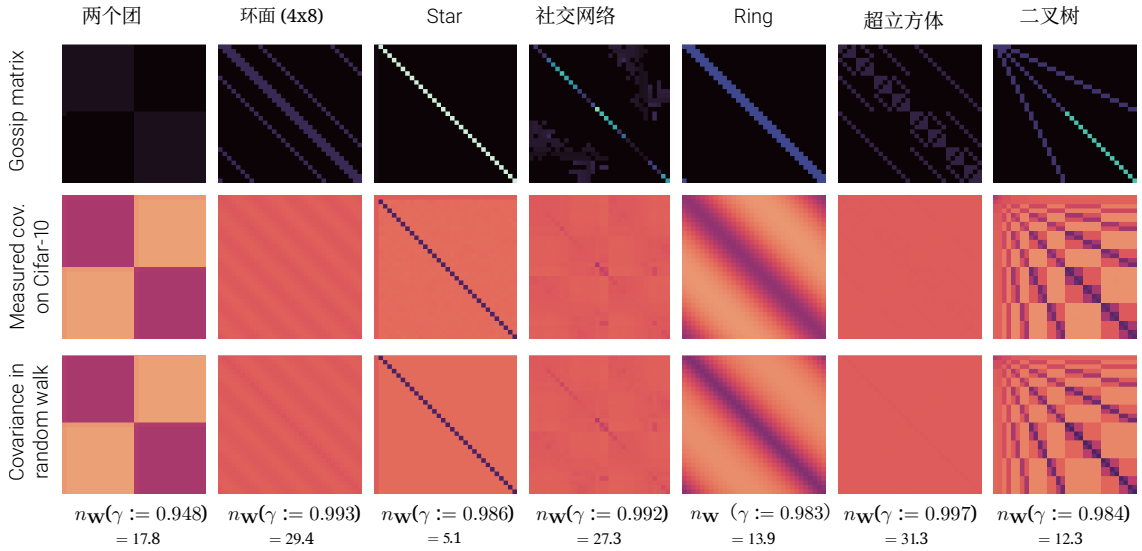


图 5：使用不同图（顶行）的工作节点之间，在 CIFAR-10 上测得的协方差（第二行）。经过 10 个训练轮次后，我们保存模型的检查点，并重复进行 100 个随机梯度下降步数的训练，从而为 32 个工作节点得到 100 个模型。我们展示了工作节点之间归一化后的协方差矩阵。这些矩阵与第3.1节（第三行）中随机游走过程的协方差非常吻合。我们列出了拟合得到的衰减参数及相应的“有效邻居数”。

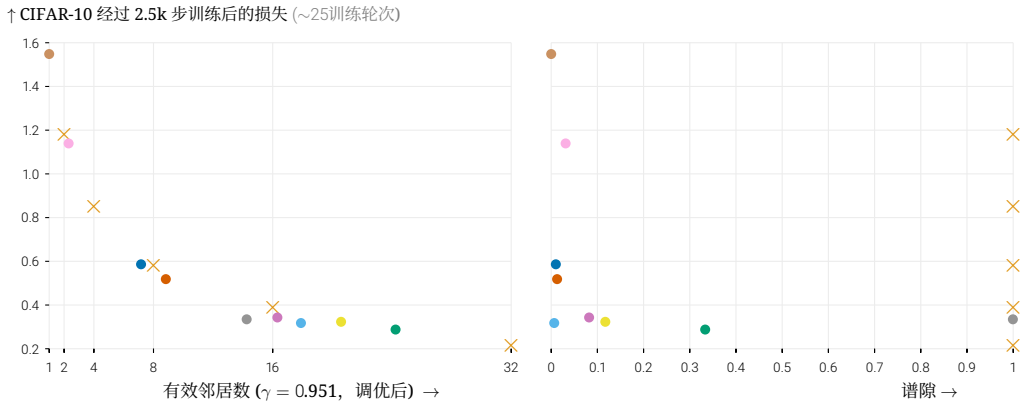


图 6：采用最优学习率时，所有研究的拓扑结构在 2.5k 步后的 CIFAR-10 训练损失。颜色与 图 4相同，而 ×表示工作节点数量各不相同的全连接图。拟合能够捕捉问题特性的衰减参数 $\gamma = 0.951$ 后，有效邻居数（左图）可通过随机游走中的方差缩减来衡量（如第3.1节所示），与这些拓扑结构的谱隙（右图）相比，它能更好地解释这些图的相对性能。

即使学习率很高，模型也仍然接近其加权邻域中的某种局部平均值。工作节点受益于一定数量的“有效邻居”，其数量可能少于整个图中的节点数；这些邻居使工作节点能够采用更大的学习率，同时在“局部邻域”内保持足够的共识。

当节点的局部函数存在异质性时，也有类似的见解：如果局部邻域之间的异质性较小，就没有必要在整个网络中强制进行全局平均。此外，在训练的早期阶段，模型都离全局最优点很远，因此也无需补偿异质性。

基于这些见解，我们建议稀疏分布式学习算法的实践者不要只关注图拓扑的谱隙，而应进一步考察实际使用的“有效邻居数”。我们也希望这些见解能促使理论研究者留意那些人为限制学习率的假设，即使这些假设在最坏情况下是紧的。事实上，谱隙在去中心化文献中无处不在，这有时会掩盖一些微妙的现象，例如我们所强调的学习率方差呈超线性下降。

我们通过实验表明，我们的结论在深度学习中也成立，但将我们的理论推广到非凸情形仍是一个重要的开放方向，可能会揭示有趣的新现象。另一个值得探索的方向是更深入地理解非正则图的有效邻居数（超越最坏情况）。

致谢与经费披露

本项目由 SNSF 资助，资助编号为 200020_200342。

我们感谢 Lie He 与我们进行了宝贵的交流，并指出了拓扑的谱隙与其经验性能之间的差异。

我们还感谢 Raphaël Berthier 的有益讨论，这些讨论帮助我们阐明了有效邻居数与谱维数之间的联系。

我们还感谢 Aditya Vardhan Varre、Yatin Dandi 和 Mathieu Even 对稿件提出的意见。

参考文献

Mahmoud Assran、Nicolas Loizou、Nicolas Ballas 和 Michael G. Rabbat。用于分布式深度学习的随机梯度推送。在 国际机器学习大会论文集，第 97 卷，第 344–353 页，2019年。

Raphaël Berthier。梯度下降与八卦算法的分析和加速。巴黎科学与文学大学博士论文，2021年。

Raphaël Berthier、Francis R. Bach 和 Pierre Gaillard。使用雅可比多项式迭代加速给定维度网络中的八卦算法。*SIAM 数学与数据科学杂志*，2(1)：24–47，2020年。

Yatin Dandi、Anastasia Koloskova、MartinJaggi 和 SebastianU. Stich。面向去中心化学习的感知数据异质性的混合方法。*CoRR*，abs/2204.06477，2022。

Allison Davis、Burleigh Bradford Gardner 和 Mary R Gardner。《*Deep South*：种姓与阶级的社会人类学研究》。南卡罗来纳大学出版社，1930年。Mathieu Even、Hadrien Hendrikx 和 Laurent Massoulie。异构延迟下的去中心化优化：一种连续时间方法。*arXiv* 预印本 *arXiv:2106.03585*，2021年。Sergey Ioffe 和 Christian Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。发表于人工智能与统计学国际会议论文集，第37卷，第448–456页，2015年。Anastasia Koloskova、Nicolas Loizou、Sadra Boreiri、Martin Jaggi 和 Sebastian U. Stich。变化拓扑与局部更新下去中心化随机梯度下降的统一理论。发表于人工智能与统计学国际会议论文集，第119卷，第5381–5393页，2020年。Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton。CIFAR-10（加拿大高等研究院）。B. Le Bars、Aurélien Bellet、Marc Tommasi 和 Anne-Marie Kermarrec。没错，拓扑在去中心化优化中至关重要：异构数据下的精细收敛分析与拓扑学习。$CoRR$，abs/2204.04452，2022。Yuanzhi Li、Colin Wei 和 Tengyu Ma。探究初始大学习率在神经网络训练中的正则化效应。发表于$NeurIPS$，第11669–11680页，2019年。Xiangru Lian、Ce Zhang、Huan Zhang、Cho-Jui Hsieh、Wei Zhang 和 Ji Liu。去中心化算法能否胜过集中式算法？以去中心化并行随机梯度下降为例。发表于$NeurIPS$，第5330–5340页，2017年。Xiangru Lian、Wei Zhang、Ce Zhang 和 Ji Liu。异步去中心化并行随机梯度下降。发表于人工智能与统计学国际会议论文集，第80卷，第3049–3058页，2018年。Tao Lin、Sai Praneeth Karimireddy、Sebastian U. Stich 和 Martin Jaggi。准全局动量：加速异构数据上的去中心化深度学习。发表于人工智能与统计学国际会议论文集，第139卷，第6654–6665页，2021年。Yucheng Lu 和 Christopher De Sa。去中心化训练的最优复杂度。发表于人工智能与统计学国际会议论文集，第139卷，第7111–7123页，2021年。Giovanni Neglia、Chuan Xu、Don Towsley 和 Gianmarco Calbi。去中心化梯度方法：拓扑是否重要？发表于人工智能与统计学国际会议，第108卷，第2348–2358页，2020年。Dominic Richards 和 Patrick Rebeschini。具有线性加速的去中心化非参数回归的最优统计速率。发表于$NeurIPS$，第1214–1225页，2019年。Dominic Richards 和 Patrick Rebeschini。分布式随机次梯度下降中的图依赖隐式正则化。《机器学习研究杂志》，21：34:1–34:44，2020年。Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。发表于国际学习表征会议，2015年。Hanlin Tang、Xiangru Lian、Ming Yan、Ce Zhang 和 Ji Liu。 d^2：基于去中心化数据的去中心化训练。发表于人工智能与统计学国际会议论文集，第80卷，第4855–4863页，2018年。

Thijs Vogels、Lie He、Anastasia Koloskova、Sai Praneeth Karimireddy、Tao Lin、Sebastian U. Stich 和 Martin Jaggi。面向异构数据去中心化深度学习的 Relaysum。发表于*NeurIPS*，第 28004–28015 页，2021年。Thijs Vogels、Hadrien Hendrikx 和 Martin Jaggi。超越谱隙：拓扑在去中心化学习中的作用。发表于*NeurIPS*，2022。Jianyu Wang、Anit Kumar Sahu、Zhouyi Yang、Gauri Joshi 和 Soummya Kar。MATCHA：通过匹配分解采样加速去中心化随机梯度下降。*CoRR*，abs/1905.09435，2019年。Han Xiao、Kashif Rasul 和 Roland Vollgraf。Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。*CoRR*，abs/1708.07747，2017年。Bicheng Ying、Kun Yuan、Yiming Chen、Hanbin Hu、Pan Pan 和 Wotao Yin。指数图可证明能高效实现去中心化深度训练。发表于*NeurIPS*，第 13975–13987 页，2021年。