

从近似高斯消元到图学习： Laplacian 理论、可迁移的定理与 Graph DeltaNet 的接口

Reading Rasmus Kyng's *Approximate Gaussian Elimination*

研究阅读笔记

以 Kyng (2017) 博士论文与 Kyng-Sachdeva (2016) 为主线
文献核对截至 2026 年 10 月 10 日；非 JMLR 已发表或录用论文

Abstract

近似高斯消元的核心并不是把高斯消元逐项截断，而是在消元产生的稠密相互作用上构造稀疏随机替代物，同时用谱序、有效电阻和矩阵鞅控制全局误差。本笔记从图 Laplacian 的能量形式出发，推导 Schur 补的变分解释、星形到团的精确消元、谱近似的继承、逆算子与扩散稳定性，以及可微线性求解的梯度公式。随后解释 Kyng-Sachdeva 的无偏团采样为何有效、为何无偏性本身还不够。面向 graph learning，区分能量保持、终端电阻保持、图池化、隐式传播和训练梯度保证；面向多前驱 DAG Graph DeltaNet，提取多端口消元、非交换路径和与非对称扰动界，而不把 Laplacian 定理错误地升级为任意图递归的保证。最后给出可复现的数值恒等式核验证及按优先级组织的研究路线。全文明确区分原文定理、标准事实、本文推导和待验证设想。

Keywords: 图 Laplacian，近似高斯消元，Schur 补，谱稀疏化，有效电阻，图学习，Graph DeltaNet

1 读什么，以及对当前研究最有价值的是什么

Kyng (2017) 的博士论文 *Approximate Gaussian Elimination* 是一组相互关联的算法与分析技术：通过近似消元构造 Laplacian 求解器、稀疏 Schur 补，并向块矩阵和有向系统扩展。其中最合适逐式学习的是 Kyng and Sachdeva (2016)：它把随机采样直接嵌入 Cholesky 消元，避免显式生成每一步的全部 fill-in。本笔记以这篇论文的原始算法为锚点，而不是把不同年代的变体混成一个算法。

对当前以多前驱 DAG、残差 Delta 更新、Graph-UT、多端口 WY 和 GPU 分块为主线的研究，最值得迁移的是**消元接口的设计方法**：先精确定义块内部被消去以后，边界必须保留的算子和源项；再决定是否、在什么度量下近似这个接口。Laplacian 提供了一个异常完整的成功范例，但其成功依赖对称性、正权重、半正定序和电阻几何。保留这些假设的模块可以直接利用谱理论；不具备这些假设的 DAG 递归需要单独的非对称分析。

全文使用四种标记：**[原文]** 表示已发表论文或博士论文中的结论；**[标准]** 表示线性代数或图谱理论事实；**[推导]** 表示在本文列出的假设下由这些事实直接推出、并非声称新定理的结果；**[设想]** 表示尚需算法、复杂度或学习实验验证的方向。证明只覆盖所写命题；第6节是原论文概率证明的路线图，不冒充完整复现其全部技术引理。

研究目标	最有用的工具	保证止于何处
稀疏图上的稳定传播	谱近似、预条件、resolvent 界	不自动保证任意非线性 GNN 或训练轨迹
保留关键节点的图缩减	Schur 补、谐波延拓、终端电阻	不是保留所有原节点信息或全部特征向量
可微 Laplacian 模块	原算子隐式微分、伴随求解	前向近似不等于梯度近似
多前驱 DAG Delta 递归	精确多端口消元、非交换路径和	不能直接套用 PSD 谱序或电阻采样
并行与 GPU 实现	块消元、依赖图、设置成本摊销	代数可并行不等于已获得 GPU 加速

Table 1: 从研究需求反向选择定理，而不是从“都在做图消元”推断可直接迁移。

2 Laplacian 的基本语言与两个不能混用的误差约定

2.1 能量、零空间和兼容的右端项

设 $G = (V, E, w)$ 为无向连通图， $n = |V|$ ， $m = |E|$ ，每条无向边只计一次；允许平行边，边权严格为正。给边任意定向只为定义关联矩阵，令 $b_e = e_u - e_v$ 、 $B \in \mathbb{R}^{m \times n}$ ，则

$$L = B^\top W B = \sum_{e \in E} w_e b_e b_e^\top = D - A, \quad x^\top L x = \sum_{e=\{u,v\}} w_e (x_u - x_v)^2. \quad (1)$$

边定向不会改变 L 。对多通道节点表示 $H \in \mathbb{R}^{n \times d}$ ，

$$\|H\|_{L,F}^2 := \text{tr}(H^\top L H) = \sum_{\{u,v\} \in E} w_{uv} \|H_u - H_v\|_2^2. \quad (2)$$

这解释了图平滑正则的准确含义：惩罚加权邻接差异，而不是无条件地让所有节点表示相同。常量信号不产生能量，因而连通图满足 $L \succeq 0$ ， $\ker L = \text{span}\{\mathbf{1}\}$ 。 $\|x\|_L$ 在全空间只是半范数，在 $\mathbf{1}^\perp$ 上才是范数。

方程 $Lx = b$ 可解的条件是 $\mathbf{1}^\top b = 0$ 。其解相差一个常量；选取均值为零的规范后， $x = L^\dagger b$ ，其中 $\dagger$ 表示 Moore–Penrose 伪逆。另一个合法办法是固定一个节点电位后求解接地的正定系统，但不能混淆两种规范下的坐标值。若图不连通，则每个连通分量的右端项均须和为零，且下面的 $n - 1$ 应替换为 $n - c$ ， c 为分量数。

2.2 谱近似究竟近似了什么

本文主要使用下列约定， $0 < \varepsilon < 1$ ：

$$(1 - \varepsilon)L \preceq \widehat{L} \preceq (1 + \varepsilon)L \iff |x^\top (\widehat{L} - L)x| \leq \varepsilon x^\top L x \quad (\forall x). \quad (3)$$

它同时控制所有信号的二次能量，而不只是采样的一批训练特征。若 $L, \widehat{L}$ 都半正定，式(3)也强制两者具有相同零空间。在 $\mathbf{1}^\perp$ 上，它等价于

$$\left\| L^{\dagger/2} (\widehat{L} - L) L^{\dagger/2} \right\|_2 \leq \varepsilon. \quad (4)$$

因此“小”指的是相对图能量的误差，不是矩阵各元素小，也不是 $\|\widehat{L} - L\|_F$ 小。高条件数系统中，这个区别直接影响求解稳定性。

原文符号警告。 Kyng (2017, 第 2 章) 使用指数约定

$$M \approx_{\eta}^{\exp} N \iff e^{-\eta} M \preceq N \preceq e^{\eta} M. \quad (5)$$

而 Kyng and Sachdeva (2016) 的主要定理以 $1 \pm \varepsilon$ 表述。前者的误差参数可沿近似链相加；后者的上、下界分别相乘。式(3)蕴含指数误差 $\eta = -\log(1 - \varepsilon)$ ；反过来，取 $\eta = \log(1 + \varepsilon)$ 足以使指数近似蕴含式(3)。只能说小误差时二者同阶，不能不经转换直接复用精确常数。

2.3 归一化 Laplacian 与改变度矩阵的陷阱

无孤立节点时，归一化 Laplacian 是 $\mathcal{L} = D^{-1/2} L D^{-1/2}$ ，其零向量为 $D^{1/2} \mathbf{1}$ ，不是 $\mathbf{1}$ 。对式(3)施加同一个合同变换 $D^{-1/2}$ ，谱序仍成立；但是若近似图重新计算了 $\hat{D}$ ， $\hat{D}^{-1/2} \hat{L} \hat{D}^{-1/2}$ 不再是该合同变换的直接结果。度归一化变化需要另行分析。类似地， $I - D^{-1} A$ 的随机游走形式一般非对称，通常要通过相似变换回到对称空间，而不是直接套用欧氏范数结论。

3 精确消元：Schur 补为什么是正确的图接口

3.1 分块恒等式、变分解释与谐波延拓

把节点分成被消去的 F 和保留的 C ，按 (F, C) 排列。假设 C 非空且每个连通分量都接触 C ，则 $L_{FF} \succ 0$ 。定义

$$S = \text{Schur}_C(L) = L_{CC} - L_{CF} L_{FF}^{-1} L_{FC}. \quad (6)$$

精确分块分解为

$$L = \begin{bmatrix} I & 0 \\ L_{CF} L_{FF}^{-1} & I \end{bmatrix} \begin{bmatrix} L_{FF} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & L_{FF}^{-1} L_{FC} \\ 0 & I \end{bmatrix}. \quad (7)$$

Proposition 1 (Schur 补的能量最小化性质, [标准]) 对任意终端信号 $y \in \mathbb{R}^{|C|}$,

$$y^{\top} S y = \min_{z \in \mathbb{R}^{|F|}} \begin{bmatrix} z \\ y \end{bmatrix}^{\top} L \begin{bmatrix} z \\ y \end{bmatrix}, \quad z_* = -L_{FF}^{-1} L_{FC} y. \quad (8)$$

Proof 令 $r = z + L_{FF}^{-1} L_{FC} y$ ，展开得到原二次型为 $r^{\top} L_{FF} r + y^{\top} S y$ 。第一项非负且当且仅当 $r = 0$ 时为零，结论成立。 ■

令谐波延拓矩阵

$$Q = \begin{bmatrix} -L_{FF}^{-1} L_{FC} \\ I \end{bmatrix}, \quad Q^{\top} L Q = S. \quad (9)$$

因此消元不是忽略内部节点，而是将其置于与边界条件相容的最低能量状态。 S 仍是无向图 Laplacian；保留的是边界可见的等效相互作用，可能比原图稠密得多。这也是电网络 Kron reduction 的含义 (Dörfler and Bullo, 2013)。

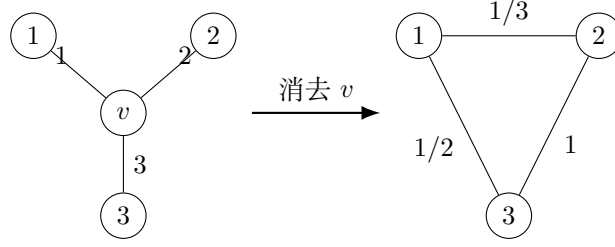


Figure 1: 精确星形-团变换。图为根据式(11)绘制的原创示意；每个终端信号的谐波延拓能量，以及任意两终端的有效电阻，都保持不变。

3.2 最值得记住的局部公式：星形变成加权团

消去节点 v ，令其不同邻居的总连接权重为 $a_1, \dots, a_r$ ， $d_v = \sum_i a_i$ 。该星形的 Laplacian 及产生的团为

$$L_{\text{star}(v)} = \begin{bmatrix} d_v & -a^\top \\ -a & \text{diag}(a) \end{bmatrix}, \quad (10)$$

$$C_v = \text{diag}(a) - \frac{aa^\top}{d_v} = \sum_{i < j} \frac{a_i a_j}{d_v} (e_i - e_j)(e_i - e_j)^\top. \quad (11)$$

新边权为 $a_i a_j / d_v$ 。一次精确消元可能创建 $\Theta(r^2)$ 个邻居对，这就是 fill-in 的源头。若把无序对求和改成全部有序对，必须补上 1/2，否则能量被多算一倍。

例如 $a = (1, 2, 3)^\top$ 时，

$$S = \begin{bmatrix} 5/6 & -1/3 & -1/2 \\ -1/3 & 4/3 & -1 \\ -1/2 & -1 & 3/2 \end{bmatrix}. \quad (12)$$

对终端信号 $y = (1, 0, -1)^\top$ ，内部最优值 $z_* = -1/3$ ，原星形最小能量和 $y^\top S y$ 都是 10/3。反之，直接取三个叶节点的诱导子图，会得到没有边的零矩阵，把这部分边界关系完全丢掉。

3.3 只保留 Schur 矩阵还不够：右端项与回代

若原任务是解 $Lx = b$ ，消元后的系统与回代是

$$Sx_C = g_C, \quad g_C = b_C - L_{CF} L_{FF}^{-1} b_F, \quad (13)$$

$$x_F = L_{FF}^{-1} (b_F - L_{FC} x_C). \quad (14)$$

当 $b \perp \mathbf{1}$ 时， $g_C \perp \mathbf{1}_C$ 。 x_C 的任意常量规范经回代形成一个原系统解，最后可以整体减去均值。图中有内部特征注入、源项或 bias 时，只替换邻接/Schur 矩阵而漏掉 g_C ，一般不会得到等价计算。这条原则在后面的多端口 DAG 消元里同样重要。

4 最可直接复用的三个谱结论

4.1 谱误差经过 Schur 补不会放大

Theorem 2 (Schur 补继承谱近似, [标准]) 设 $L, \hat{L}$ 满足式(3), 且相同划分的内部主块可逆。令 $S = \text{Schur}_C(L)$ 、 $\hat{S} = \text{Schur}_C(\hat{L})$, 则

$$(1 - \varepsilon)S \preceq \hat{S} \preceq (1 + \varepsilon)S. \quad (15)$$

Proof 对固定 y , 将式(3)作用于全部 $[z; y]$ 并对 z 取最小值。下界使用每个 z 均成立的不等式, 上界可代入 L 的最优延拓。由式(8)得到结论。 ■

这是“先稀疏化再消元”的重要理论接口。它不是说多层各自再引入一次 ε 误差后总误差仍为 ε : 若层层重新近似, 误差界要相乘; 指数约定下则相加。工程上必须区分误差的传递与新误差的注入。

4.2 逆算子、解误差与预条件数

Proposition 3 (逆的稳定性, [标准/推导]) 在共同零空间的正交补上, 式(3)蕴含

$$\frac{1}{1 + \varepsilon} L^\dagger \preceq \hat{L}^\dagger \preceq \frac{1}{1 - \varepsilon} L^\dagger. \quad (16)$$

对 $b \perp \mathbf{1}$, 令 $x = L^\dagger b$ 、 $\hat{x} = \hat{L}^\dagger b$, 则

$$\|\hat{x} - x\|_L \leq \frac{\varepsilon}{1 - \varepsilon} \|x\|_L. \quad (17)$$

Proof 限制到 $\mathbf{1}^\perp$ 并写 $\hat{L} = L^{1/2}(I + E)L^{1/2}$, 其中 $\|E\|_2 \leq \varepsilon$ 。矩阵 $I + E$ 的特征值位于 $[1 - \varepsilon, 1 + \varepsilon]$; 取逆即得式(16)。又 $\|(I + E)^{-1} - I\|_2 \leq \varepsilon/(1 - \varepsilon)$, 在能量坐标中得到式(17)。 ■

因子 $1/(1 - \varepsilon)$ 不能随意删掉: 当 $\hat{L} = (1 - \varepsilon)L$ 时, 非零 b 使该界取等号。以 $\hat{L}$ 作为 L 的预条件器, 相关的非零广义特征值条件数满足

$$\kappa \leq \frac{1 + \varepsilon}{1 - \varepsilon}, \quad \|e_k\|_L \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|e_0\|_L. \quad (18)$$

因此可以只构造常数精度的稀疏因子, 再迭代求解原方程至高精度, 而不必把因子本身构造造成最终求解容差。若直接要求分解误差趋近零, ε^{-2} 的代价可能很高。处理奇异因子时, 要采用接地系统或兼容子空间的投影/规范; 一般不能把普通逆的合同公式机械替换成伪逆公式。

4.3 有效电阻与 leverage score: 什么边是谱意义上的重要边

定义

$$R_{uv} = (e_u - e_v)^\top L^\dagger (e_u - e_v), \quad \tau_e = w_e R_{uv}. \quad (19)$$

R_{uv} 衡量两节点之间的等效电阻, τ_e 是边相对于整个图的 leverage score。两个基本恒等/不等式为

$$0 < \tau_e \leq 1, \quad \sum_e \tau_e = n - 1. \quad (20)$$

第一式来自 $w_e b_e b_e^\top \preceq L$ 的白化；第二式来自 $\sum_e w_e b_e^\top L^\dagger b_e = \text{tr}(L^\dagger L) = n - 1$ 。桥边的 leverage 为 1，不因它的权重小就不重要；相反，存在许多替代路径的边可能具有较低 leverage。这个分数表示图几何的相对重要性，不等价于某个监督任务的特征归因分数。

对终端 $u, v \in C$ ，精确 Schur 补满足

$$(e_u - e_v)^\top L^\dagger (e_u - e_v) = (e_u^C - e_v^C)^\top S^\dagger (e_u^C - e_v^C). \quad (21)$$

可用单位电流注入和式(13)证明，因为此时 $b_F = 0$ ，终端电位差不变。它并不意味着 $(L^\dagger)_{CC} = S^\dagger$ ；准确形式需在终端上中心化： $\Pi_C (L^\dagger)_{CC} \Pi_C = S^\dagger$ ，其中 $\Pi_C = I - \mathbf{1}_C \mathbf{1}_C^\top / |C|$ 。若再做谱近似，电阻按式(16)的倒数因子变化，而不是机械写成同一个 $1 \pm \varepsilon$ 。

Spielman and Srivastava (2011) 的全局谱稀疏化把有效电阻变成采样分布：独立采样 q 次，令 $p_e = \tau_e / (n - 1)$ ，每次将抽中的边赋权 $w_e / (qp_e)$ 。于是

$$\mathbb{E} \hat{L} = L, \quad q = O\left(\frac{n \log(n/\delta)}{\varepsilon^2}\right) \implies (3) \text{ 以至少 } 1 - \delta \text{ 的概率成立}. \quad (22)$$

这里抑制绝对常数；使用足够精确的电阻上界也可相应调整采样量。计算电阻本身不是免费的，不能在复杂度论证中先用稠密伪逆算出全部电阻，再宣称整个流程近线性。Kyng-Sachdeva 的局部算法重要之处之一，正是避免把每一步都变成完整的电阻计算。

4.4 行列式、生成树与图结构学习中的正则项

Laplacian 的奇异性并不妨碍定义其非零特征值乘积 $\det_+(L)$ 。矩阵树定理将它与加权生成树配分函数联系起来 (Kyng, 2017, 第 2.4、4 章)：

$$Z_{\text{tree}}(w) = \sum_{T \text{ 为生成树}} \prod_{e \in T} w_e, \quad \det_+(L) = n Z_{\text{tree}}(w). \quad (23)$$

因此当图保持连通、支撑固定且边权为正时，

$$\frac{\partial \log \det_+(L)}{\partial w_e} = R_e, \quad \frac{\partial \log Z_{\text{tree}}}{\partial \log w_e} = w_e R_e = \mathbb{P}_T(e \in T). \quad (24)$$

这里 $\mathbb{P}_T(T) \propto \prod_{e \in T} w_e$ 。第一式可对 $\log \det(L + \mathbf{1}\mathbf{1}^\top/n)$ 微分证明；第二式直接对树配分函数微分。由此，leverage 同时是谱采样分数和随机生成树的边缘概率。但这些边缘概率不意味着树内各边独立；独立按同样边缘概率采边，通常不是生成树分布。

这给 graph structure learning 一个不同于消息传递的应用。若中心化信号 $x^{(j)} \in \mathbf{1}^\perp$ 由精度矩阵为 L 的高斯模型描述，平均负对数似然（略去常数）为

$$\mathcal{J}(w) = \frac{1}{2N} \sum_{j=1}^N (x^{(j)})^\top L x^{(j)} - \frac{1}{2} \log \det_+(L) + \lambda \Omega(w). \quad (25)$$

[推导] 对可微正则 Ω ，其边权梯度等于经验边差平方均值的一半，减去 $R_e/2$ ，再加正则梯度。只最小化平滑能量容易鼓励权重缩小；log-determinant 项提供概率模型所需的归一化约束。若允许图断开，则零空间维数变化，不能不修改模型便沿用上述似然和导数。

精确消元还可分解行列式。选根节点 $r \in C$ 并接地，便有

$$\det L_{V \setminus \{r\}, V \setminus \{r\}} = \det L_{FF} \det S_{C \setminus \{r\}, C \setminus \{r\}}. \quad (26)$$

不过从谱近似估计 log-determinant 时必须注意维度：式(3)仅直接给出

$$|\log \det_+(\widehat{L}) - \log \det_+(L)| \leq -(n-1) \log(1-\varepsilon). \quad (27)$$

这是广义特征值逐个取对数后的求和界。常数级谱精度并不自动对应维度无关的小似然误差；若论文主张高精度配分函数或图生成概率，还需要专门的误差预算，而不是只报告谱稀疏度。

5 Kyng–Sachdeva：用稀疏随机团替代稠密 fill-in

5.1 先理解完整分解，再理解随机部分

对当前 Laplacian S_t 消去节点 v ，令 $d = (S_t)_{vv}$ 、 $c = S_t[:, v]/d$ 。当 $d > 0$ 时，记录一个秩一因子 $dcc^\top$ ，精确剩余矩阵是 $S_t - dcc^\top$ ；这等价于删去 v 的星形，再加上式(11)的团。对连通图而言最后一个节点的对角为零，应单独处理，不做除零操作。Kyng and Sachdeva (2016) 的随机步骤只有一个原则：用 $\widehat{C}_v$ 替代 C_v ，但保留精确抽取的当前列因子。

Theorem 4 (近似稀疏 Cholesky, [原文]) 设 L 是 n 个节点、 m 条边的连通正权无向多重图 Laplacian。对 $0 < \varepsilon \leq 1/2$ 和失败指数 $a > 1$ ，Kyng and Sachdeva (2016, Theorem 3.1) 的算法返回置换 P 、下三角矩阵 R 、非负对角矩阵 D ，使

$$\mathbb{P} \left[(1-\varepsilon)L \preceq PRDR^\top P^\top \preceq (1+\varepsilon)L \right] \geq 1 - 2n^{-a}. \quad (28)$$

其运行时间和 R 的非零元数的期望均为 $O(a^2 \varepsilon^{-2} m \log^3 n)$ 。

原文用 $\delta > 1$ 表示这里的 a ；本文把 $\delta \in (0, 1)$ 留给通常的失败概率，以免误读。该定理保证的是因子稀疏，不意味着把乘积 $PRDR^\top P^\top$ 显式展开后仍有相同的非零元数。求解器应用因子和三角求解，不能展开稠密乘积后再讨论加速。

5.2 原始 CliqueSample 的准确分布与无偏性证明

设当前节点 v 关联 r 条多重边，边 i 连接邻居 u_i ，权为 w_i ；此处 r 是边条数， $d = \sum_i w_i$ 是加权重，两者不能混用。执行 r 次独立采样，每次先以 w_i/d 抽取第一条边，再以 $1/r$ 均匀抽取第二条边 j 。若 $u_i \neq u_j$ ，输出一条连接两邻居的边，权为

$$h_{ij} = \frac{w_i w_j}{w_i + w_j}; \quad \text{若 } u_i = u_j \text{ 则输出零贡献.} \quad (29)$$

注意它不是直接以 $w_i w_j$ 抽一对再照抄精确团边权。对一对不同的边索引 i, j ，一次采样抽到无序对的概率是 $(w_i + w_j)/(dr)$ ，因此其 r 次总期望权为

$$r \cdot \frac{w_i + w_j}{dr} \cdot \frac{w_i w_j}{w_i + w_j} = \frac{w_i w_j}{d}. \quad (30)$$

将连接相同邻居的多重边贡献合并，就得到精确团；所以 $\mathbb{E}[\widehat{C}_v \mid S_t, v] = C_v$ 。输出至多 r 条边，局部采样可在线性工作量内完成。算法真正解决了“生成 r^2 个候选再稀疏化”这个潜在瓶颈，而不是只把最终存储量变小。

Algorithm 1 KS16 近似消元的结构性伪代码（不是生产级求解器）**输入：** 连通正权无向图 G ，精度 $\varepsilon \leq 1/2$ ，失败指数 $a > 1$ **输出：** 置换及近似 Cholesky 因子

- 1: $\rho \leftarrow \lceil 12(1+a)^2 \varepsilon^{-2} \ln^2 n \rceil$
- 2: 每条权重 w 的边拆成 ρ 条权重 w/ρ 的平行边
- 3: 选取节点的均匀随机排列；令当前图矩阵 $S \leftarrow L$
- 4: **for** 排列中除最后一个以外的节点 v **do**
- 5: $d \leftarrow S_{vv}$ ；若 $d = 0$ ，记录零对角并跳过该节点
- 6: 记录 $c \leftarrow S[:,v]/d$ 及 d
- 7: 用式(29)的加权/均匀双抽样构造 $\hat{C}_v$
- 8: $S \leftarrow S - L_{\text{star}(v)} + \hat{C}_v$ ，删除已消去坐标
- 9: **end for**
- 10: 将记录的列、对角与置换组织为 P, R, D ；处理零对角终点

5.3 为何要拆边，以及为何消元顺序是定理的一部分

拆成平行边不改变 L ，但降低每一次随机贡献的相对尺度。固定原始图的有效电阻，把一条当前边的尺度写为 wR_{uv}^L 。拆分后的初始边均满足该值不超过 $1/\rho$ 。如果两条当前星形边也满足这个界，则由有效电阻的三角不等式，

$$h_{ij}R_{u_i u_j}^L \leq h_{ij}(R_{u_i v}^L + R_{v u_j}^L) \leq \frac{w_i w_j}{w_i + w_j} \left(\frac{1}{\rho w_i} + \frac{1}{\rho w_j} \right) = \frac{1}{\rho}. \quad (31)$$

这揭示了调和型权重的第二个作用：它不仅给出正确的期望，还使新生成的边继续满足小 leverage 上界。**这是 Laplacian 特有的几何闭包，不是对任意矩阵秩一项都成立的事实。**

均匀随机消元顺序用于控制剩余图上被处理的平均度、累积工作量和方差。实际实现常偏好低度消元，但不能把定理4中的随机排列直接换成任意确定性的最小度规则，然后沿用原来的概率与时间结论。类似地，原始局部采样的输出不必每次都连接全部邻居；高概率的全局谱结论与逐步确定性的连通性保证不同。后续实用算法在采样与数据结构上作了专门设计，不能与原始算法视为逐行相同 (Gao et al., 2023)。

6 为什么局部无偏能够累积成全局谱保证**6.1 关键转换：不是乘一串近似比，而是追踪加性误差**

把已消去坐标补零后，令 $\hat{S}_k$ 表示剩余图， $d_t c_t c_t^\top$ 表示之前记录的因子。定义

$$Z_k = \sum_{t \leq k} d_t c_t c_t^\top + \hat{S}_k, \quad Z_0 = L. \quad (32)$$

一次操作精确记录当前消元列，仅把团替换成随机团，所以

$$Z_k - Z_{k-1} = \hat{C}_k - C_k, \quad \mathbb{E}[Z_k - Z_{k-1} \mid \text{过去历史及当前主元}] = 0. \quad (33)$$

尽管后续团取决于过去的随机结果，误差仍构成适应性矩阵鞅。把“每一步是独立噪声”作为证明前提是不对的；正确的是给定过去后的条件均值为零。

用固定的原始 L 白化： $Y_k = L^{\dagger/2}(Z_k - L)L^{\dagger/2}$ 。目标是 $\|Y_k\|_2 \leq \varepsilon$ ，而不是每步在不断变化的 $\hat{S}_k$ 度量下单独证明小误差后未经转换就相加。由于式(31)，单次采样边以及其条件期望的白化矩阵都位于 $[0, I/\rho]$ ，相应中心化增量的算子范数不超过 $1/\rho$ 。

6.2 矩阵 Freedman 不等式及其使用条件

对自伴鞅差 X_t , 假设 $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$ 且 $\|X_t\|_2 \leq R$. 令

$$Y_k = \sum_{t \leq k} X_t, \quad V_k = \sum_{t \leq k} \mathbb{E}[X_t^2 | \mathcal{F}_{t-1}]. \quad (34)$$

Tropp (2011) 的矩阵 Freedman 不等式经正、负两尾联合, 可写成

$$\mathbb{P}(\exists k : \|Y_k\|_2 \geq t, \|V_k\|_2 \leq \sigma^2) \leq 2d \exp\left(-\frac{t^2}{2(\sigma^2 + Rt/3)}\right), \quad (35)$$

其中 d 是矩阵维数。它需要控制可预测二次变差, 不是只控制终点噪声的期望或逐步最大值。对于每个主元下的独立边样本, 可利用 $X^2 \preceq \|X\|_2 X$ (对 PSD 的未中心化样本) 把条件方差控制为团矩阵的倍数。

原始分析还利用: 在剩余 r_k 个节点上均匀选择主元时, 团不超过对应星形, 而所有星形之和为当前 Laplacian 的两倍, 因此

$$\mathbb{E}_v[C_v | \hat{S}_k] \preceq \frac{2}{r_k} \hat{S}_k. \quad (36)$$

调和级数 $\sum_k 1/r_k = O(\log n)$ 随之出现。在尚未离开谱误差良好事件时, $\hat{S}_k \preceq Z_k \preceq (1+\varepsilon)L$, 于是白化后的期望方差具有约 $\log n / \rho$ 的尺度。

这一步还不是完整证明。对方差的期望上界不能直接替代式(35)要求的路径级高概率界; 而“尚未离开良好事件”本身又涉及待证明的量。KS16 用停止过程及其适应性采样分析处理这些相互依赖, 包括对方差累积的进一步控制。本节只给出结构性链条, 精确常数与尾概率应回到原文证明, 而不能凭上面两行就宣称完成了定理4。

6.3 对图学习证明工作的启发与反例

可以迁移的证明习惯是: 先找到精确的望远镜差分, 再设计小增量和可控制的条件方差, 最后选择与任务相符的度量。不能仅凭 $\mathbb{E}\hat{L} = L$ 就说求解无偏, 因为一般

$$\mathbb{E}[(\mu I + \hat{L})^{-1} X] \neq (\mu I + L)^{-1} X. \quad (37)$$

仅在一维, 正随机数的倒数就已经是非线性的。更不能据此推出带非线性激活、多层递归、损失函数或随机训练过程的输出无偏。局部无偏是概率证明的输入条件, 不是所有下游任务的最终保证。

7 图缩减与 graph learning: 保留的是哪一种信息

7.1 稀疏 Schur 补, 而不是任意图池化

Kyng (2017, Theorem 1.2.5; 第 3.4 节) 给出稀疏近似 Schur 补构造。用其指数误差约定, 对保留集 C 、 $k = |C|$ 、 $0 < \eta \leq 1/2$, 可得到 $\hat{S} \approx_\eta^{\text{exp}} \text{Schur}_C(L)$, 失败概率为 $O(\delta)$, 且

$$\text{nnz}(\hat{S}) = O(k\eta^{-2} \log(n/\delta)), \quad \text{构造时间 } \tilde{O}(m + n\eta^{-2}). \quad (38)$$

这里 $\tilde{O}$ 隐藏关于 n 、失败概率及相应低阶对数的因子; 输出随 k 缩小, 并不意味着原定理的构造代价可写成 $\tilde{O}(m + k\eta^{-2})$ 。单纯提前停止近似消元, 也不自动得到该尺寸的边界矩阵; 稀疏 Schur 构造还要控制并稀疏化剩余结构。

对学习任务，选择终端集 C 与在固定 C 上保能量是两个不同问题。定理对给定终端的算子质量负责，不会自动告诉我们哪些节点最适合作为语义代表、读出位置或监督锚点。把终端选择做成可学习决策，需要补充任务目标、离散选择的训练方法和外推评估。

7.2 谐波半监督学习与仅终端损失的精确约化

当终端标签/表示固定为 Y_C ，最小化图 Dirichlet 能量得到

$$H_F^* = -L_{FF}^{-1}L_{FC}Y_C, \quad \min_{H_F} \frac{1}{2} \text{tr}(H^\top L H) = \frac{1}{2} \text{tr}(Y_C^\top S Y_C). \quad (39)$$

这是经典谐波半监督学习的线性代数核心 (Zhu et al., 2003)。更一般地，若损失只依赖终端，

$$\min_{H_F, Y_C} \ell_C(Y_C) + \frac{\lambda}{2} \text{tr}(H^\top L H) = \min_{Y_C} \ell_C(Y_C) + \frac{\lambda}{2} \text{tr}(Y_C^\top S Y_C). \quad (40)$$

这给出一个可靠的 graph-learning 研究接口：在边界损失不变的条件下压缩正则项。用谱近似的 $\hat{S}$ 替代 S ，可逐点控制正则能量；但最优参数或预测误差还需要损失曲率、可辨识性等假设。若内部节点上也有监督、重建损失、非线性模块或不可消去的特征，式(40)就不再描述完整目标。

7.3 Schur、诱导子图和 Galerkin 投影不是同一个操作

诱导子图只保留原来的终端间边；Schur 补加入内部路径产生的等效边；一般的粗化采用延拓矩阵 P ，构造 $P^\top L P$ ，只对 $\text{range}(P)$ 中的信号解释为原能量。Schur 对应的是特定谐波延拓 Q ，并且在固定终端值时对内部自由度取能量最小值，而不是任意聚类分配矩阵。Loukas (2019) 的谱/割保证研究也提示：降低维度以后，应明确所保留的子空间与误差定义，而不是笼统宣称“整个谱被保留”。

即使 $Q^\top L Q = S$ ，原信号的欧氏质量也变为 $\|Qy\|_2^2 = y^\top Q^\top Q y$ 。因此保留能量不等于保持普通特征值问题的归一化；相应的受限 Rayleigh 商是 $y^\top S y / (y^\top Q^\top Q y)$ 。关于低频特征子空间的进一步保证，还需要适当的子空间覆盖、质量矩阵与谱间隙条件。

7.4 一个特别容易误用的等式：加质量项与消元不交换

很多传播层使用 $\mu I + \tau L$ ，其精确约化应该直接消元这个完整算子。通常

$$\text{Schur}_C(\mu I + L) \neq \mu I_C + \text{Schur}_C(L). \quad (41)$$

考虑单位边权三节点路径，保留两端并令 $\mu = 1$ ，直接计算得到

$$\text{Schur}_C(I + L) = \begin{bmatrix} 5/3 & -1/3 \\ -1/3 & 5/3 \end{bmatrix}, \quad I_C + \text{Schur}_C(L) = \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix}. \quad (42)$$

两者之差为全 $1/6$ 矩阵。另有恒等式 $Q^\top(\mu I + L)Q = \mu Q^\top Q + S$ ，但这个固定于 L 的 Q 也一般不是 $\mu I + L$ 的能量最优延拓。对扩散、正则化和隐式 GNN，这意味着“先压缩 Laplacian，之后随便补一个单位阵”并不自动等价于原模型。

8 隐式传播、稳定性与可微求解：可以直接用的推导

8.1 正则化 resolvent 的误差界

设 $A = \mu I + \tau L$, $\mu > 0$, $\tau \geq 0$, 并考虑多右端项传播

$$AH = X, \quad \hat{A} = \mu I + \tau \hat{L}, \quad \hat{A}\hat{H} = X. \quad (43)$$

这是一个定义清楚的线性层；若原损失采用 $\mu \|H - X\|_F^2 / 2$, 相应右端项应改为 μX 。这里固定 A, X , 不把训练中的变化偷偷放进同一个误差符号。

Proposition 5 (隐式传播稳定性, [推导]) 若式(3)成立, 则

$$\|\hat{H} - H\|_{A,F} \leq \frac{\varepsilon}{1 - \varepsilon} \|H\|_{A,F}, \quad (44)$$

$$\|\hat{H} - H\|_F \leq \frac{\varepsilon}{\mu(1 - \varepsilon)} \|X\|_F. \quad (45)$$

Proof 因 $\tau L \preceq A$, 有 $(1 - \varepsilon)A \preceq \hat{A} \preceq (1 + \varepsilon)A$ 。将式(17)的正定版本逐列应用得到第一式。再用 $A \succeq \mu I$ 及 $\|A^{-1}X\|_{A,F} \leq \|X\|_F / \sqrt{\mu}$ 得到第二式。 ■

若损失对 H 是 K -Lipschitz, 损失差不超过式(45)右端乘以 K 。这是固定模型和输入的前向误差界, 不是泛化界, 也没有证明近似训练与原训练最终找到同一个参数。更精细的分析可利用

$$\hat{A}^{-1} - A^{-1} = -\hat{A}^{-1}\tau(\hat{L} - L)A^{-1} \quad (46)$$

考察输入落在哪些图频率方向上, 而不是仅报告最坏情形上界。

对多项式滤波 $p(L) = \sum_{j=0}^q a_j L^j$, 一个直接的欧氏算子界是

$$\|p(\hat{L}) - p(L)\|_2 \leq \|\hat{L} - L\|_2 \sum_{j=1}^q |a_j| j M^{j-1}, \quad M = \max\{\|L\|_2, \|\hat{L}\|_2\}. \quad (47)$$

它来自矩阵幂差的非交换望远镜展开。高阶多项式、较大系数或多层 Lipschitz 常数可放大误差；谱近似并不凭空消除这些因素。

8.2 边权梯度：两个线性求解就能写出精确形式

令 $G = \nabla_H \ell(H)$, 伴随变量 U 满足 $AU = G$ 。固定 μ, τ 时, 对输入与图的微分有

$$dH = A^{-1}dX - \tau A^{-1}(dL)H. \quad (48)$$

由于 $\partial L / \partial w_e = b_e b_e^\top$,

$$\frac{\partial \ell}{\partial w_e} = -\tau \langle b_e^\top U, b_e^\top H \rangle = -\tau \langle U_u - U_v, H_u - H_v \rangle. \quad (49)$$

该式表明边权梯度取决于前向信号差与伴随信号差的对齐, 而不是仅取决于边权或有效电阻。如果用 $w_e = \text{softplus}(\theta_e)$ 保持正权, 再乘以 $\text{sigmoid}(\theta_e)$ 。若 X, μ, τ 或损失显式依赖边权, 必须额外加入对应链式项。

8.3 前向与反向的求解容差应该一起预算

设 $\bar{H}, \bar{U}$ 是近似解, 残差 $r = X - A\bar{H}$ 、 $s = G - A\bar{U}$ 。对任意矩阵右端项,

$$\|\bar{H} - H\|_{A,F} = \|r\|_{A^{-1},F}, \quad \|\bar{U} - U\|_{A,F} = \|s\|_{A^{-1},F}. \quad (50)$$

定义 δ_H, δ_U 为上述两误差, $R_e^A = b_e^\top A^{-1} b_e$ 。由能量 Cauchy-Schwarz 不等式可得

$$|\bar{g}_e - g_e| \leq \tau R_e^A \left(\|U\|_{A,F} \delta_H + \|H\|_{A,F} \delta_U + \delta_H \delta_U \right), \quad (51)$$

其中 g_e 由式(49)定义。证明只需展开两个内积之差, 再使用 $\|b_e^\top Z\|_2 \leq \sqrt{R_e^A} \|Z\|_{A,F}$ 。 R_e^A 是正则化算子的对应二次型, 不应与无质量项的 R_e 混同。

式(50)的反向右端项是精确 $G = \nabla \ell(H)$ 。实际代码常用 $\bar{G} = \nabla \ell(\bar{H})$, 这时还需计入 $\|\bar{G} - G\|_{A^{-1},F}$; 若梯度为 L_g -Lipschitz, 此项不超过 $L_g \|\bar{H} - H\|_F / \sqrt{\mu}$ 。只报告前向残差, 而不检查反向求解及损失梯度的变化, 不能宣称梯度已经受控。

8.4 谱近似不推出参数导数近似

令原图与参数无关, $L_\theta = L_0$, 而近似图为

$$\hat{L}_\theta = (1 + \varepsilon \sin(\omega\theta)) L_0. \quad (52)$$

它对所有 θ 都满足 $1 \pm \varepsilon$ 谱近似, 但 $\partial_\theta \hat{L}_\theta = \varepsilon \omega \cos(\omega\theta) L_0$, 可随 ω 任意大。在 $\theta = 0$, 两图甚至完全相同, 导数却不同。因此**即使谱保证对所有参数一致成立, 也不能单凭它推出训练梯度一致。**

如果采样概率 $p_\theta(\xi)$ 随参数变化, 期望目标的微分还包括分布项:

$$\nabla_\theta \mathbb{E}_{\xi \sim p_\theta} f_\theta(\xi) = \mathbb{E}[\nabla_\theta f_\theta(\xi) + f_\theta(\xi) \nabla_\theta \log p_\theta(\xi)], \quad (53)$$

在允许交换微分与期望的条件下成立。忽略第二项不自动得到无偏梯度。较稳妥的实现路线是: 把随机近似因子作为原算子 A 的预条件器, 前向/伴随都迭代求解原系统, 隐式微分原方程。这样不需要把预条件器误当成模型定义; 仍须记录构造成本、有限容差与图更新后的有效性。

9 与 DAG Graph DeltaNet 的真正接口: 多端口、路径和与误差度量

9.1 先固定递归语义, 不把问题改写成静态回归

以下形式用来说明消元机制, 不声称代替当前项目中尚待固定的全部架构细节。对 DAG 节点 v , 设状态 $S_v \in \mathbb{R}^{d_v \times d_k}$, 多前驱合并及 Delta 更新为

$$M_v = \sum_{u \in \text{pred}(v)} \alpha_{uv} S_u, \quad (54)$$

$$S_v = M_v + \beta_v (v_v - M_v k_v) k_v^\top = M_v A_v + \beta_v v_v k_v^\top, \quad A_v = I - \beta_v k_v k_v^\top. \quad (55)$$

这里 v_v 是 value 向量, 不是节点索引; 假设一次前向中 α, k, v, β 由输入预先确定, 而不依赖尚未计算的递归状态。若门控依赖 S_u , 整体系统一般非线性, 下面的固定线性消元就不能原样成立。讨论精确并行化时, 这个假设必须出现在方法章节, 而不能留在实现细节中。

使用按列向量化 $s_v = \text{vec}(S_v)$, 相同状态宽度下有

$$\mathcal{K}_{vu} = \alpha_{uv}(A_v^\top \otimes I_{d_v}), \quad b_v = \text{vec}(\beta_v v_v k_v^\top), \quad \mathcal{T}s = b, \quad \mathcal{T} = I - \mathcal{K}. \quad (56)$$

在拓扑排序后, $\mathcal{K}$ 为严格块下三角。若最长有向路径有 D 条边, $\mathcal{K}^{D+1} = 0$, 因此

$$(I - \mathcal{K})^{-1} = I + \mathcal{K} + \dots + \mathcal{K}^D. \quad (57)$$

每个块是沿有向路径按顺序相乘的矩阵贡献之和。它是原递归的精确重写, 不是把 Delta 更新替换成 Graph Ridge, 也不需要引入一个新的静态最小二乘训练目标。

9.2 非交换性与拓扑排序: 顺序不可以随便交换

即使 A_v 都对称, 也不意味着相互交换。例如

$$A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_2 = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad A_1 A_2 \neq A_2 A_1. \quad (58)$$

两者都可由单位 key、 $\beta = 1$ 的 Delta 因子产生。因此“并行”应该意味着重组计算并保留有向路径上的有序乘积, 而不是把各门控当成可交换标量。序列 DeltaNet 的 WY/chunkwise 表达说明低秩更新可被代数重排为适合矩阵运算的形式 (Yang et al., 2024); 多前驱图还要显式表示合流的多端口依赖, 不能只把序列下标改成节点下标。

若节点重编号对应块置换 $\mathcal{P}$, 有 $\mathcal{K}' = \mathcal{P}\mathcal{K}\mathcal{P}^\top$ 、 $b' = \mathcal{P}b$, 则 $s' = \mathcal{P}s$ 。这给出固定线性递归的置换等变性。不同合法拓扑调度计算同一个方程, 不改变路径语义; 但用原始节点编号挑 chunk、裁剪路径或分配随机样本, 可能破坏逐样本等变性。随机算法需区分固定耦合随机性下的等变与仅分布意义的等变。

9.3 块消元给出精确多端口传输, 而非 PSD 能量最小化

将一组内部状态记为 F , 外部/边界状态记为 C , 一般的非对称消元仍然成立:

$$\begin{aligned} \mathcal{T}_{\text{eff}} &= \mathcal{T}_{CC} - \mathcal{T}_{CF}\mathcal{T}_{FF}^{-1}\mathcal{T}_{FC} \\ &= I - \mathcal{K}_{CC} - \mathcal{K}_{CF}(I - \mathcal{K}_{FF})^{-1}\mathcal{K}_{FC}, \end{aligned} \quad (59)$$

$$b_{\text{eff}} = b_C - \mathcal{T}_{CF}\mathcal{T}_{FF}^{-1}b_F. \quad (60)$$

$\mathcal{T}_{FF}$ 可逆, 因为诱导子图仍无环; 式(59)保留经过内部路径的全部有序传输。但一般没有对称性、Laplacian 符号结构或式(8)的能量最小化解释。

对一个 chunk, 把输入端口状态堆为 p , 内部满足 $\mathcal{T}_{FF}s_F = b_F + \mathcal{K}_{FF}p$, 输出定义为 $y = D_C p + C_C s_F$ 。则

$$y = \underbrace{(D_C + C_C \mathcal{T}_{FF}^{-1} \mathcal{K}_{FF})}_{T_c} p + \underbrace{C_C \mathcal{T}_{FF}^{-1} b_F}_{g_c}. \quad (61)$$

这就是最值得从消元视角借用的接口: 一个多端口仿射算子 (T_c, g_c) , 而不是单一递归状态或只有传输矩阵的接口。Graph-UT / Multi-port WY 可以寻求它的精确结构化表示和高效应用; 不必先引入任何随机近似。若当前模型还需要 chunk 内节点的输出或梯度, 应保留回代/重计算所需信息, 而不是认为端口映射已经包含所有读出。



$$T_c = D_c + C_c \mathcal{T}_{FF}^{-1} \mathcal{K}_{FP}, \quad g_c = C_c \mathcal{T}_{FF}^{-1} b_F$$

精确分块先保语义；是否压缩接口，是后续独立选择。

Figure 2: 面向多前驱 DAG 的消元接口。与 Laplacian 消元共享分块恒等式，但不共享半正定能量保证。图为本文根据式(61)绘制。

9.4 非对称系统应使用什么误差界

设近似消元最终得到 $\hat{\mathcal{T}} = \mathcal{T} + E$ ，原系统解 $s = \mathcal{T}^{-1}b$ 。对某个一致的诱导范数，若

$$\eta = \|\mathcal{T}^{-1}E\| < 1, \quad (62)$$

则 Neumann 展开给出

$$\|\hat{s} - s\| \leq \frac{\eta}{1 - \eta} \|s\| + \frac{\|\mathcal{T}^{-1}\Delta b\|}{1 - \eta}, \quad (\mathcal{T} + E)\hat{s} = b + \Delta b. \quad (63)$$

这是一个可直接使用的 [推导]，但它依赖 $\mathcal{T}^{-1}$ 的放大作用，不是 Laplacian 的谱序结论。局部端口算子误差小，仍可能被下游传播放大。 $\mathcal{T}$ 的全部特征值都是 1 也无济于事：非正规下三角矩阵的奇异值条件数与逆范数可以很大。

若 $0 \leq \beta_v \|k_v\|_2^2 \leq 2$ ，则 $\|A_v\|_2 \leq 1$ 。但与 k_v 正交的方向仍有特征值 1，故不保证严格收缩。若 $\alpha_{uv} \geq 0$ 且每行权重和不超过 1，在块最大范数下，合并-门控不放大单步状态差，但每步新误差仍可沿深度累积。若进一步有严格上界 $\gamma < 1$ ，则 $\|(I - \mathcal{K})^{-1}\| \leq 1/(1 - \gamma)$ ；这是额外的架构稳定性假设，也可能引入遗忘，不应为了得到漂亮界而悄悄改变模型语义。

把方程乘成 $\mathcal{T}^\top \mathcal{T}s = \mathcal{T}^\top b$ 虽能产生 SPD 正规方程，却通常不产生 Laplacian 或对角占优矩阵，而且会平方条件数并可能增加 fill-in。因而“先做正规方程再调用 Kyng”不是无成本的理论桥梁，也不是本文建议替代当前递归主线的方法。

9.5 若将来要近似端口：矩形矩阵采样的正确起点

[设想] 假设已经得到某个固定端口算子的可枚举分解 $T_c = \sum_{i=1}^N B_i$ 。独立按 $p_i > 0$ 抽取 q 个索引，可定义

$$\hat{T}_c = \frac{1}{q} \sum_{j=1}^q \frac{B_{I_j}}{p_{I_j}}, \quad \mathbb{E} \hat{T}_c = T_c. \quad (64)$$

这里的 B_i 不需要对称或 PSD。对矩形误差使用自伴扩张

$$\mathcal{D}(E) = \begin{bmatrix} 0 & E \\ E^\top & 0 \end{bmatrix}, \quad \|\mathcal{D}(E)\|_2 = \|E\|_2, \quad (65)$$

便可以应用矩阵集中不等式 (Tropp, 2011)。例如独立样本时，可取增量界与方差参数

$$R = \frac{1}{q} \left(\max_i \frac{\|B_i\|_2}{p_i} + \|T_c\|_2 \right),$$

$$\sigma^2 = \frac{1}{q} \max \left\{ \left\| \sum_i \frac{B_i B_i^\top}{p_i} - T_c T_c^\top \right\|_2, \left\| \sum_i \frac{B_i^\top B_i}{p_i} - T_c^\top T_c \right\|_2 \right\}, \quad (66)$$

$$\mathbb{P}(\|\hat{T}_c - T_c\|_2 \geq t) \leq (d_{\text{out}} + d_{\text{in}}) \exp \left[-\frac{t^2}{2(\sigma^2 + Rt/3)} \right]. \quad (67)$$

这些是固定矩阵分解的标准概率工具，不是已经得到的 Graph DeltaNet 高效算法。真正需要解决的是： B_i 是否能廉价访问、其数量是否指数增长、采样权重如何估计、 g_c 如何共同处理、局部误差如何满足式(62)，以及学习门控变化后是否还保持保证。若为了采样显式枚举所有 DAG 路径，已经失去了分块方法要解决的计算优势。

10 推广到有向图和向量边：哪些桥存在，哪些桥不存在

10.1 Eulerian 有向 Laplacian 与一般 DAG 不是同一个类

采用 $A_{ij} = w(i \rightarrow j)$ 的约定，令 $L_{\text{dir}} = D_{\text{out}} - A^\top$ ，其列和为零。Eulerian 平衡进一步要求每个节点入度权重等于出度权重，从而 $L_{\text{dir}} \mathbf{1} = 0$ 。在适当连通条件下， $U = (L_{\text{dir}} + L_{\text{dir}}^\top)/2$ 是对称 Laplacian，可用如下非对称误差量：

$$\left\| U^{\dagger/2} (\hat{L}_{\text{dir}} - L_{\text{dir}}) U^{\dagger/2} \right\|_2 \leq \varepsilon. \quad (68)$$

同时必须满足相应的左右零空间兼容条件。只控制二次型是不够的，因为 $x^\top L_{\text{dir}} x = x^\top U x$ 看不到反对称部分；同一个三节点有向环及其反向环就有相同的 U ，却有不同方向传输。

Kyng 博士论文第 5 章与后续稀疏 LU 工作采用更精细的有向消元、平衡性与内部块条件来控制误差 (Kyng, 2017; Cohen et al., 2018)。完整分解分析还可能使用构造的辅助度量 F ，不能把最终 LU 保证简单地替换为无向的 $1 \pm \varepsilon$ Loewner 序。对有向图，消去节点生成的是入邻居到出邻居的有向相互作用；维护入/出平衡本身就是算法约束。

Proposition 6 (无非平凡正权 Eulerian DAG, [标准]) 有限无环有向图若每个节点的加权入度等于加权出度，且边权均为正，则图中没有边。

Proof DAG 存在源点，其入度为零，平衡性迫使出度权重也为零。删除该孤立源点后重复，最终所有边均不存在。 ■

所以当前多前驱 DAG 不能仅因为“也是 directed graph”，就直接适用 Eulerian Laplacian 求解定理。若做强连通扩充或添加反向边，必须重新检查是否改变了原递归语义。

10.2 Connection Laplacian 是更接近矩阵消息的特殊结构

对每条无向边给定正交传输 $O_{ij} \in \mathbb{R}^{r \times r}$ ，满足 $O_{ji} = O_{ij}^\top$ ，连接 Laplacian 的能量为

$$x^\top \mathcal{L}_{\text{conn}} x = \sum_{\{i,j\}} w_{ij} \|x_i - O_{ij} x_j\|_2^2. \quad (69)$$

核验项目	测得误差/数值	参照或上界
星形消元矩阵相对误差	0	精确恒等式
团采样期望相对误差	9.10×10^{-17}	精确无偏
谐波延拓能量相对误差	2.32×10^{-16}	精确恒等式
终端有效电阻最大相对误差	2.66×10^{-15}	精确保持
$\sum_e \tau_e - (n-1)$ 的绝对值	1.42×10^{-14}	0
Schur 谱误差 ε_S	0.06940	原图误差 0.09224
隐式传播相对能量误差	0.03712	0.10161
取 $\hat{L} = 0.8L$ 时解误差	0.25000	0.25000
质量项与消元不交换: 差的 Frobenius 范数	0.33333	1/3
隐式边权梯度与中心差分的绝对误差	9.25×10^{-10}	数值微分检查
DAG 终端精确消元相对误差	6.60×10^{-17}	精确恒等式
DAG 扰动解相对误差	0.03537	0.11111

Table 2: 本次随附脚本的实际输出, 完整精度见 `verification_results.json`。这些结果核验推导和实现, 不验证 KS16 的整体随机失败概率、近线性复杂度或 Graph DeltaNet 学习表现。

其对角块为加权重乘 I_r , 非对角块为 $-w_{ij}O_{ij}$ 。它满足块对角占优 (bDD) 条件

$$M_{ii} \succeq I_r \sum_{j \neq i} \|M_{ij}\|_2. \quad (70)$$

Kyng et al. (2016) 研究了这个结构及更一般的 bDD 系统, 得到稀疏消元与多重网格型求解结果。读复杂度时有两个限制: 块宽 r 在相应近线性表述中视为常数; 该文 Theorem 1.3 中 $O(n)$ 块非零元因子的存在/多项式时间构造, 不能与 Theorem 1.2 的快速求解器混成“近线性时间直接构造那个更强稀疏因子”的结论。

普通 Delta 因子 $I - \beta k k^\top$ 一般不是正交矩阵。甚至“边映射是收缩”也不足以保证 bDD: 能量 $(x_1 - ax_2)^2$, $0 < a < 1$, 对应矩阵

$$\begin{bmatrix} 1 & -a \\ -a & a^2 \end{bmatrix} \succeq 0, \quad a^2 < a, \quad (71)$$

第二行不对角占优。因此 connection Laplacian 是研究可学习几何/正交边传输的扎实起点, 但不是任意矩阵消息、任意门控或任意 PSD 能量的免费求解器。添加足够大的正则项可能恢复某些结构, 却会改变原目标, 应作为明确的新模型选择而非技术性等价变形。

11 数值核验与建议的研究推进顺序

11.1 已执行的核验: 只验证代数, 不冒充学习或加速实验

随附 `verify_formulas.py` 用 NumPy/SciPy 的 float64 稠密运算实现独立核验, 随机种子为 20261010。随机图有 $n = 24$ 、 $m = 50$, 终端数为 7。表2记录本次实际运行结果; 不同 BLAS/平台下最后几位可能有差异。对采样器, 程序枚举全部双抽样结果来验证精确期望, 并非用少量 Monte Carlo 样本推断无偏性。

谱误差实验将每条边乘以 $[0.85, 1.15]$ 中的独立正因子, 没有删除边, 因而不是稀疏化实验。矩阵谱误差通过 $\mathbf{1}^\perp$ 上的广义特征值计算, 避免把 Laplacian 的零特征值纳入条件数。

DAG 核验使用小型固定系数块三角系统，不是完整的 Graph DeltaNet/WY 实现。数值一致只能支持这些实例，不能替代上述定理证明。

11.2 第一优先级：精确 Graph-UT / 多端口 WY 的正确性与并行性

当前主线宜先保持精确。把逐节点多前驱 Delta 递归作为 oracle，对照 Graph-UT 与多端口 chunk 表达，测试不同拓扑宽度、深度、合流度、端口数、chunk 划分及状态宽度。在 float64 下比较前向误差与随机方向梯度；对小规模实例再做有限差分。置换节点编号并耦合相同随机性，验证等变性与合法拓扑调度不变性。

论文方法的关键不应仅是“能写出块矩阵”，而应说明哪些工作跨 chunk 可独立完成，哪些端口状态必须等待前驱，哪些输出能转为大矩阵运算；同时明确不显式展开所有路径。数学等价、低同步深度和实际 GPU 高吞吐是三个不同证据层级。目前本笔记没有实现或测量 Triton/CUDA 内核，因此不提供任何已获加速倍数。

11.3 第二优先级：把 Laplacian 保证用于一个真正匹配的模块

在确实采用正权无向图能量的任务上，可增加一个边界读出、谐波插值或隐式扩散模块，并保持当前 Delta 递归主干不变。比较精确直接求解、原算子迭代求解、随机近似因子预条件，以及直接替换近似算子的差异。后两者分别是“近似算法求原模型”和“精确求另一个近似模型”，应单独报告。

对于多个特征通道和多轮训练，要区分因子一次设置成本、多右端项应用成本、反向成本及边权变化导致的重建成本。至少报告能量误差、原方程残差、梯度误差、任务损失/指标、边/因子非零元数、峰值显存和端到端时间。若只测复用固定图时的求解，而学习过程中图每步变化，就不能把同一个摊销结论用于训练。

11.4 第三优先级：有可控收益时再研究近似端口压缩

近似端口可以作为第二阶段方向，但需先测精确多端口表示的真正瓶颈：是端口宽度、稀疏 gather、局部三角求解、同步依赖，还是显存带宽？仅在压缩确实针对主成本时，再研究式(64)及式(63)。合理的目标是“给定深度/稳定性假设下的输出与梯度误差预算”，而不是预先宣称得到适用于任意 DAG 的近线性 Laplacian 型求解器。

后续文献可用于定位实现难点。Baumann and Kyng (2024) 通过低度节点集合等机制研究近似消元的并行 work/depth；Liang et al. (2026) 的 2026 年修订预印本专门处理随机预条件器中的依赖追踪。这些工作支持“稀疏消元的并行调度本身是一项研究问题”的判断，却不能证明密集 WY 内核与稀疏消元具有相同瓶颈或性能。Gao et al. (2023) 则适合对照理论采样和实际鲁棒性之间的距离。

11.5 回读原文的路线与最终判断

结论。对 Laplacian graph learning，最有用的公式是能量最小化式(8)、星形-团式(11)、Schur 误差继承式(15)、逆稳定性式(17)和隐式梯度式(49)。对当前 DAG Graph DeltaNet，最有用的不是把这些谱保证改个名字，而是精确多端口接口式(61)及其非对称误差条件式(62)。两条线可共存：前者提供有明确能量语义的辅助模块，后者继续保持递归 Delta 主线的非交换路径语义。先证明并实现精确并行，再决定近似什么，是更稳妥且更容易形成可审查贡献的推进顺序。

原始材料	回读时要解决的问题
Kyng 博士论文, 第 2 章	核对指数谱近似、伪逆、零空间和 Schur 补的约定。
KS16, 第 3 节与算法页	核对主定理的概率/期望限定; 逐步推导双抽样权重。
KS16 的误差分析部分	追踪当前随机图、固定原图白化、停止事件与方差控制。
Kyng 博士论文, 第 3.4 节	区分稀疏因子、部分消元与稀疏边界 Schur 输出。
Kyng 博士论文, 第 2.4、4 章	核对生成树、leverage 边缘概率及行列式与消元的关系。
Kyng 博士论文, 第 5 章及有向 LU	核对 Eulerian 条件、左右零空间及辅助误差度量。
Connection Laplacian 论文	核对 bDD、固定块宽, 以及快速求解与稀疏因子存在性。

Table 3: 按照“能量语义 $\rightarrow$ 随机消元 $\rightarrow$ 学习接口 $\rightarrow$ 适用边界”的顺序回读, 比直接从全部技术引理开始更适合当前研究。章节号对应本文引用版本; 纸本页码与 PDF 页序可能不同。

Appendix A. 内部节点恢复: 终端等效不是全部预测等效

A.1 完全相同的 Schur 补, 也能给出不同的内部值

考虑一个内部节点连接两个终端的星形, 边权为 a, b 。消元后的两终端等效边权为 $ab/(a+b)$, 而给定终端值 y_1, y_2 的内部谐波值是

$$h_F = \frac{ay_1 + by_2}{a + b}. \quad (72)$$

边权 $(a, b) = (2, 2)$ 与 $(4, 4/3)$ 都产生权重为 1 的终端边, 即两者的 S 完全相同。可是当 $(y_1, y_2) = (1, 0)$ 时, 内部预测分别为 $1/2$ 与 $3/4$ 。所以从 S 的精确保持或近似保持, 不能单独恢复所有内部表示。这是对“池化无损”的一个最小反例: 无损必须相对于明确的观察位置与任务定义。

A.2 完整图的谱近似可以额外控制谐波恢复误差

Proposition 7 (固定边界下的恢复稳定性, [推导]) 设 $L, \hat{L}$ 在同一节点集上满足式(3), 固定同一个 Y_C , 令 $H, \hat{H}$ 分别为两图的谐波延拓。若 $L_{FF} \succ 0$, 则

$$\|\hat{H} - H\|_{L,F} \leq \frac{\varepsilon}{1 - \varepsilon} \|H\|_{L,F}, \quad (73)$$

$$\|\hat{H}_F - H_F\|_F \leq \frac{\varepsilon}{(1 - \varepsilon)\sqrt{\lambda_{\min}(L_{FF})}} \|H\|_{L,F}. \quad (74)$$

Proof 记 $E = \hat{H} - H$, 则 $E_C = 0$, 而谐波最优条件给出 $\text{tr}(E^\top \hat{L} \hat{H}) = 0$ 与 $\text{tr}(E^\top L H) = 0$ 。因此

$$(1 - \varepsilon) \|E\|_{L,F}^2 \leq \text{tr}(E^\top \hat{L} E) = -\text{tr}(E^\top (\hat{L} - L) H) \leq \varepsilon \|E\|_{L,F} \|H\|_{L,F}. \quad (75)$$

最后一步由式(4)逐列应用 Cauchy-Schwarz 得到。除去 $E = 0$ 的平凡情况后约去其能量范数, 得到第一式。又因 $E_C = 0$, $\|E\|_{L,F}^2 = \text{tr}(E_F^\top L_{FF} E_F)$, 得到第二式。■

两个假设差别是实质性的: 前一反例仅保持终端 S , 本命题保持整个原图的谱几何。式(74)还

揭示了终端选择的作用：若某些内部区域与边界耦合很弱，接地块的最小特征值很小，即使能量误差有界，逐点或欧氏预测误差仍可能大。对稀疏标签学习，终端覆盖和算子近似质量应分开评估。

Appendix B. DAG 递归的伴随：精确并行实现应该核对哪些梯度

B.1 非对称线性系统的微分

对式(56)，令 $g = \nabla_s \ell$ ，解转置伴随系统

$$\mathcal{T}^\top u = g, \quad d\ell = u^\top db + u^\top (d\mathcal{K})s. \quad (76)$$

这是由 $\mathcal{T}ds = db + (d\mathcal{K})s$ 直接得到的恒等式。这里不能像对称 Laplacian 那样不加区分地复用相同方向的求解；转置系统沿反向拓扑依赖传播。对单个块有 $\nabla_{\mathcal{K}_{vu}} \ell = u_v s_u^\top$ ，但若多个块共享同一组门控参数，应先按链式法则汇总，而不能当作互不相关的自由矩阵。

B.2 Delta 更新本身的局部反向公式

将总伴随还原为与 S_v 同形状的 U_v ，并令 G_u 为损失对节点状态的直接贡献，则

$$U_u = G_u + \sum_{v: u \in \text{pred}(v)} \alpha_{uv} U_v A_v^\top. \quad (77)$$

在门控和混合系数的直接依赖固定时，设 $r_v = v_v - M_v k_v$ ，可得

$$\frac{\partial \ell}{\partial \alpha_{uv}} = \text{tr}(U_v^\top S_u A_v), \quad \nabla_{v_v} \ell = \beta_v U_v k_v, \quad (78)$$

$$\frac{\partial \ell}{\partial \beta_v} = \langle r_v, U_v k_v \rangle, \quad \nabla_{k_v} \ell = \beta_v (U_v^\top r_v - M_v^\top U_v k_v). \quad (79)$$

key 的梯度有两项，因为 k_v 同时出现在预测残差和写入方向中。若 key 做归一化、 α 来自 softmax、 β 来自 sigmoid，继续对这些参数化反传；若 loss 或其他模块还有直接依赖，也应加上它们。以上是递归数学的参考梯度，不依赖某一种 WY 内核实现。

精确多端口表示的训练等价性，至少应对照式(76)–(79)。前向一致并不能排除自定义 backward 遗漏源项、混合权重或 key 的第二项。保存内部状态、checkpoint 重计算或保存结构化因子，是内存与计算量的选择，不能改变应求导的原方程。随附脚本核验一般 DAG 伴随的方向导数和一个独立局部 Delta 更新的四类参数梯度；这仍不等于已经验证完整 Graph-UT / WY 训练实现。

References

- Yves Baumann and Rasmus Kyng. A framework for parallelizing approximate gaussian elimination. In *Proceedings of the 36th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pages 195–206, 2024. doi: 10.1145/3626183.3659987.
- Michael B. Cohen, Jonathan Kelner, Rasmus Kyng, John Peebles, Richard Peng, Anup B. Rao, and Aaron Sidford. Solving directed Laplacian systems in nearly-linear time through sparse LU factorizations. In *IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018. URL <https://arxiv.org/abs/1811.10722>.

- Florian Dörfler and Francesco Bullo. Kron reduction of graphs with applications to electrical networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 60(1):150–163, 2013. URL <https://arxiv.org/abs/1102.2950>.
- Yuan Gao, Rasmus Kyng, and Daniel A. Spielman. Robust and practical solution of Laplacian equations by approximate elimination. arXiv:2303.00709, 2023. URL <https://arxiv.org/abs/2303.00709>. Cited under the 2023 preprint title.
- Rasmus Kyng. *Approximate Gaussian Elimination*. PhD thesis, Yale University, 2017. URL <https://rasmuskyng.com/rjkyng-dissertation.pdf>.
- Rasmus Kyng and Sushant Sachdeva. Approximate gaussian elimination for Laplacians: Fast, sparse, and simple. In *IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, 2016. URL <https://arxiv.org/abs/1605.02353>. Theorem and algorithm statements in this note follow arXiv:1605.02353v1.
- Rasmus Kyng, Yin Tat Lee, Richard Peng, Sushant Sachdeva, and Daniel A. Spielman. Sparsified cholesky and multigrid solvers for connection Laplacians. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016. URL <https://arxiv.org/abs/1512.01892>.
- Tianyu Liang, Chao Chen, Yotam Yaniv, Hengrui Luo, David Tench, Xiaoye S. Li, Aydin Buluç, James Demmel, and Jie Chen. Tackling parallelization challenges of randomized preconditioners with dependency tracking. arXiv:2505.02977v3, 2026. URL <https://arxiv.org/abs/2505.02977v3>. Version revised September 7, 2026.
- Andreas Loukas. Graph reduction with spectral and cut guarantees. *Journal of Machine Learning Research*, 20(116):1–42, 2019. URL <https://jmlr.org/papers/v20/18-680.html>.
- Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. URL <https://arxiv.org/abs/0803.0929>.
- Joel A. Tropp. Freedman’s inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624. URL <https://arxiv.org/abs/1101.3039>.
- Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL <https://arxiv.org/abs/2406.06484>.
- Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 912–919, 2003. URL <https://pages.cs.wisc.edu/~jerryzhu/pub/zgl.pdf>.