arXiv:2302.03611v1 [math.CO] 7 Feb 2023

分类热带线段上的树拓扑变化

SHELBY COX

Abstract. 系统发育树空间在热带几何中自然地以热带格拉斯曼流形的形式出现。因此，热带几何提出了一种自然的热带路径概念：在热带格拉斯曼流形中，两棵树之间的热带线段。此前有人猜想，沿着这样的线段，树拓扑会通过一种称为最近邻交换（NNI）的组合操作发生变化。我们给出了这一猜想的反例，但证明了对于一般位置树，沿热带线段发生的树拓扑变化要么是 NNI 移动，要么是“四分支演化枝重排”移动。此外，我们证明，沿热带线段发生的 NNI 移动次数最多可达n^2；而随机选取两棵端点树时，移动次数的平均值为 $O(n(\log\ n)^4)$。相比之下，将一棵树转换为另一棵树所需的 NNI 移动平均次数为 $O(n\log\ n)$。

1. Introduction

系统发育树是以叶节点标记的度量树，用于表示一组生物之间的进化关系。本文中的系统发育树也都是有根的。根表示这些生物的共同祖先，叶节点表示现存生物，而边的长度表示已发生的突变量。由 n 个现存生物（即叶节点）组成的系统发育树空间是一个拓扑空间，我们将其记为 Tree$_n$。

Tree$_n$ 空间已通过拓扑结构与度量结构的多种组合得到实现：作为 CAT(0)度量空间（Billera、Holmes 和 Vogtman [1]）、作为热带簇（Speyer 和 Sturmfels [2], ；Ardila 和 Klivans [3]）、作为边乘积空间（Kim [4],Moulton 和 Steel [5], ；Gill 等人 [6]），以及作为 Wald空间（Gabra 等人 [7]）。本文聚焦于将 Tree$_n$ 实现为热带簇的一种特定方式；有关 Tree$_n$ 构造及相关结果的精彩综述，请参阅 [7, 第 1] 节。

在某种嵌入下，Tree$_n$ 被赋予热带簇的结构——超度量空间——而在这种实现中，Tree$_n$ 是热带凸的，这意味着， n 叶节点上的两棵系统发育树之间的热带线段 上的任意一点，本身都可以解释为一棵 n 叶节点上的系统发育树。在这一框架下，Lin 和 Yoshida 研究了样本均值 [8], ，Yoshida 等人则发展出一种主成分分析方法 [9]。在这些结果中，树空间里的热带多胞体发挥着关键作用，然而，对于由树顶点构成的热带多胞体中会出现哪些树，人们仍知之甚少。因此，本文的首要目标是回答以下问题：

问题 1.1. 沿着热带线段，树拓扑（即树结构）会如何变化?

日期：2023年2月8日。

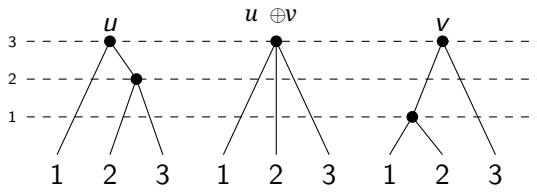


FIGURE 1. 具有单次 NNI 移动的热带线段上的转折点。

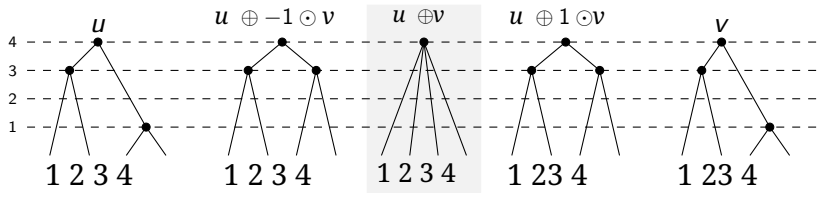


FIGURE 2. 具有四分支演化枝重排的热带线段上的转折点。

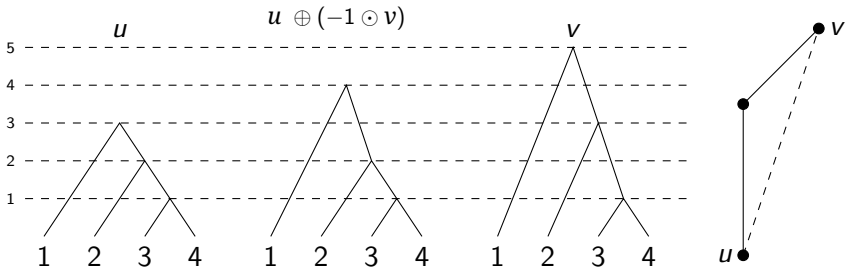


FIGURE 3. 具有恒定树拓扑的热带线段（左），以及 u 与 v之间的热带线段（右侧实线）和直线段（右侧虚线）上的转折点。

此前曾有人猜想，热带线段上的树会通过最近邻交换（NNI）发生变化。这并不完全正确。例如，在 [10]中已经证明，两棵具有四个叶节点的树之间的热带直线可能以大于 0 的概率经过星形树，而这并非 NNI 移动。不过，实际情况也不会糟糕太多。

热带线段是经典线段的串接。树拓扑可能在经典线段相互连接的点处发生变化，但不会在任何经典线段的内部发生变化。我们将这些点称为转折点（无论树拓扑是否发生变化）。在第4节中，我们证明以下结果：

定理4.1。树拓扑变化只会发生在转折点处；在每个转折点处，可能发生以下三种情况之一：

(1)（最近邻交换）一个内部顶点有三个子节点，见图1。(2)（四分支演化枝重排）一个内部顶点有四个子节点，见图 2。(3)（无拓扑变化）所有内部顶点都有两个子节点，见图3。

四分支群重排移动可以通过三次最近邻交换移动实现，因此值得探究两个一般树之间的热带线段上会发生多少次 NNI 移动——每个转折点可能贡献 0、1 或 3 次 NNI 移动。我们将热带路径上发生的 NNI 移动次数称为热带 ***NNI*** 数 （将一次四分支簇重排计为三次 NNI 移动）。

问题 1.2. n个叶节点上的随机树对之间的热带 NNI 数是多少？它与***NNI*** 距离相比如何？也就是说，将一棵树变换为另一棵树所需的最少 NNI 移动次数是多少？

所谓在 $[n]$个叶节点上随机选取的系统发育树对，是指从$[n]$ 个叶节点上的（有限多个）二叉树拓扑中均匀随机地选取树拓扑，并且相应的树度量足够一般，具体定义见第3节。已知，对于 n 个叶节点上的随机系统发育树对，平均 NNI 距离为 $O(n\log(n))$，而计算最少 NNI 移动次数是 NP 难题 [11]。定理4.1表明，热带线段上的转折点数量是 NNI 移动次数三倍的一个上界。首先，我们给出一个例子，说明热带 NNI 数可能远大于最少 NNI 移动次数。

定理 5.1. 对于每个 n，都存在定义在 n 个叶节点上的一般位置树对 T_1、T_2，使得从 T_1 到 T_2 的热带 NNI 数为 $\binom{n-1}{2}$。

接下来，我们证明平均情形下的上界。

推论 6.2. 在 n 个叶节点上随机选取两棵树，它们之间的热带线段上发生的 NNI 移动次数的期望为 $O(n(\log(n))^4)$。

本文结构如下。第2节介绍树空间、热带数学和最近邻交换（NNI）的背景与约定。第3节给出转折点的一种非标准定义，该定义将在本文余下部分中发挥作用。第4节对两棵一般树之间热带线段上发生的树拓扑变化进行分类，并证明定理4.1。第5节介绍一类一般位置树对，其热带 NNI 数远高于预期的热带 NNI 数（证明定理 5.1）。第6节界定在 n 个叶节点上随机选取的两棵树之间热带线段上的预期 NNI 距离（证明定理 6.1）。

2. Background

2.1. 热带半环。 热带半环是研究树的理想场所。我们将在热带最大加法半环 $(\mathbb{R}\cup\{-\infty\},\ \oplus,\ \odot)$ 中进行研究，其中热带加法和乘法分别定义为取最大值和经典加法：

$$a\oplus b:=\max\{a,b\},\ a\odot b:=a+b.$$

注意，$-\infty$ 是加法的单位元，而 0 是乘法的单位元。热带除法就是经典减法；热带加法逆元通常不存在。从几何上说，我们将在热带射影空间 $\mathbb{R}^n/\mathbb{R}\mathbb{1}$ 中进行研究，其中 $\mathbb{1}=(1，1，\ldots\ldots，1)$ 是全 1 向量。

定义 2.1 ([12])。 u_1的热带凸包、……、$u_k\in\mathbb{R}^n$ 是所有由 u_i 构成的热带线性组合的集合：

$$\operatorname{tconv}(u_1,\ldots,u_k)=\{\lambda_1\odot u_1\oplus\cdots\oplus\lambda_k\odot u_k\mid\lambda_i\in\mathbb{R}\}.$$

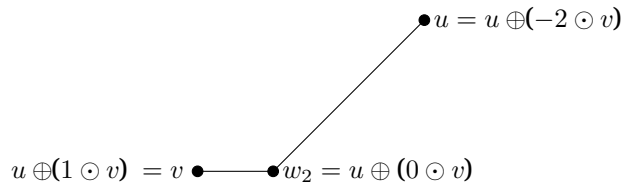


FIGURE 4. 从 u 到 v的热带线段。

注意，若 $u'_i = \alpha_i \odot u_i$，则

$$\begin{aligned}\operatorname{tconv}(u_1, \dots, u_k) &= \{\lambda_1 \odot u_1 \oplus \cdots \oplus \lambda_k \odot u_k \mid \lambda_i \in \mathbb{R}\} \\ &= \{(\lambda_1 - \alpha_1) \odot (\alpha_1 \odot u_1) \oplus \cdots \oplus (\lambda_k - \alpha_k) \odot (\alpha_k \odot u_k) \mid \lambda_i \in \mathbb{R}\} \\ &= \{\lambda'_1 \odot u'_1 \oplus \cdots \oplus \lambda'_k \odot u'_k \mid \lambda'_i \in \mathbb{R}\} \\ &= \operatorname{tconv}(u'_1, \dots, u'_k)\end{aligned}$$

即，tr 热带凸包在热带射影环面中定义良好，$\mathbb{R}^m/\mathbb{R}\mathbb{1}$。

热带线段是两个点 u、v 的热带凸包。它至多由 $n-1$ 条经典（即欧几里得）线段首尾连接而成，并且可以使用定义 2.2 中的算法计算。

定义 2.2（热带线段算法）。

（1）**输入：** u，$v \in \mathbb{R}^n$。（2）$\Lambda(u, v) = [u_i - v_i \mid i \in [n]]$，按从小到大排序。（3）$w_i = u \oplus (\lambda_i \odot v)$，其中 λ_i 是 Λ 中第 i 小的元素。

（4）**输出：** 从 w_i 到 w_{i+1} 的欧几里得线段的首尾连接。

定义 2.3。在上述算法中，$\Lambda(u, v) = \{u_i - v_i \mid i \in [n]\}$ 是转折点标量的集合，而 w_i 是从 u 到 v 的热带线段的转折点。

例 2.4。令 $u = (3, 3, 1)$ 和 $v = (3, 2, 3)$。则 $\Lambda = [-2, 0, 1]$，热带直线上的转折点为：

$$\begin{aligned}u \oplus (-2 \odot v) = (3,3,1) \oplus (1,0,1) = (3,3,1) &= u \\ u \oplus (0 \odot v) = (3,3,1) \oplus (3,2,3) = (3,3,3) &= (0,0,0) \\ u \oplus (1 \odot v) = (3,3,1) \oplus (4,3,4) = (4,3,4) &= v.\end{aligned}$$

该热带线段如图4所示。

2.2。 系统发育树与超度量。

定义 2.5。具有 n 个叶节点的系统发育树 是一个有根[1], 度量树，其内部边长度为正，叶标签集为 $[n] = \{1, 2, \dots, n\}$。我们将该树的根（一个内部顶点）记为 ρ。

定义 2.6。给定 T 的两个叶节点 i, j，i 和 j 的最近公共祖先 是路径 $i \rightsquigarrow j$、$\rho \rightsquigarrow i$、$\rho \rightsquigarrow j$ 交集中的内部顶点。我们用 $\mathrm{LCA}_T(i, j)$ 表示 T 中 i, j 的最近公共祖先。

[1]一般而言，系统发育树不一定是有根的。

给定一棵度量树 T 及其两个顶点 a、b，我们将从 a 到 b 的距离定义为从 a 到 b 的唯一路径上各边长度之和。将此距离记为$d_T(a,b)$。

我们尤其关注等距树。

定义 2.7。 如果系统发育树的每个叶节点到h 的距离都相同 ρ，则称其为等距树 。

对于有根等距树，内部顶点的高度是良定义的，因此为内部顶点的高度引入一种记号会很有用。

定义 2.8。 设 x是 T的一个内部顶点。对于任意满足 $\mathrm{LCA}_T(i,j) = x$的 $i,j \in [n]$，

$$h_T(x) = \frac{1}{2} d_T(i,j)$$

我们可以为每棵系统发育树 T（无论是否等距）关联一个记录叶节点间距离的向量。将此向量记为 u^T（或简记为 u）：

$$u = u^T = (d_T(1,2), d_T(1,3), \ldots, d_T(n-1,n)) \in \mathbb{R}^{\binom{n}{2}}.$$

这称为 T的树度量 。我们常常交替使用 u 和 T 。因此，$i \in [n]$可以称为 u的叶，$\mathrm{LCA}_u(i,j) = \mathrm{LCA}_T(i,j)$，且 $h_u(x) = h_T(x)$。

反过来，给定 $\mathbb{R}^{\binom{n}{2}}$中的一个向量，我们想知道它何时是来自系统发育树的叶间距离的向量，何时是来自等距树的叶间距离的向量。这分别对应四点条件和三点条件：

定理 2.9（四点条件 [13]）。 $u \in \mathbb{R}^{\binom{n}{2}}_{\geq 0}$ 是系统发育树的树度量T，当且仅当对于所有互不相同的 i、j、k、$l \in [n]$，$\max\{u_{ij}+u_{kl}, u_{ik}+u_{jl}, u_{il}+u_{jk}\}$ 至少在两个位置上取到。

定理 2.10（三点条件 [13]）。 $u \in \mathbb{R}^{\binom{n}{2}}_{\geq 0}$ 是等距树的树度量 T，当且仅当对于所有互不相同的 i、j、$k \in [n]$，$\max\{u_{ij}, u_{ik}, u_{jk}\}$ 至少在两个位置上取到。

稍后在进行树计数时，在根ρ旁添加一个标记为 0 的叶会很方便，这样一般树的所有内部顶点的度都恰好为 3。不过，我们不会将到叶 0 的距离纳入树度量向量。

定义 2.11。 超度量是满足三点条件的 $\mathbb{R}^{\binom{n}{2}}$ 中的一个向量。

注记 1。 超度量 $u \in \mathbb{R}^{\binom{n}{2}}_{\geq 0}$ 是 $[n]$ 上的一个度量。

我们通常通过给出其树度量作为 $\mathbb{R}^{\binom{n}{2}}$ 中的一个向量（超度量）来描述等距树。此外，将超度量视为热带射影空间中的元素是很自然的，也就是说，对超度量向量添加 $\lambda\mathbb{1}$ 不会改变等距树的结构。当我们对超度量向量添加 $\lambda\mathbb{1}$ 时，相应等距树的变化体现为外部（与叶相邻的）边长发生变化。

以下定理对于我们研究热带树空间至关重要：

定理 2.12 ([13])**.** 超度量空间是热带凸的。

这意味着，两个超度量（或等距树）之间的热带直线仍处于超度量空间中。

2.3. **系统发育树空间**。 我们所说的树拓扑，是指树作为叶标记图的组合结构。我们将通过三种方式对其进行定义：使用边分割（BHV 定义）、使用 argmax 函数（度量定义），最后使用热带格拉斯曼流形上的扇结构（热带定义）。首先从 BHV 定义开始:

定义 2.13。 给定一棵系统发育树 T 和一条边 e， T 在 e 处的边分割是指 e 在 T的叶节点上诱导出的划分。

定义 2.14 （BHV 树拓扑）。 如果两棵树具有相同的边分割集合，则它们具有相同的拓扑 。

我们也可以使用树度量来确定树拓扑。若 T 是一棵树， u 是其对应的超度量，且 $S \subset \binom{[n]}{2}$，则定义 $\operatorname{argmax}_u^1$ 为 S 中使 u_{ij} 取得最大值的指标子集。接着，定义 $\operatorname{argmax}_u^m\{S\}$ 为 S 中达到 u_{ij} 在指标 S 中第 m 大取值的指标子集。形式化地，我们在下文定义***argmax*** 函数。

定义 2.15。 令 $\operatorname{argmax}_u(S) = \{s \in S : u_s = \max_{x\in S} u_x\}$ ，对任意 $S \subseteq \binom{[n]}{2}$。

定义 2.16. 两个系统发育树 u, v 具有相同的树拓扑，当且仅当

$$\operatorname{argmax}_u\{ij, ik, jk\} = \operatorname{argmax}_v\{ij, ik, jk\}$$

对所有 $i < j < k \in [n]$均成立。

最后，我们利用热带格拉斯曼流形上的扇结构来定义树拓扑。热带化线性空间具有多种扇结构。这里我们考虑 Speyer 和 Sturmfels 在 [2]中首次描述的扇结构。我们称由超度量 u, v 表示的两棵树具有相同的树拓扑，当且仅当它们位于该扇结构中同一个锥的相对内部。下面总结了树拓扑的各种定义。

定义 2.17. 由 $u_1, u_2 \in \mathcal{U}_n$表示的、具有 n 个叶节点的两棵等距树 T_1, T_2 ，具有相同的（粗）树拓扑，当且仅当以下任一等价陈述成立:

(1) u_1, u_2 均位于$\mathcal{U}_n$上的粗扇结构中同一个锥的相对内部；或 **(2)** 对每个三元组 $i < j < k \in [n]$， $\operatorname{argmax}_{u^1}\{ij, ik, jk\} = \operatorname{argmax}_{u^2}\{ij, ik, jk\}$ ；或 **(3)** T_1, T_2 作为叶标记树彼此同构；或 **(4)** T_1, T_2 确定相同的边分割（BHV 定义）。

我们定义系统发育树空间。

定义 2.18。 系统发育树空间 是由具有 n 个叶节点的等距系统发育树构成的集合；等价地，它是满足三点条件的 n 个点上的度量集合:

$$\begin{aligned}\operatorname{Tree}_n^{\mathrm{tr}} &= \{(d_T(i,j))_{i<j} \in \mathbb{R}^{\binom{n}{2}}/\mathbb{R}\mathbb{1} \mid T \text{ a phylogenetic tree on } n \text{ leaves}\} \\ &= \{u \in \mathbb{R}^{\binom{n}{2}}/\mathbb{R}\mathbb{1} \mid \max(u_{ij}, u_{ik}, u_{jk}) \text{ is achieved at least twice for } i < j < k.\}\end{aligned}$$

定理 2.19 ([2], 定理 4.2)。 树空间是由单纯形锥组成的并，每个锥都包含表示固定树拓扑 τ 的所有点。

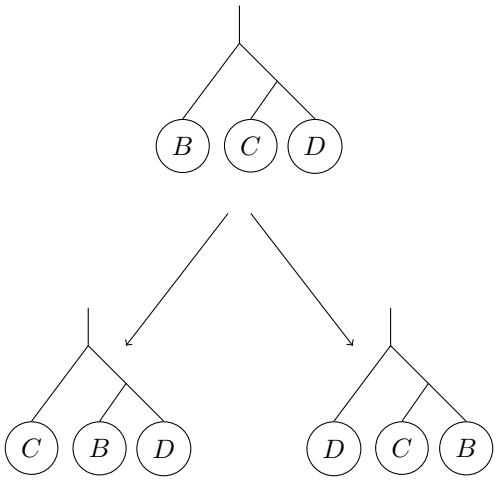


FIGURE 5. 可能的 NNI 移动

注 2。 Ardila 和 Klivans 在 [3] 中描述了热带格拉斯曼流形上的另一种扇结构。它比我们上文描述的扇结构更细；两棵树 u, v 位于这一更细结构的同一个锥中，并且具有相同的排序树拓扑，当且仅当它们组合同构，且两棵树中所有 $n-1$ 内部顶点的先后顺序相同，或者等价地，

$$\operatorname{argmax}_u^k \binom{[n]}{2} = \operatorname{argmax}_v^k \binom{[n]}{2}, \text{ for all } k \in \mathbb{N},$$

其中，$\operatorname{argmax}_u^k(S)$ 表示 S 中满足 u 在 S 中的元素里取得第 k 大值的元素。

请注意，在图 2 中，第二棵树 $w = u \oplus -1 \odot v$ 位于细扇结构的一个余维 1 锥中（因为该树是二叉分支，但两个内部顶点处于相同高度）。更确切地说， w 与 $u \oplus v$ 之间的线段（图 2 中的第二棵和第三棵树）位于代表不同**排序** 树拓扑的两个胞腔的交集内：$w_{12} \le w_{34} \le w_{13} = w_{14} = w_{23} = w_{24}$ 和 $w_{34} \le w_{12} \le w_{13} = w_{14} = w_{23} = w_{24}$。

转折点的数量与 NNI 距离接近，因为每个转折点都位于粗扇结构中余维至多为二的锥内，因此至多需要三次 NNI 移动。在更精细的扇结构中，热带线段可以穿过余维高达 $n/2$ 的锥，这表明单个转折点处可能需要许多有序 NNI 移动。此外，在粗扇结构中，热带线段仅在点处与非极大锥相交；而在更精细的扇结构中，热带线段可能与非极大锥沿整个经典线段相交。

2.4.最近邻交换。转换系统发育树的一种常见方法是最近邻交换（NNI），该方法由 [14] 和 [15] 独立提出。NNI 移动是对树叶节点进行划分的任意三个支系的所有可能重排。有根树的一条内部边会将叶节点划分为三个支系：B、 C 和 D。图 5 所示的移动之一就是一次 NNI。

定义 2.20。 系统发育树的支系 T 是从一个给定的固定内部顶点向下延伸的所有叶节点的集合。

n 叶节点的 NNI 图，其顶点表示所有由 n 个叶节点构成的树拓扑；若两个顶点所代表的树拓扑仅相差一次 NNI 移动，则它们之间有边相连。注意，在 NNI 图中，我们不考虑树的边长。

NNI 图是一个经过充分研究的图。1997年，DasGupta 等人证明了，在 NNI 图中计算最短路径是 NP 完全问题 [16]。而 [17] 提供了一种多项式时间$O(n^2)$算法，可将最短 NNI 路径的长度近似到 4 log(n) 倍以内 $+4$。已知 NNI 图的直径为 $O(n\log(n))$，随机选取的两个顶点之间的期望距离也是 $O(n\log(n))$。我们尤其关注的是，热带 NNI 数是否接近 [17] 中给出的 NNI 距离多项式时间近似值。

沿热带线段可能发生的另一种移动是四分支群重排移动，即四个分支可以同时汇合。图 2 给出了一个示例。树拓扑中的这种变换也可以通过连续执行三次 NNI 移动来实现。

3. Redefining Turning Points for Trees

我们首先从端点树的内部顶点角度重新诠释热带线段算法（定义 2.2）中的转折点标量，这将使我们能够在不提及树度量的情况下描述转折点。

定义 3.1。 给定一对系统发育树 T_1, T_2，其叶节点为 n 个，将 T_1, T_2 的本质对 定义为

$$\Pi(T_1, T_2) := \{(x_1, x_2) \mid x_k \in T_k, \exists i, j \in [n], i \neq j \text{ s.t. } x_k = \mathrm{LCA}_{T_k}(i, j)\}$$

在接下来的各节中，我们会发现，考虑具有足够一般超度量的树会使结果简洁得多。下面的定义将说明“足够一般的超度量”中“一般”的含义。

定义 3.2。 具有 n 个叶节点、距离向量为 u 的等距树 T，若满足以下任一等价条件，则称其为一般：

(1) u 位于 $\mathrm{Tree}_n^{\mathrm{tr}}$ 的某个全维锥的相对内部；(2) T 恰有 $n-1$ 个内部顶点（包括根）；(3) T 的所有内部顶点的度均为 3，唯有根的度为 2。

注意，如果 u 是一般位置，那么对于任意 $\lambda \in \mathbb{R}$， $\lambda \odot u$ 也都是一般位置，因为热带乘法只影响悬挂边长。

非一般位置的树是未解析的。

定义 3.3. 如果树 T 未解析（位于内部顶点 v），即某个内部顶点 v 有两个以上子节点。

如果树中某些边长相同，那么在 T_1 和 T_2 之间的热带线段上也可能遇到高度未解析的树。因此，我们还要求 T_1 和 T_2 作为一对是一般位置的。我们称这样的一对为一般位置对，并在定义 3.4 中正式定义这种树对。

定义 3.4. 具有 n 个叶节点、距离向量分别为 u 和v 的一对等距树 T_1， T_2，称为一般位置对，如果：

(1) u 和 v 均为一般情形，且

(2) 对于 T_i 的所有互不相同的内部顶点 x_i, y_i， $h_{T_1}(x_1) - h_{T_2}(x_2) \neq h_{T_1}(y_1) - h_{T_2}(y_2)$。

对于一般树对，转折点的数量与超度量的具体选择无关。

引理 3.5. 设 T_1, T_2 是叶集为 $[n]$ 的系统发育树。那么，对于在 T_k 上充分一般地选择的任意超度量 u^k （即 u^1, u^2 是一般位置对），

$$\Lambda(u_1, u_2) = \{2(h_{u_1}(x_1) - h_{u_2}(x_2)) \mid (x_1, x_2) \in \Pi(T_1, T_2)\}$$

证明. 根据定义 2.2，T_1 与 T_2 之间热带线段的转折点出现在 $u^1 \oplus \lambda_{ij} \odot u^2$ 处，其中 $\lambda_{ij} = u^1_{ij} - u^2_{ij}$。

我们可以用内部顶点上的高度来改写上面的表达式。

$$\lambda_{ij} = 2(h_{T_1}(x_1) - h_{T_2}(x_2)), \text{ where } x_k = \mathrm{LCA}_{T_k}(i, j)$$

反过来，给定内部顶点 $x_1 \in T_1$ 和 $x_2 \in T_2$， $\mu_{x_1x_2} = 2(h_{T_1}(x_1) - h_{T_2}(x_2))$ 是转折点对应的标量，当且仅当对于某个 i， $j \in [n]$，有 $\mu_{x_1x_2} = \lambda_{ij}$；这又当且仅当存在 i， $j \in [n]$，使得对于 $k = 1$，2， $x_k =$ 是 $\tau_k(i,\ j)$ 的最近公共祖先。□

推论 3.6。 设 T_1、T_2 为叶集为 $[n]$ 的系统发育树。则对于在 T_k 上任意充分一般地选取的超度量 u_k，$\#\Lambda(u_1、\ u_2) = \#\Pi(T_1、\ T_2)$。

因此，两个组合（非度量）树结构之间的转折点数量可以定义为：

定义 3.7。 树 T_1、 T_2 的热带交换数 记为 tI(T_1， T_2)，它是对 T_1， T_2 上任意选取的一般位置权重 u_1， u_2 而言， T_1 与 T_2 之间的转折点数量。

4. The Tropical Line and NNI

现在我们来探讨本文的第一个主要问题：在两棵树 T_1 和 T_2 之间的热带线段上，可能发生哪些树拓扑变化?

事实证明，即使对于一般位置对，也可能存在一个并非单个 NNI 的转折点。例如，图 2 中的端点树构成一般位置对，但热带线段会经过一棵具有一个拥有四个后代的内部顶点的树（这与在这些树上具体选取何种度量无关）。本节将证明，当端点树的选取充分一般时，可能发生的唯一变换是：单个 NNI，或单个四分支演化枝重排。

4.1. 可能的转折点。

本节的主要定理如下：

定理4.1。 在一般树对之间的热带线段的每个转折点处，中间树具有以下三种形式之一：

(1) 除一个有 *3* 个子节点的内部顶点外，其余部分均为三价树（即经过余维*1*胞腔的 *NNI* 中点），示例见图*1*；(2) 除一个有 *4* 个子节点的内部顶点外，其余部分均为三价树（即经过余维 *2* 胞腔的四分支演化枝重排中点），示例见图 *2*；(3) 一般树（仍处于顶维胞腔中），示例见图*3*。

推论4.2. 在两棵一般位置树 u, v 之间的热带直线上，不存在具有一个拥有 *5* 个或更多子节点的内部顶点的树。

以下定义将有助于证明该定理：

定义 4.3. 对于 u，v 一对具有 n 个叶节点的树，令 $G(u \geq v)$ 为一个顶点为$[n]$、边为 $(i,\ j)$ 的图，使得 $u_{ij} \geq v_{ij}$。

注意，$G(u \geq v) \cup G(u \leq v) = K_n$。

引理 4.4. 若 G 和 H 是 K_5的子图，且满足 $G \cup H = K_5$，则 G、H中至少有一个包含奇圈。

证明。假设 G 和 H 均不包含奇圈。那么 G、H 是定义在五个顶点上的二分图。设 A，B，C 是 G 中颜色相同的不同顶点。那么，$\{A,B\}$、$\{A,C\}$和 $\{B,C\}$ 这几对中至少有一对在 H中颜色恒定。因此，G和 H都不包含连接该顶点对的边。这与假设 $G \cup H = K_5$矛盾。□

定理*4.1* 的证明。反设存在某个 $\lambda \in \Lambda(u,\ v)$，使得树 $w = u \oplus \lambda \odot v$ 有一个内部顶点拥有 5 个或更多子节点。将该内部顶点记为 w_0。从 w_0 的每个子节点中各选取一个后代叶节点，并将它们记为 1，2，3，4，5，.....。那么 $10 = \binom{5}{2}$ 的距离相同：

$$\max(u_{12}, \lambda + v_{12}) = \max(u_{13}, \lambda + v_{13}) = \cdots = \max(u_{45}, \lambda + v_{45}).$$

引理 4.4 表明，$G(u \geq \lambda + v)$、$G(u \leq \lambda + v)$ 中必有一个包含奇圈。这意味着，对于 $k = 1$ 或 $k = 2$:

$$u_{12} = u_{23} = \cdots = u_{2k+1,1}$$

（或者对称地，对 $\lambda \odot v$ 也成立）。不失一般性，假设涉及 u 的等式成立。那么，除最后一个距离外，其余距离都表明，对于所有 $i \in \{1, 3, \ldots, 2k+1\}$ 和 $j \in \{2, 4, \ldots, 2k\}$，$u_0 = \mathrm{LCA}_u(i, j)$，这意味着奇数编号和偶数编号的叶节点分别是不同子节点的后代。但最后一个距离 $u_{2k+1,1}$ 说明 1、$2k+1$ 也有最近公共祖先 u_0，因此该内部顶点至少有 3 个子节点，这与 u 是一般的这一假设相矛盾。

因此，u、$\lambda \odot v$ 中至少有一个不是一般的。由于 $\lambda \odot v$ 是一般的当且仅当 v 是一般的，这与我们的假设矛盾：u、v 是一般的（即三价）树。因此，一般位置对 u、v 之间的热带直线上，不存在具有 5 个或更多子节点的内部顶点的中间树。□

命题 4.5. 一般树对之间的热带直线上，不存在具有两个内部顶点且每个顶点都有三个或更多子节点的中间树。

以下引理将有助于证明该命题。

引理 *4.6.* 设 i、j、k、$l \in [n]$ 为叶节点，u、v 为以 $[n]$ 个叶节点为标记的一般等距树。若 $LCA_u(i, j) = LCA_u(k, l)$，且 $LCA_v(i, j) = LCA_v(k, l)$，则 $LCA_{u \oplus v}(i, j) = LCA_{u \oplus v}(k, l)$。

引理 *4.6* 的证明。等式 $\mathrm{LCA}_u(i,\ j) = \mathrm{LCA}_u(k,\ l)$ 等价于：

$$d_u(i, j) = d_u(k, l) \geq d_u(i, k), d_u(j, l), d_u(i, l), d_u(j, k).$$

在 $u \oplus v$ 中，

$$d_{u \oplus v}(i, j) = \max\{d_u(i, j), d_v(i, j)\} = \max\{d_u(k, l), d_v(k, l)\} = d_{u \oplus v}(k, l)$$

and

$$d_{u \oplus v}(i, j) = \max\{d_u(i, j), d_v(i, j)\} \geq \max\{d_u(x, y), d_v(x, y)\} = d_{u \oplus v}(x, y)$$

对所有 $xy \in \{jk, jl, il, jk\}$。因此，$\mathrm{LCA}_{u \oplus v}(i,\ j) = \mathrm{LCA}_{u \oplus v}(k,\ l)$。□

命题 *4.5*的证明。假设 $w = u \oplus (\lambda \odot v)$ 有两个内部顶点，且每个顶点至少有三个子节点。将第一个高阶内部顶点的三个后代命名为 1、2 和 3，将第二个高阶内部顶点的三个后代命名为 4、5 和 6。那么

$$w_{12} = w_{13} = w_{23}$$

and

$$w_{45} = w_{46} = w_{56}.$$

将 $w_{12} = w_{13} = w_{23}$用 u和 $\lambda \odot v$表示展开后，我们可以看到

$$\max(u_{12}, \lambda + v_{12}) = \max(u_{13}, \lambda + v_{13}) = \max(u_{23}, \lambda + v_{23}). \tag{1}$$

我们将证明这些等式蕴含：对于某个 $ij \subset \{1,2,3\}$，有 $u_{ij} = \lambda + v_{ij}$；对于某个 $k\ell \subset \{4,5,6\}$，有 $u_{k\ell} = \lambda + v_{k\ell}$；这又蕴含 $u_{ij} - v_{ij} = \lambda = u_{k\ell} - v_{k\ell}$，与假设 u，v 是一般位置对矛盾。

超度量的定义告诉我们，u_{12}、u_{13}，u_{23} 中有两个相等且大于第三个；不失一般性，假设 $u_{12} = u_{13} \geq u_{23}$。 u 是一般位置这一事实意味着该不等式是严格的：$u_{12} = u_{13} > u_{23}$。对 $\lambda \odot v$使用类似论证，可得三种可能：

$$\lambda + v_{12} = \lambda + v_{13} > \lambda + v_{23} \tag{2}$$

$$\lambda + v_{12} = \lambda + v_{23} > \lambda + v_{13} \tag{3}$$

$$\lambda + v_{13} = \lambda + v_{23} > \lambda + v_{12} \tag{4}$$

然而，若 (2) 成立，则 $w_{13} = \max\{u_{13}, \lambda + v_{13}\} > \max\{u_{23}, \lambda + v_{23}\} = w_{23}$，这与 $w_{13} = w_{23}$矛盾。因此，(3) 或 (4) 必有一个成立（同时置换 u 和 v 的指标后，它们是等价的），所以不失一般性，假设$\lambda + v_{12} = \lambda + v_{23} > \lambda + v_{13}$。

现在必有 $w_{13} = u_{13}$。否则，$w_{13} = \lambda + v_{13} < \lambda + v_{23} \leq w_{23}$，这与 $w_{13} = w_{23}$矛盾。同理，若 $w_{23} \neq \lambda + v_{23}$，则 $w_{23} = u_{23} < u_{13} \leq w_{13}$，这又与 $w_{23} = w_{13}$矛盾。因此可得：

$$w_{12} = w_{23} = w_{13} = u_{13} = u_{12} > u_{23}, \tag{5}$$

$$w_{12} = w_{13} = w_{23} = \lambda + v_{23} = \lambda + v_{12} > \lambda + v_{13} \tag{6}$$

$$\implies \lambda + v_{12} = u_{12} = u_{13} = \lambda + v_{23} > u_{23}, \lambda + v_{13}. \tag{7}$$

特别是 ， $\lambda + v_{12} = u_{12}$。将 123 替换为 45 后，按照相同的论证 6:

$$\lambda + v_{45} = u_{45} = u_{46} = \lambda + v_{56} > u_{56}, \lambda + v_{46}.$$

这意味着 $u_{12} - v_{12} = \lambda = u_{45} - v_{45}$，因此 $u_{12} - v_{12} = u_{45} - v_{45}$，这与 u，v 是一般位置对的假设相矛盾。□

5. A Very Long Tropical Line

尽管定理 6.1 将平均热带 NNI 数限制为 $O(n(\log n)^4)$，但特定的树对可能具有大得多的热带 NNI 数。图 6 展示了一对含有 $[n]$ 个叶节点的树，它们相差 $n-2$ 次 NNI 移动，但其热带线段上有 $\binom{n-1}{2} \sim n^2$ 次单独的 NNI 移动。分别将下方的树记为 u^n 和 v^n。

定理 5.1。从 v^n 到 u^n的热带线段上，有 $\binom{n-1}{2}$ 次单独的 *NNI* 移动，且没有四簇重排。



FIGURE 6. u^n, v^n 是具有 $[n]$ 个叶节点且热带 NNI 数为 $\binom{n-1}{2}$的一般树对。只需 $n-2$ 次 NNI 移动即可将 u^n 变换为 v^n。

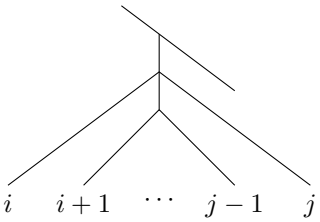


FIGURE 7. 转折点树的一部分 $w = u \oplus (\lambda_{ij} \odot v)$。当 $i+1 = j$ 时，i 和 j之间的子树中没有叶节点，并且 w 是一般位置对。

证明。首先注意，$(\text{LCA}_u(i,j), \text{LCA}_v(i,j)) = \left(u_{\min(i,j)}, v_{\max(i,j)}\right)$。因此，$\#\Pi(u,v) = \binom{n}{2}$。此外，$u$, v 是一对一般位置对，因此 $\#\Lambda(u,v) = \#\Pi(u,v)$，唯一可能的转折点就是定理4.1中列出的那些。

固定 $i < j$。根据定义，u_{ij} 仅取决于 $\min(i,j) = i$，而 v_{ij} 仅取决于 $\max(i,j) = j$。我们将证明，在标量为 $\lambda = u_{ij} - v_{ij}$ 的转折点处，如果 $|i-j| > 1$，则存在一次 NNI 移动；如果 $|i-j| = 1$，则存在一棵一般树。中间树 w如图7所示。

现在令 $k \in [i+1, j-1] \cap [n]$。在树 u 中，我们有

$$u_{ij} = u_{ik} > u_{kj},$$

在树 v 中，我们有

$$v_{ij} = v_{kj} > v_{ik}.$$

回顾一下，我们令 $\lambda = u_{ij} - v_{ij}$。令 $w = u \oplus (\lambda \odot v)$。那么

$$u_{ik} = u_{ij} = v_{ij} + \lambda = v_{jk} + \lambda > v_{ik} + \lambda, u_{jk}$$

这意味着 $w_{ij} = w_{ik} = w_{jk}$，即，$\text{LCA}_w(i,\ j) = \text{LCA}_w(i,\ k) = \text{LCA}_w(j,\ k)$。这意味着当 $|i-j| > 1$ 时，w 是非泛型的。

注意，若 $k \notin [i,j] \cap [n]$，则 $\text{LCA}_w(k,\ i) \neq \text{LCA}_w(i,\ j)$。通过比较$w_{ki}$ 或 w_{jk} 与 w_{ij}即可看出这一点。若 $k < i$，则 $u_{ij} < u_{jk}$，且 $v_{ij} = v_{jk}$，因此

$$w_{jk} = \max(u_{jk}, v_{jk} + \lambda) = \max(u_{jk}, v_{ij} + \lambda) = \max(u_{jk}, u_{ij}) = u_{jk} > w_{ij}.$$

类似的论证表明，对于 $k > j$，$w_{jk} > w_{ij}$。因此，只有在 $|i-j| > 1$时，w 才是非泛型的。我们只需排除四分支簇重排的情形。

如果 k_1, $k_2 \in [i+1, j-1] \cap [n]$，那么

$$u_{ij} = u_{ik_2} > u_{k_1k_2}, \text{ and } v_{ij} = v_{k_1j} > v_{k_1k_2}$$

这意味着 $w_{k_1k_2} < w_{ij}$，因此 w 发生了 NNI，而不是四分支簇重排。 □

警告： 热带 NNI 数不是一种度量，因为它不满足三角不等式。例如，从 v^n 到 u^n 存在一条由热带线段连接而成的路径，沿此路径仅发生 $n-2$ 次 NNI 移动（可能的最少次数），而这少于从 v^n 到 u^n 的热带线段上发生的 NNI 移动次数。

6. Expected Length of the Tropical Line Segment

既然我们已经知道，热带线段上的树拓扑通过单次 NNI 移动或四分支群重排移动发生变化，那么我们可以询问：在两棵随机选取的树之间的热带线段上会发生多少次 NNI 移动。本节的目标是证明定理 6.1，从而界定转折点的数量。随机选取的 $[n]$ 叶节点系统发育树对，意味着这些树的拓扑是从 $[n]$ 叶节点的（有限多个）二叉树拓扑中均匀随机选取的，并且相应的超度量构成一个一般位置对。

定理 6.1。 在 n 个叶节点上随机选取两棵系统发育树，它们之间的热带线段的转折点期望数量为 $O\left(n(\log(n))^4\right)$。

推论 6.2。 在 n 个叶节点上随机选取两棵系统发育树，它们之间的热带 *NNI* 数期望为 $O\left(n(\log(n))^4\right)$。

6.1. 样本空间、记号与概率等价性。 我们希望界定n 个叶节点上的两棵系统发育树之间的热带线段的转折点数量。经过一系列转换，我们将问题归结为研究带有一个标记内部顶点的无标号有根二叉平面树空间。

定义 6.3。 如果为每个顶点的子节点指定一个顺序，则称有根树是**平面的**。对于有根二叉平面树，我们将内部顶点的左子节点和右子节点分别称为左子节点和右子节点，并用 $L(v)$（ $R(v)$）分别表示左（右）子节点的后代。令 $\ell(v) = |L(v)|$ 和 $r(v) = |R(v)|$。

我们将考虑以下树的样本空间：

$$\begin{aligned}
\mathcal{T}_{[n]} &= \{T : T \text{ is a rooted binary tree with } n \text{ labeled leaves}\}.\\
\mathcal{T}_{[n]}^{\text{planar}} &= \{T : T \text{ is a planar rooted binary planar trees with } n \text{ labeled leaves}\}.\\
\mathcal{T}_{[n],v}^{\text{planar}} &= \{(T,v) : T \in \mathcal{T}_{[n]}^{\text{planar}}, v \text{ is an internal vertex of } T\}.\\
\mathcal{T}_{n}^{\text{planar}} &= \{T : T \text{ is a rooted binary planar tree with } n \text{ unlabeled leaves}\}.\\
\mathcal{T}_{n,v}^{\text{planar}} &= \{(T,v) : T \in \mathcal{T}_{n}^{\text{planar}}, v \text{ is an internal vertex of } T\}.\\
\mathcal{T}_{n,v}^{\text{planar}}(a,b) &= \{(T,v) \in \mathcal{T}_{n,v}^{\text{planar}} : \ell(v) = a, r(v) = b\}.
\end{aligned}$$

现在，我们用上述最后一种树的样本空间来重新表述转折点的数量。这个命题清楚地表明，转折点的数量之所以低于朴素上界 $\binom{n}{2}$，仅仅是因为不同的叶节点对（i，j）在 T 中可能具有相同的共同祖先。

定义 6.4. 给定 $A_i, B_i \subseteq [n]$，其中 $A_i \cap B_i = \emptyset$ 和 $|A_i| = a_i$，$|B_i| = b_i$，令 $Q(a_1, b_1, a_2, b_2)$ 表示以下交集非空的概率：

$$A_1 \cap A_2 \neq \emptyset, B_1 \cap B_2 \neq \emptyset.$$

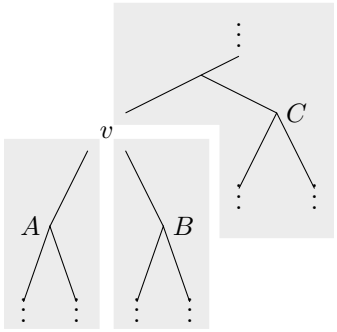


FIGURE 8. 在一个内部顶点处分割一棵二叉树，得到三棵二叉树。

命题 6.5。

$$E\left(\#\Pi(T_1,T_2) \mid T_k \in \mathcal{T}_{[n]}\right) \le 2(n-1)^2 E\left(Q\left(|L(v_1)|,|R(v_1)|,|L(v_2)|,|R(v_2)|\right) \mid {}^{(T_k,v_k)\in\mathcal{T}_{n,v}^{\text{planar}}}_{k=1,2}\right).$$

证明。 根据定义，(*LHS*) 是一对内部顶点 $x_1 \in T_1$、 $x_2 \in T_2$ 分别为其各自的树中某些叶节点 i,j 的最近公共祖先的概率，再乘以配对 (x_1,x_2) 的数量。每棵树 T_k 都有 $n-1$ 个内部顶点，因此共有 $(n-1)^2$ 种可能的配对。因此可得，(LHS) 等于：

$$(n-1)^2 P((x_1,x_2) \in \Pi(T_1,T_2) \mid (T_k,x_k) \in \mathcal{T}_{[n],v}^{\text{planar}}, k=1,2). \tag{8}$$

我们还可以用平面叶标记树来改述（8）：

$$(n-1)^2 P\left({}^{L(x_1)\cap L(x_2)\neq\emptyset, R(x_1)\cap R(x_2)\neq\emptyset}_{\text{or}\quad L(x_1)\cap R(x_2)\neq\emptyset, R(x_1)\cap L(x_2)\neq\emptyset} \mid (T_k,x_k)\in\mathcal{T}_{[n],v}^{\text{planar}}\right). \tag{9}$$

我们可以 仅使用其中一个条件来简化（9）。于是，（9）有如下上界 ：

$$2(n-1)^2 P\left(L(x_1)\cap L(x_2)\neq\emptyset, R(x_1)\cap R(x_2)\neq\emptyset \mid (T_k,x_k)\in\mathcal{T}_{[n],v}^{\text{planar}}\right). \tag{10}$$

最后，我们忽略叶标签并引入变量 Q，于是 (10) = RHS。 □

6.1.1. 树计数。 本节中，我们对树进行计数，为上一小节证明的期望值界给出一个显式公式。

引理 6.6。 [18,例 5.3。12]

$$|\mathcal{T}_n^{\text{planar}}| = \frac{1}{n}\binom{2n-2}{n-1}.$$

引理 6.7。

$$|\mathcal{T}_{n,v}^{\text{planar}}| = \frac{n-1}{n}\binom{2n-2}{n-1}.$$

证明。 具有 n 个叶节点的有根二叉树有 $n-1$ 个内部顶点（包括根），因此

$$|\mathcal{T}_{[n],v}^{\text{planar}}| = (n-1)\cdot\frac{1}{n}\binom{2n-2}{n-1} = \frac{n-1}{n}\binom{2n-2}{n-1}. \quad \square$$

引理 6.8。

$$|\mathcal{T}_{n,v}^{\text{planar}}(\ell(v)=a, r(v)=b)| = \frac{1}{ab}\binom{2a-2}{a-1}\binom{2b-2}{b-1}\binom{2(n-a-b)}{n-a-b}.$$

证明。 我们要计算具有 a 个左后代和 b 个右后代的标记内部顶点 v 的 n 叶二叉平面树的数量。我们可以将三棵二叉平面树 A、B 和 C 粘合起来，构造这样的树；这三棵树分别有 a、b 和 $c = (n-a-b)$ 个叶节点，如图 8 所示。具体来说，我们选取树 C 的一个叶作为标记顶点 v，然后将树 A（树 B）分别接到 v 的左（右）侧作为后代。由此可得：

$$\begin{aligned}|\mathcal{T}^{\text{planar}}_{n,v}(a,b)| &= |\mathcal{T}^{\text{planar}}_{a}||\mathcal{T}^{\text{planar}}_{b}|(n-a-b+1)|\mathcal{T}^{\text{planar}}_{n-a-b+1}| && \text{by Lemma 6.6}\\ &= \frac{1}{a}\binom{2a-2}{a-1}\frac{1}{b}\binom{2b-2}{b-1}\frac{(n-a-b+1)}{n-a-b+1}\binom{2(n-a-b)}{n-a-b}\\ &= \frac{1}{ab}\binom{2a-2}{a-1}\binom{2b-2}{b-1}\binom{2(n-a-b)}{n-a-b}. && \square\end{aligned}$$

6.2. 对总和进行界定。 令 $P(a,\ b;\ n)$ 表示从 $\mathcal{T}_{\text{平面}n,\ v}$ 中选取树-顶点对 $(T,\ v)$ 的概率，其中 $L(v) = a$，$R(v) = b$。在命题 6.5 中，我们将要计算的数量（对具有 $[n]$ 个叶节点的树对取平均的热带 NNI 数）界定为以下期望，该期望用 $P(a_i,\ b_i;\ n)$ 和 $Q(a_1,\ b_1,\ a_2,\ b_2;\ n)$ 表示。

$$S_n := \sum_{\substack{a_i+b_i+c_i=n\\ a_i,b_i\geq 1, c_i\geq 0}} Q(a_1,b_1,a_2,b_2;n)P(a_1,b_1;n)P(a_2,b_2;n) \tag{11}$$

我们将通过界定 $P(a,\ b;\ n)$ 和 $Q(a_1,\ b_1,\ a_2,\ b_2;\ n)$，来界定式 (11)。中心二项式系数的一些界将会派上用场。

命题 6.9 ([19])。

$$\frac{4^n}{\sqrt{\pi\left(n+\frac{1}{3}\right)}} \leq \binom{2n}{n} \leq \frac{4^n}{\sqrt{\pi\left(n+\frac{1}{4}\right)}}$$

引理 6.10。 对于 $n \geq 1$，我们有以下界

$$\frac{4^{n-1}}{n\sqrt{\pi\left(n-\frac{2}{3}\right)}} \leq \left|\mathcal{T}^{\text{planar}}_{n}\right| \leq \frac{4^{n-1}}{n\sqrt{\pi\left(n-\frac{3}{4}\right)}}.$$

$$\frac{4^{n-1}(n-1)}{n\sqrt{\pi\left(n-\frac{2}{3}\right)}} \leq \left|\mathcal{T}^{\text{planar}}_{n,v}\right| \leq \frac{4^{n-1}(n-1)}{n\sqrt{\pi\left(n-\frac{3}{4}\right)}}.$$

证明。 将命题 6.9 中的界代入我们在引理 6.6 中为 $\left|\mathcal{T}^{\text{planar}}_{n}\right|$ 得到的公式，即可得到这些界。

$$\left|\mathcal{T}^{\text{planar}}_{n}\right| = \frac{1}{n}\binom{2n-2}{n-1} \leq \frac{4^{n-1}}{n\sqrt{\pi\left(n-1+\frac{1}{4}\right)}} = \frac{4^{n-1}}{n\sqrt{\pi\left(n-\frac{3}{4}\right)}}$$

$$\left|\mathcal{T}^{\text{planar}}_{n}\right| = \frac{1}{n}\binom{2n-2}{n-1} \geq \frac{4^{n-1}}{n\sqrt{\pi\left(n-1+\frac{1}{3}\right)}} = \frac{4^{n-1}}{n\sqrt{\pi\left(n-\frac{2}{3}\right)}}$$

对于 $\left|\mathcal{T}^{\text{planar}}_{n,\ v}\right|$ 的界可通过类似的论证得到，同时注意到正如我们在引理 6.7 $\left|\mathcal{T}_{\text{平面}n,\ v}\right| = (n-1)\left|\mathcal{T}_{\text{平面}n}\right|$ 中所看到的那样。□

推论 6.11。

$$P(a,b;n) \le \frac{1}{2\pi} \frac{\sqrt{n}}{a\left(a-\frac{3}{4}\right)^{\frac{1}{2}} b\left(b-\frac{3}{4}\right)^{\frac{1}{2}} \left(c+\frac{1}{4}\right)^{\frac{1}{2}}}$$

证明。 我们可以将 $P(a,b;\ n)$ 表示为树-顶点对的数量 $(T,v)\ \in \mathcal{T}_{n\,v}^{\text{planar}}$,与 $\ell(v)\ = a,\ r(v)\ = b$ 之比，分母为 $\mathcal{T}_{\text{平面}n,v}$ 中所有树的数量。

$$P(a,b;n) = \frac{\left|\mathcal{T}_{n,v}^{\text{planar}}(a,b)\right|}{\left|\mathcal{T}_{n,v}^{\text{planar}}\right|} = \frac{\left|\mathcal{T}_{a}^{\text{planar}}\right| \left|\mathcal{T}_{b}^{\text{planar}}\right| (c+1) \left|\mathcal{T}_{c+1}^{\text{planar}}\right|}{\left|\mathcal{T}_{n,v}^{\text{planar}}\right|}$$

应用引　　理 6.10 中的界于上述表达式，可得所声称的不等式　　.

$$P(a,b;n) \le \frac{1}{4\pi} \frac{(n-1)\sqrt{n}}{a\left(a-\frac{3}{4}\right)^{\frac{1}{2}} b\left(b-\frac{3}{4}\right)^{\frac{1}{2}} \left(c+\frac{1}{4}\right)^{\frac{1}{2}} n} \le \frac{1}{2\pi} \frac{\sqrt{n}}{a\left(a-\frac{3}{4}\right)^{\frac{1}{2}} b\left(b-\frac{3}{4}\right)^{\frac{1}{2}} \left(c+\frac{1}{4}\right)^{\frac{1}{2}}}$$

□

引理 6.12。 设 A_1 和 A_2 是 $[n]$的子集，分别从大小为 a_1 和 a_2 的所有子集中均匀随机选取。那么，

$$P(A_1 \cap A_2 \neq \emptyset) \le \frac{a_1 a_2}{n}$$

证明。 $1 \in A_1 \cap A_2$为 $(a_1/n)\cdot(a_2/n)$ 的概率。因此， $[n]$ 中至少有一个元素属于 A_1, A_2 交集的概率至多为 $n(a_1/n)(a_2/n)\ = a_1 a_2/n$。 □

推论 6.13。

$$Q(a_1,b_1,a_2,b_2) \le \min\left(\frac{a_1 a_2}{n}, \frac{b_1 b_2}{n}\right) \le \frac{(a_1 a_2 b_1 b_2)^{1/2}}{n}$$

定义 6.14。

$$\tilde{Q}(a_1,b_1,a_2,b_2;n) := \frac{(a_1 a_2 b_1 b_2)^{1/2}}{n}.$$

$$\tilde{Q}_0(a,b;n) := \left(\frac{ab}{n}\right)^{\frac{1}{2}}.$$

˜ ˜ ˜推论 6.13 表明 $Q(a_1,b_1,a_2,b_2;\ n)\ \le Q(a_1,b_1,a_2,b_2;\ n)$ $= Q_0(a_1,b_1;n)Q_0(a_2,b_2;\ n)$。由于我们尚未掌握某些边界情况中$P(a,b;\ n)$的界，因此还不能计算（11）中的完整总和，但我们可以对一个重要的部分和给出界。

命题 6.15。

$$S_n \le \left(\sum_{\substack{a+b\le n \\ a,b\ge 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)\left(c+\frac{1}{4}\right)^{\frac{1}{2}}} \right)^2. \tag{12}$$

证明。 我们代入推论 6.11 和推论 6.13 中计算出的界，并进行化简。这里的关键思路是，Q 上界的对称性使我们能够将六个变量中的和表示为三个变量中和的平方。

$$\begin{aligned} S_n &= \sum_{\substack{a_i+b_i+c_i=n \\ a_i,b_i\geq 1, c_i\geq 0}} Q(a_1,b_1,a_2,b_2)P(a_1,b_1;n)P(a_2,b_2;n) \\ &\leq \sum_{\substack{a_i+b_i+c_i=n \\ a_i,b_i\geq 1, c_i\geq 0}} \tilde{Q}(a_1,b_1,a_2,b_2;n)P(a_1,b_1;n)P(a_2,b_2;n) \\ &= \left(\sum_{\substack{a+b\leq n \\ a,b\geq 1}} \tilde{Q}_0(a,b;n)P(a,b;n) \right)^2 \end{aligned}$$

现在，我们利用推论 6.11 中关于 $P(a,\ b;\ n)$ 的界，对上面平方中的和的各项进行展开并估计。

$$\begin{aligned} \tilde{Q}_0(a,b;n)P(a,b;n) &\leq \frac{a^{\frac{1}{2}}b^{\frac{1}{2}}}{\sqrt{n}} \frac{\sqrt{n}}{a\left(a-\frac{3}{4}\right)^{\frac{1}{2}} b\left(a-\frac{3}{4}\right)^{\frac{1}{2}}\left(c+\frac{1}{4}\right)^{\frac{1}{2}}} \\ &= \frac{1}{a^{\frac{1}{2}}\left(a-\frac{3}{4}\right)^{\frac{1}{2}} b^{\frac{1}{2}}\left(b-\frac{3}{4}\right)^{\frac{1}{2}}\left(c+\frac{1}{4}\right)^{\frac{1}{2}}} \end{aligned}$$

应用算术-几何平均不等式可以进一步简化：

$$\frac{1}{a^{\frac{1}{2}}\left(a-\frac{3}{4}\right)^{\frac{1}{2}}} \leq \frac{1}{a-\frac{1}{2}}$$

因此，我们得到关于 S_n 的以下界。

$$S_n \leq \left(\sum_{\substack{a+b\leq n \\ a,b\geq 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)\left(c+\frac{1}{4}\right)^{\frac{1}{2}}} \right)^2.$$

□

以下引理将有助于完成命题 6.15 右侧总和的计算。

引理 6.16。

$$\sum_{\substack{a+b=d \\ a,b\geq 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)} \leq \frac{2(2+\log\left(d-\frac{3}{2}\right))}{d-1}.$$

证明。 首先，当 $a+b=d$ 时，我们有：

$$\frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)} = \frac{1}{d-1}\left(\frac{1}{a-\frac{1}{2}} + \frac{1}{b-\frac{1}{2}}\right).$$

此外，

$$\sum_{a=1}^{d-1} \frac{1}{a-\frac{1}{2}} = 2 + \frac{1}{2-\frac{1}{2}} + \cdots + \frac{1}{d-\frac{3}{2}} \leq 2 + \log\left(d-\frac{3}{2}\right).$$

由此可得

$$\sum_{\substack{a+b=d\\a,b\geq 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)} = \frac{1}{d-1}\sum_{\substack{a+b=d\\a,b\geq 1}} \frac{1}{a-\frac{1}{2}} + \frac{1}{b-\frac{1}{2}} = \frac{2}{d-1}\sum_{a=1}^{d-1}\frac{1}{a-\frac{1}{2}} \leq \frac{2(2+\log\left(d-\frac{3}{2}\right))}{d-1}.$$

□

回顾以下不等式。

引理 6.17 (切比雪夫求和不等式，[20])。若 a_n 是递减数列，且 b_n 是递增数列，则

$$\frac{1}{n}\sum_{k=1}^{n} a_k b_k \leq \left(\frac{1}{n}\sum_{k=1}^{n} a_k\right)\left(\frac{1}{n}\sum_{k=1}^{n} b_k\right)$$

命题 6.18。$\sqrt{S_n}$ 是 $O\left((\log(n))^2/n^{\frac{1}{2}}\right)$。

证明。我们首先应用引理 6.16 对平方括号内的和进行估计。

$$\begin{aligned}\sum_{\substack{a+b\leq n\\a,b\geq 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)\left(c+\frac{1}{4}\right)^{\frac{1}{2}}} &= \sum_{c=0}^{n-2}\left(c+\frac{1}{4}\right)^{-\frac{1}{2}} \sum_{\substack{a+b=n-c\\a,b\geq 1}} \frac{1}{\left(a-\frac{1}{2}\right)\left(b-\frac{1}{2}\right)}\\ &\leq \sum_{c=0}^{n-2}\frac{2\left(2+\log\left(n-c-\frac{3}{2}\right)\right)}{\left(c+\frac{1}{4}\right)^{\frac{1}{2}}(n-c-1)}\\ &\leq \sum_{c=0}^{n-2}\frac{2\left(2+\log\left(n-c-1\right)\right)}{\left(c+\frac{1}{4}\right)^{\frac{1}{2}}(n-c-1)} \qquad (*)\end{aligned}$$

数列 $\left(c+\frac{1}{4}\right)^{-\frac{1}{2}}$ 关于 c 递减，而数列 $(2+\log(n-c-1))/(n-c-1)$ 在区间 $\left[0, n-\frac{1}{e}-1\right]$ 上关于 c 递增。因此，我们可以应用切比雪夫求和不等式 [20] 来估计 $(*)$ 中的和。

$$\begin{aligned}\sum_{c=0}^{n-2}\frac{1}{\left(c+\frac{1}{4}\right)^{\frac{1}{2}}}\frac{2\left(2+\log\left(n-c-1\right)\right)}{(n-c-1)} &\leq \frac{1}{n-1}\left(\sum_{c=0}^{n-2}\left(c+\frac{1}{4}\right)^{-\frac{1}{2}}\right)\left(\sum_{c=0}^{n-2}\frac{2(2+\log(n-c-1))}{(n-c-1)}\right)\\ &\sim \frac{1}{n-1}\left(\sum_{c=0}^{n-2}\left(c+\frac{1}{4}\right)^{-\frac{1}{2}}\right)\left(\sum_{c=0}^{n-2}\frac{\log(n-c-1)}{(n-c-1)}\right)\\ &\sim \frac{n^{\frac{1}{2}}(\log(n))^2}{n-1} \sim \frac{(\log(n))^2}{n^{\frac{1}{2}}}\end{aligned}$$

在第三行中，我们通过将这些和视为单调函数的左黎曼和或右黎曼和，使用由此得到的渐近估计。

$$\int_0^{n-1}\left(x+\frac{1}{4}\right)^{-\frac{1}{2}}dx \leq \sum_{c=0}^{n-2}\left(c+\frac{1}{4}\right)^{-\frac{1}{2}} \leq 2+\int_0^{n-2}\left(x+\frac{1}{4}\right)^{-\frac{1}{2}}dx \sim n^{\frac{1}{2}} \tag{13}$$

$$\int_{-1}^{n-2}\frac{\log(n-x-1)}{n-x-1}dx \leq \sum_{c=0}^{n-2}\frac{\log(n-c-1)}{(n-c-1)} \leq \int_0^{n-2}\frac{\log(n-x-1)}{n-x-1}dx \sim (\log(n))^2 \tag{14}$$

□

定理 *6.1* 的证明。本节的结果证明了以下不等式：

$$E\left(\#\Pi(T_1,T_2) \mid T_1,T_2 \in \mathcal{T}_{[n]}\right) \leq 2(n-1)^2 S_n.$$

此外，根据命题 6.18，S_n 是 $O\left((\log(n))^4/n\right)$。由此可得，$(n-1)^2 S_n$ 是$O\left(n(\log(n))^4\right)$，正如所声称的那样。□

7. Conclusion

本文分类了两棵树之间的热带线段上可能发生的树拓扑变化，这是在超度量空间中解释统计学习模型的重要一步。我们还给出了两棵树之间的热带线段上拓扑变化平均次数的一个上界，这使我们能够粗略比较 NNI 图中的距离与热带线段上发生的树拓扑变化次数。不过，这个上界很可能并不紧。具体而言，我们猜想推论 6.13 中对 $Q(a_1, b_1, a_2, b_2; n)$ 所采用的上界可以改进。尽管如此，这是比较 Tree$_n$ 不同实现中的测地线上树拓扑变化的一步。

Acknowledgements

我衷心感谢我的导师 David Speyer 和 Ruriko Yoshida给予我的指导与支持。我还要感谢匿名审稿人提出的详尽意见，以及 Anna Brosowsky 和 Sameer Kailasa 与我进行的有益交流。本材料基于以下资助支持的研究：美国国家科学基金会研究生研究奖学金（资助编号 DGE-1841052），以及美国国家科学基金会（资助编号 1855135）。

References

[1] Louis J. Billera、Susan P. Holmes 和 Karen Vogtmann。系统发育树空间的几何。*Adv. in Appl. Math.*，27(4)：733–767，2001。[2] David Speyer 和 Bernd Sturmfels。热带格拉斯曼流形。*Adv. Geom.*，4(3)：389–411，2004。[3] Federico Ardila 和 Caroline J. Klivans。拟阵的伯格曼复形与系统发育树。*J.Combin. Theory Ser. B*，96(1)：38–49，2006。[4] Junhyong Kim。切片高维橙子：系统发育估计的几何。*Molecularphylogenetics and evolution*，17(1)：58–75，2000。[5] Vincent Moulton 和 Mike Steel。剥离系统发育“橙子”。*Advances in Applied Mathematics*，33(4)：710–727，2004。[6] Jonna Gill、Svante Linusson、Vincent Moulton 和 Mike Steel。系统发育树边乘积空间的正则分解。*Advances in Applied Mathematics*，41(2)：158–176，2008。[7] M. K. Garba、T. M. W. Nye、J. Lueg 和 S. F. Huckemann。系统发育树的信息几何。*J. Math. Biol.*，82(3)：第19号论文，39页，2021。[8] Bo Lin 和 Ruriko Yoshida。热带费马-韦伯点。*SIAM J. Discrete Math.*，32(2)：1229–1245，2018。[9] Ruriko Yoshida、Leon Zhang 和 Xu Zhang。热带主成分分析及其在系统发育学中的应用。*Bull. Math. Biol.*，81(2)：568–597，2019。[10] Ruriko Yoshida 和 SHELBY COX。热带线段上的树拓扑。*Vietnam Journal ofMathematics*，第1–25页，2022。[11] Bhaskar DasGupta、Xin He、Ming Li、John Tromp 和 Louxin Zhang。最近邻交换及相关距离，第 1402–1405 页。Springer New York，New York，NY，2016。[12] Mike Develin 和 Bernd Sturmfels。热带凸性。*Documenta Mathematica*，9：1–27，2004。

[13] Diane Maclagan 和 BerndSturmfels。热带几何导论，数学研究生教程丛书第161卷。美国数学学会，普罗维登斯，罗得岛州，2015。[14] G William Moore、M Goodman 和 J Barnabas。从加法假设的角度出发，采用迭代方法解决由分子数据集提出的树状图问题。理论生物学杂志，38(3)：423–457，1973。[15] David F Robinson。三价标记树的比较。组合理论杂志*B*辑，11(2)：105–119，1971。[16] Bhaskar DasGupta、Xin He、Tao Jiang、Ming Li、John Tromp 和 Louxin Zhang。系统发育树之间的距离。载于*SODA*，第97卷，第427–436页。Citeseer，1997。[17] Ming Li、John Tromp 和 Louxin Zhang。关于最近邻交换距离的一些注记。载于国际计算与组合学会议，第343–351页。施普林格，1996。[18] Richard P. Stanley。计数组合学。第*2*卷，剑桥高等数学研究丛书第62卷。剑桥大学出版社，剑桥，1999。Gian-Carlo Rota 作序，Sergey Fomin 撰写附录1。[19] robjohn（ https://math.stackexchange.com/users/13854/robjohn）。初等中心二项式系数估计。数学 Stack Exchange。URL：https://math.stackexchange.com/q/932509（版本：2019-10-07）。[20] G. H. Hardy、J. E. Littlewood 和 G. P´olya。不等式。剑桥数学文库。剑桥大学出版社，剑桥，1988。1952年版本的重印本。