arXiv:2603.08542v1 [math.ST] 9 Mar 2026

植入匹配的贝叶斯推断：局部后验近似与无限体积极限

Zhou Fan* Timothy L. H. Wee† Kaylee Y. Yang‡

摘要

我们研究在临界标度$\|X_i - Y_{\pi^*(i)}\|_2 \asymp n^{-1/d}$下，对两个相关随机点集 $\{X_i\}_{i=1}^n$ 与 $\{Y_i\}_{i=1}^n$ 之间未知匹配 π^* 进行贝叶斯推断的问题，分别考察所有点均可观测的精确匹配模型，以及部分点可能缺失的部分匹配模型。在最简单的 $d = 1$设定下，本文探讨以下问题：（1）匹配上的后验分布是否可以通过局部算法近似；（2）该后验分布的边缘统计量是否在 $n \to \infty$ 时具有定义良好的极限。对于部分匹配，我们对两个问题都给出了肯定回答：当 n较大时，会出现关联衰减。对于精确匹配，我们证明，只有在对点进行全局排序之后，后验分布才能在局部得到近似；此外，要定义边缘统计量的大-n 极限，需要基于流的概念，对数据的泊松点过程极限中的点进行谨慎的索引。将这些结果推广到 $d \geq 2$维仍是一个开放问题。

目录

1 引言 **2**

1.1 结果概述 .3 1.2 相关工作 .4

2 主要结果 **5**

2.1 模型.5 2.2 部分匹配的结果 .7 2.2.1 后验边缘分布的局部计算 .7 2.2.2 后验分布的渐近性 .8 2.3 精确匹配的结果.10 2.3.1 后验边缘分布的局部计算 .10 2.3.2 后验分布的渐近性 .11

3 定理 2.9 和 2.11 的证明概述 **12**

A 命题 2.1 和 2.2 的证明 **14**

作者按字母顺序排列。
*耶鲁大学统计与数据科学系，zhou.fan@yale.edu
†佐治亚理工学院数学学院，timothy.wee@gatech.edu
‡耶鲁大学统计与数据科学系，yingxi.yang@yale.edu

B 精确匹配的证明 **16**
gB.1 定理 2.9 的证明.17
B.1.1 平均势 .17B.1.2 点过程的正则性与吉布斯测度的局部性 21B.1.3 相关性衰减.29
B.2 定理 2.11 的证明 .33
B.2.1 弱收敛到极限点过程 .33B.2.2 算法 2 的局部近似 .37B.2.3 命题 2.12 的证明 . 41B.2.4 命题 2.14 的证明 .42

C 部分匹配的证明 **44**
C.1 定理 2.4 的证明.44C.1.1 平均势 .44C.1.2 点过程的正则性与吉布斯测度的局部性. . . . . 47C.1.3 相关性衰减.51C.2 定理 2.7 的证明.52C.2.1 到极限点过程的局部弱收敛 .52C.2.2 命题 2.8 的证明 .54

1 引言

在科学与工程的许多领域中，识别两组对象之间未知的匹配或对齐是一项基础任务；相关应用包括单细胞基因组学中的批次校正与数据整合 [37], 、粒子追踪与轨迹推断 [27], 、图像匹配[25], 、记录链接与数据库对齐 [32], 以及网络对齐 [5]。针对这些应用，已经开发出许多估计潜在匹配的方法，通常结合局部或近邻特征化 [21], 、用于降维的主成分分析/典型相关分析/非负矩阵分解 [35, 22], 、线性指派或最大后验估计 [10, 40], 和/或最优传输 [33, 17]等思路。然而，尽管所有这类方法都会给出一个“最可能”匹配的估计，大多数方法并未对底层匹配的不确定性提供经过校准的量化结果。

为了更好地从理论上理解这些问题中不确定性量化在算法上的可行性，本文引入并研究了两个用于匹配实数集 $\mathbb{R}^d$ 中点集的简单贝叶斯模型。在一种精确匹配 模型中，我们假设完整观测到两个点集 $X_1, \dots, X_n \in \mathbb{R}^d$ 和 $Y_1, \dots, Y_n \in \mathbb{R}^d$，它们通过一个潜在双射$\pi^*$: $[n] \to [n]$ 关联，其形式为

$$X_i = \bar{X}_{\pi^*(i)} \qquad \text{and} \qquad Y_i = \bar{Y}_i \tag{1}$$

其中（X_1， Y_1）.....，（X_n， Y_n）是独立同分布的数据对。我们还将考虑一种部分匹配 模型，其中 $\{X_1, \dots, X_n\}$ 和 $\{Y_1, \dots, Y_n\}$ 的子集可能独立地未被观测到。我们将从统计和计算两个角度，研究在所有候选匹配上采用均匀先验分布时，对 π^* 进行后验推断（等价地，对给定的观测值X_1，.....， X_n 和 Y_1，.....， Y_n进行均匀随机排序）。

在高斯设定下， $\bar{X}_i \sim \mathcal{N}(0, I_d)$与 $\bar{Y}_i - \bar{X}_i \sim \mathcal{N}(0, \sigma^2 I_d)$的精确匹配已得到深入研究, [13, 23], 对 π^*的完全恢复和近乎完全恢复的信息论极限，以及最大似然估计的估计误差，给出了精确刻画。在最简单的固定维度 d设定下，这些工作的结果表明， π^*实现近乎完全恢复的阈值为 $\sigma \ll n^{-1/d}$。在本文中，我们将研究临界缩放区间 $\|\bar{Y}_i - \bar{X}_i\|_2 \asymp n^{-1/d}$，其中每个

点 X_i 通常会与多于一个点 Y_j 匹配，并且当 $n \to \infty$时，其后验概率不会趋于零。在这一设定下，我们提出两个基本问题：

1. 从算法角度看，能否使用一种高效的局部算法来近似后验分布 $P(\pi \mid X, Y)$ 的“边缘分布”——也就是说，对每个点 X_i 或 Y_j，求其正确匹配对象的后验概率向量——而该算法只需观察 X_i 或 Y_j 周围最近的 $O(1)$ 个点？

2. 从统计角度看，当 $n \to \infty$ 时，能否刻画这些后验边缘分布的极限经验分布，从而理解 π^* 的贝叶斯最优推断的性质，例如贝叶斯估计中未匹配点所占比例，或用于不确定性量化的贝叶斯可信集平均大小？

显然，这些问题与关联衰减的性质以及后验测度 $P(\pi \mid X, Y)$ 是否存在无限体积极限有关。我们将针对一般维数 d 引入模型，在这一设定下，这些问题自然会出现；不过，本文的结果仅涉及维数 $d = 1$ （此时答案已有些微妙，尤其是对于精确匹配模型）。下文将讨论把我们的分析推广到维数 $d \geq 2$ 时可能遇到的其他挑战。

1.1 结果概要

部分匹配。在部分匹配模型中，我们假设对于潜在样本量 N，每个点 $X_1, \ldots, X_N$ 和 $Y_1, \ldots, Y_N$ 均以固定概率 p 独立随机地被观测到。我们研究部分双射 π^* 上的后验分布，这些部分双射将每个观测到的点 $X_1, \ldots, X_{N_X}$ 与另一集合中对应的点 $Y_1, \ldots, Y_{N_Y}$ 或空标签 $\varnothing$ 相匹配。与精确匹配模型相比，这对 π^* 施加了较宽松的双射约束。我们证明，对于维数 $d = 1$，在关于 (X_i, Y_i) 的联合分布满足温和假设的条件下，其中$|\bar{Y}_i - \bar{X}_i| \asymp n^{-1/d} = n^{-1}$，后验分布 $P(\pi \mid X, Y)$ 在适当意义下呈现关联衰减。

这使我们能够对问题 1 给出肯定的回答：对于每个点 X_i，只需计算以 X_i 为中心、大小为 $O(n^{-1})$ 的窗口内各点之间部分双射的局部后验分布，就可以近似得到后验边缘匹配概率。当 $n \to \infty$ 时，后验边缘分布的经验分布弱收敛到一个自然的无限体积极限，该极限定义在观测数据的泊松点过程极限上，也由此简单地回答了问题 2。

精确匹配。在所有点 $X_1, \ldots, X_n$ 和 $Y_1, \ldots, Y_n$ 均已观测到、且要求推断精确双射的精确匹配模型中，我们对问题 1 给出了部分肯定的回答：后验边缘分布可以通过一种算法很好地近似，该算法首先对 $\{X_i\}_{i=1}^n$ 和 $\{Y_i\}_{i=1}^n$ 执行全局排序步骤，然后计算排序索引相同的 $O(1)$ 个点之间匹配的局部后验分布。我们的分析表明，一种更简单直接的方法并不可行：例如，计算 X_i 附近最近的 k 个点与其最近邻 Y_j 附近最近的 k 个点之间双射的局部后验分布，即使在极限 $k \to \infty$ 下，也无法正确近似后验分布律。

这种全局排序的必要性与无限体积极限中吉布斯测度的关联衰减不完全有关。在一个相关的模型中，将 $\mathbb{R}$ 上的可数无限点集匹配到其自身，[8]证明存在一个由极端无限体积吉布斯测度组成的无限集，每个测度对应于一个守恒量的整数值；该守恒量称为流，并会导致长程依赖。在我们的模型中，我们定义了相对于 π^* （参见定义2.13）的类似流的概念，并通过证明后验分布边缘

分布收敛到一个无限体积极限，为相对于 π^* 的、定义在极限泊松点过程上的匹配，从而回答问题 2。

因此，对于部分匹配模型和精确匹配模型，这些问题的答案在维数为 $d = 1$ 时就已不同；在精确匹配模型中，守恒流变量会阻碍相关性衰减。将我们的分析扩展到维数为 $d \geq 2$ 时，还将面临其他挑战：

- 对于精确匹配，当 $d \geq 2$时，点的排序或排列并没有直接的对应概念，这引出了一个问题：如何将全局信息用作局部近似后验边缘分布的输入，从而取得成功。

- 对于部分匹配和精确匹配，我们分析中的边界变量 Γ_x（π）在维数为 $d = 1$时构成马尔可夫链，而在维数为 $d \geq 2$时则构成更一般的马尔可夫随机场；这引出了一个问题：吉布斯测度的长程相关性是否会出现其他相变现象。

我们认为这些都是有趣的开放问题，将留待未来研究。

1.2 相关工作

植入匹配的统计推断。 精确匹配模型中的最大似然估计或最大后验估计（1）可以理解为完全加权二分图上的一个置换优化问题，其中随机邻接矩阵的元素是匹配 X_i 到 Y_j 的代价。关于这一问题的变体，其中边权重为独立的情形也已在 [28, 18],中得到研究，并取得了极为精确的结果。

在粒子追踪背景下研究我们所设定的相关边权重的早期工作见于 [11], ，随后 [13, 23]针对具有不同噪声方差 σ^2 和维数 d的高斯模型，给出了完全恢复和近乎完美恢复的精确结果。 [30]研究了带有打乱的响应变量的线性回归中的置换恢复。 [14],还研究了部分匹配模型的变体，其动机来自具有不完全重叠的数据库对齐问题。

一个相关问题是几何随机图匹配：在该问题中，人们只能观测数据点之间的成对距离或相似度 k（X_i， X_j）以及 k（Y_i， Y_j）。关于这一领域的相关文献和近期结果，可参见 [38] 。

置换上的吉布斯测度。经典的Mallows模型属于这一框架，至今仍是一个活跃的研究领域 [29, 39, 26, 2]。 [9]研究了另一种图子型尺度极限下，匹配模型的吉布斯测度的渐近行为和大偏差。

我们的模型缩放方式和后验测度与数理物理中玻色–爱因斯坦凝聚背景下研究的空间随机置换模型更为密切相关[19, 36, 7]。在维数为 $d = 1$时， [8] 对具有凸噪声势 $V(\cdot)$ 的一大类此类模型的无限体积吉布斯测度进行了刻画，表明极端吉布斯态与一个称为流的整数参数相对应。我们的精确匹配模型的无限体积极限可以看作将这一设定扩展到非凸噪声势和两个不同点集的情形。据我们所知，在维数为 $d \geq 2$ 时，类似的完整刻画似乎仍未解决。对于正则格上的置换，在周期边界条件下存在无限体积极限由 [6] 证明，在高温下则由[4]证明。 [20]研究了其中一些模型的带状结构。

2 主要结果

2.1 模型

数据模型。 令 $\Lambda\colon [0,1]^d \to [0,\infty)$ 为 $[0,1]^d$ 上的概率密度，并令 $V\colon \mathbb{R}^d \to \mathbb{R}$ 为满足 $V(\varepsilon) = V(-\varepsilon)$ 的噪声势函数，其中 $\Lambda(\cdot)$ 和 $V(\cdot)$ 均为固定值，且不依赖于 n。对于每个 n，考虑一对 (X_0, Y_0) 的可交换联合密度 $\in [0,1]^d \times [0,1]^d$，其定义为

$$p_n(x,y) = \frac{1}{\mathcal{Z}_n}\sqrt{\Lambda(x)\Lambda(y)}\exp\Big(-V(n^{1/d}(x-y))\Big) \tag{2}$$
$$\mathcal{Z}_n = \int_{[0,1]^d}\int_{[0,1]^d}\sqrt{\Lambda(x)\Lambda(y)}\exp\Big(-V(n^{1/d}(x-y))\Big)\,\mathrm{d}x\,\mathrm{d}y.$$

高斯噪声的情形对应于二次势函数 $V(\cdot)$（除了在 $[0,1]^d$ 边界处进行不重要的截断）。我们始终采用这样的缩放约定：$\int_{\mathbb{R}^d}\exp(-V(\varepsilon))\,\mathrm{d}\varepsilon = 1$，因此 $q(\varepsilon) = \exp(-V(\varepsilon))$ 在 $\mathbb{R}^d$ 上定义了一个概率密度。记 $\varepsilon = n^{1/d}(y-x)$，并令

$$p_n(x) = \int_{[0,1]^d} p_n(x,y)dy, \qquad p_n(\varepsilon \mid x) = \frac{n^{-1}p_n(x, x+n^{-1/d}\varepsilon)}{p_n(x)}$$

为 X_0 的边缘密度，以及在此模型中给定 X_0 时 $n^{1/d}(Y_0 - X_0)$ 的条件密度。下面的基本命题表明，当 $n\to\infty$ 时，$\Lambda(x)$ 和 $q(\varepsilon) = \exp(-V(\varepsilon))$ 可分别视为这些边缘密度和条件密度的极限。

命题 2.1. 假设 Λ 在 $[0,1]^d$ 上连续且有正下界，并且 $V(\cdot)$ 在 $\mathbb{R}^d$ 上有下界。那么当 $n\to\infty$ 时：

(a) 对每个固定的 $x \in (0,1)^d$，$p_n(x) \to \Lambda(x)$。

(b) 对每个固定的 $x \in (0,1)^d$ 和有界域 $S \subset \mathbb{R}^d$，上确界 $\sup_{\varepsilon\in S}|p_n(\varepsilon \mid x) - q(\varepsilon)| \to 0$。

(c) $n\cdot\mathcal{Z}_n \to 1$。

精确匹配模型。 令 $(X_1,Y_1),\dots,(X_n,Y_n)$ 为具有分布 (X_i,Y_i) 的 n 独立同分布对，其中 $\sim p_n(x,y)$。令 $\pi^*\colon [n]\to[n]$ 为一个潜在双射，其先验分布为均匀先验分布

$$\pi^* \sim \mathsf{Unif}(\mathbb{S}_n)$$

其中 $\mathbb{S}_n$ 表示所有双射的集合 $\pi\colon [n]\to[n]$（即 $[n]$ 上所有置换构成的对称群）。我们观测到数据 $X = (X_i)_{i=1}^n$ 和 $Y = (Y_i)_{i=1}^n$，其中

$$X_i = \bar{X}_{\pi^*(i)} \qquad \text{and} \qquad Y_i = \bar{Y}_i. \tag{3}$$

因此，潜在对应关系为 $X_i \leftrightarrow Y_{\pi^*(i)}$。我们的目标是根据给定 (X, Y) 的后验分布律，对 π^* 进行推断，其形式为

$$P(\pi \mid X,Y) = \frac{1}{\mathcal{Z}_{\text{exact}}}\exp\Big(-\sum_{i=1}^n V(n^{1/d}(X_i - Y_{\pi(i)}))\Big) \tag{4}$$

其中 $\mathcal{Z}_{\text{exact}} = \sum_{\pi\in\mathbb{S}_n}\exp(-\sum_{i=1}^n V(n^{1/d}(X_i - Y_{\pi(i)})))$。

我们指出，对 π^* 采用均匀先验分布的假设是自然的；这相当于假设观测数据 $X = (X_i)_{i=1}^n$ 和 $Y = (Y_i)_{i=1}^n$ 的排列是均匀随机的，或者等价地说，观测到的是 $\mathbb{R}^d$ 上的无序点集 $\{X_1,\dots,X_n\}$ 和 $\{Y_1,\dots,Y_n\}$。

部分匹配模型。我们将这一框架推广到允许存在未匹配点的部分匹配情形。给定 $N \geq 1$，令 $(X_1, Y_1) \ldots, (X_N, Y_N)$ 为 N 独立同分布对，其分布为 $(X_i, Y_i) \sim p_n(x, y)$。（这里 N 将是随机的，而 n 是一个模型参数，其大小与 $\mathbb{E}[N]$ 成正比。）每个值 X_i 或 Y_i 均以概率 $p \in (0, 1)$ 独立地被观测到。因此，令 $\mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing \subseteq [N]$ 分别表示如下指标集：其中两个 $\{X_i, Y_i\}$ 均被观测到、仅 X_i 被观测到、仅 Y_i 被观测到，或 $\{X_i, Y_i\}$ 均未被观测到。对于每个 $i = 1, \ldots, N$，我们有

$$i \in \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing \quad \text{with probability} \quad p^2, p(1-p), p(1-p), (1-p)^2 \text{ respectively.}$$

令 $N_X = |\mathcal{S}_{XY}| + |\mathcal{S}_X|$ 表示 X_i 的观测值数量，并类似地令 $N_Y = |\mathcal{S}_{XY}| + |\mathcal{S}_Y|$。令

$$\pi_X^* : [N_X] \to \mathcal{S}_{XY} \cup \mathcal{S}_X \qquad \text{and} \qquad \pi_Y^* : [N_Y] \to \mathcal{S}_{XY} \cup \mathcal{S}_Y \tag{5}$$

为两个潜在双射映射。我们观测到数据

$$(N_X, N_Y), \quad X_i = \bar{X}_{\pi_X^*(i)} \text{ for } i = 1, \ldots, N_X, \quad \text{and} \quad Y_j = \bar{Y}_{\pi_Y^*(j)} \text{ for } j = 1, \ldots, N_Y \tag{6}$$

我们用 $\mathrm{dom}(\pi^*) = (\pi_X^*)^{-1}(\mathcal{S}_{XY}) \subseteq [N_X]$ 和 $\mathrm{range}(\pi^*) = (\pi_Y^*)^{-1}(\mathcal{S}_{XY}) \subseteq [N_Y]$ 表示作为索引集、代表相匹配观测数据值的集合，并用

$$\pi^* : \mathrm{dom}(\pi^*) \to \mathrm{range}(\pi^*), \qquad \pi^*(i) = (\pi_Y^*)^{-1}(\pi_X^*(i))$$

表示这些匹配点之间的潜在双射 $X_i \leftrightarrow Y_{\pi^*(i)}$。我们约定 π^* 隐含地包含其定义域和值域的信息，并记作

$$\pi^*(i) = \varnothing \text{ if } i \in [N_X] \setminus \mathrm{dom}(\pi^*), \quad (\pi^*)^{-1}(j) = \varnothing \text{ if } j \in [N_Y] \setminus \mathrm{range}(\pi^*).$$

我们的目标是基于 $X = (X_i)_{i=1}^{N_X}$ 和 $Y = (Y_j)_{j=1}^{N_Y}$ 对 π^* 进行推断。

给定样本量参数 n，我们考虑关于 N 的一个特定先验分布，

$$N \sim \mathrm{Poisson}\left((1-p)^{-2} \mathcal{Z}_n^{-1}\right) \tag{7}$$

其中，$\mathcal{Z}_n$ 是密度 $p_n(x, y)$ 的归一化常数。根据命题 2.1(c)，当 n 较大时，它近似于先验 $N \sim$ 泊松$(n/(1-p)^2)$。给定 $[N]$ 的划分 $(\mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing)$ 后，我们考虑在所有可能的双射映射 (5) 上，为 π_X^*, π_Y^* 采用独立均匀先验；这同样编码了对观测数据 $(X_i)_{i=1}^{N_X}$ 和 $(Y_j)_{j=1}^{N_Y}$ 采用均匀随机排序的假设。

先验（7）对 N 的这种略显不同寻常的选择，缘于其能使 $\pi^* = (\pi_Y^*)^{-1} \circ \pi_X^*$ 的诱导后验分布具有以下特别简单的形式。

命题 *2.2.* 令 $\mathbb{S}_n$ 为从 $[N_X]$ 到 $[N_Y]$ 的所有部分双射 π 所构成的空间，即从子集 $\mathrm{dom}(\pi) \subseteq [N_X]$ 到子集 $\mathrm{range}(\pi) \subseteq [N_Y]$ 的双射映射，且两者基数相等。若 $i \notin \mathrm{dom}(\pi)$，记 $\pi(i) = \varnothing$；若 $j \notin \mathrm{range}(\pi)$，记 $\pi^{-1}(j) = \varnothing$。令 $U_n(x) = \log p_n(x)\sqrt{\Lambda(x)}$，并定义其上的哈密顿量 $\mathbb{S}_n$

$$H_n(\pi \mid X, Y) = \sum_{i \in [N_X] : \pi(i) \neq \varnothing} V(n^{1/d}(X_i - Y_{\pi(i)})) - \sum_{i \in [N_X] : \pi(i) = \varnothing} U_n(X_i) - \sum_{j \in [N_Y] : \pi^{-1}(j) = \varnothing} U_n(Y_j). \tag{8}$$

那么，π^* 的后验分布形式为

$$P(\pi \mid X, Y) = \frac{1}{\mathcal{Z}_{partial}} \exp\Big(-H_n(\pi \mid X, Y)\Big), \quad \mathcal{Z}_{partial} = \sum_{\pi \in \mathbb{S}_n} \exp\Big(-H_n(\pi \mid X, Y)\Big).$$

算法 1： 部分匹配后验分布的局部近似

for $i \leftarrow 1$ **to** N_X **do**

Let $x_-^{(i)} = \max(0, \frac{\lfloor nX_i \rfloor - M}{n})$ and $x_+^{(i)} = \min(1, \frac{\lfloor nX_i \rfloor + M}{n})$.

Let $\mathbb{S}^{(i)}$ be the space of all partial bijections π from $S_X^{(i)} := \{j : X_j \in [x_-^{(i)}, x_+^{(i)}]\}$ to $S_Y^{(i)} := \{j : Y_j \in [x_-^{(i)}, x_+^{(i)}]\}$ (i.e. bijections between subsets of equal cardinality).

Compute the local posterior law over $\pi \in \mathbb{S}^{(i)}$ given by $P^{(i)}(\pi \mid X, Y) = \frac{1}{\mathcal{Z}^{(i)}} \exp(-H_n(\pi \mid X, Y))$, $\mathcal{Z}^{(i)} = \sum_{\pi \in \mathbb{S}^{(i)}} \exp(-H_n(\pi \mid X, Y))$ where $H_n(\cdot)$ denotes the restriction of the Hamiltonian (8) to $S_X^{(i)}, S_Y^{(i)}$.

Let

$$\widehat{P}_i(j) = \begin{cases} P^{(i)}(\pi(i) = j \mid X, Y) & \text{if } Y_j \in [x_-^{(i)}, x_+^{(i)}], \\ 0 & \text{otherwise,} \end{cases} \qquad \widehat{P}_i(\varnothing) = P^{(i)}(\pi(i) = \varnothing \mid X, Y).$$

end

return $\widehat{P}_1, \ldots, \widehat{P}_{N_X}$

2.2 部分匹配的结果

在本文余下部分　本文固定 $d = 1$，并假设 $\Lambda(\cdot)$和 $V(\cdot)$满足以下性质　.

假设 2.3。在维数 $d = 1$:

(i)（ Λ的性质。）　Λ 是定义在 $[0, 1]$, 上的连续概率密度函数，并且存在常数 $\Lambda_{\min}$，$\Lambda_{\max} > 0$，使得对于所有 $x \in [0, 1]$，$\Lambda_{\min} \le \Lambda(x) \le \Lambda_{\max}$。

(ii)（ V的性质。）对于所有 $\varepsilon \in \mathbb{R}$，$V(\varepsilon) = V(-\varepsilon)$，并且存在常数 C，c，$\delta > 0$、$V_{\min} \in \mathbb{R}$以及非递减函数 f，g：$[0, \infty) \to \mathbb{R}$，使得

$$C(1 + |\varepsilon|^C) \ge g(|\varepsilon|) \ge V(\varepsilon) - V_{\min} \ge f(|\varepsilon|) \ge c|\varepsilon|^{1+\delta} \text{ for all } \varepsilon \in \mathbb{R}.$$

2.2.1 后验边缘分布的局部计算

在部分匹配模型中，对于每个索引 $i \in [N_X]$ 和 $j \in \{\varnothing\} \cup [N_Y]$，我们简写为

$$P_i(j) = P(\pi(i) = j \mid X, Y) \tag{9}$$

表示以下边际后验概率：X_i 与 $\varnothing$ 或 Y_j匹配。因此 $P_i(\varnothing) + \sum_{j=1}^{N_Y} P_i(j) = 1$。给定用户指定的局部性参数 $M > 0$（在 n中为常数），算法 1 描述了对这些边缘分布进行局部近似的过程 P_1, ..., P_{N_X}。[1] 该过程是将哈密顿量（8）限制在每个点 $O(1/n)$ 周围大小为 X_i的窗口内，并通过对该窗口中各点的部分匹配进行后验分布的局部计算，从而近似 P_i。

下面的定理给出了算法 1 的定量 TV 保证。

定理 2.4。假设 2.3成立。令 P_1、.....、P_{N_X} 表示算法 1的输出，该算法针对某些 K、$L > 0$，取 $M = KL$ 运行。则存在一个常数 $C > 0$，其取决于 $\Lambda(\cdot)$ $V(\cdot)$，但不取决于 K、L，使得对于任意 $\kappa > 1$，在 $(X$、$Y)$ 上以至少 $1 - \exp(-(\log n)^\kappa)$ 的概率，对于所有 $n \ge n_0(\kappa$、K、$L)$，

$$\frac{1}{N_X} \sum_{i=1}^{N_X} \mathsf{TV}(P_i, \widehat{P}_i) \le CL^{-\delta/2} + C\left(1 - e^{-CL \log L}\right)^{K/3}. \tag{10}$$

[1]由对称性可知，一个　可以应用相同的算法来近似 Y_1、.....、Y_{N_Y} 的边缘匹配概率　.

特别地，这意味着对于任意 $\iota > 0$，都可以选取 $L \equiv L(\iota)$，然后选取足够大的 $K \equiv K(L, \iota)$，从而确保式（10）中的平均全变差误差不超过 ι，且当 n 足够大时以高概率成立。[2]

2.2.2 后验分布的渐近行为

令 PPP(μ) 表示实数集 $\mathbb{R}$ 上速率为 $\mu > 0$ 的齐次泊松点过程的分布律。给定 (X, Y)，对索引 $I \in [N_X]$进行均匀随机选择时，重缩放点过程$\{n(X_i - X_I)\}_{i\in[N_X]}$和 $\{n(Y_j - X_I)\}_{j\in[N_Y]}$将随着 $n \to \infty$ 几乎必然弱收敛至以下耦合泊松点过程的分布律（参见推论 C.7）。

定义 2.5. 给定 $p \in (0,1)$ 以及实数集 $[0,1]$，上的概率密度 $\Lambda(x)$，令 $(\mathbf{X}_{\Lambda,p}, \mathbf{Y}_{\Lambda,p})$ 为实数集 $\mathbb{R}$ 上的两个随机点过程，并令$\pi^*_{\Lambda,p}$ 为从点集 $\mathrm{dom}(\pi^*_{\Lambda,p}) \subseteq \mathbf{X}_{\Lambda,p}$ 到点集 $\mathrm{range}(\pi^*_{\Lambda,p}) \subseteq \mathbf{Y}_{\Lambda,p}$的双射，构造如下：

1. 抽样 $x \sim \Lambda$。
2. 抽样 $\mathbf{X}_{\Lambda,\,p} = \{0\} \cup \widetilde{\mathbf{X}}_{\Lambda,\,p}$ 其中 $\widetilde{\mathbf{X}}_{\Lambda,\,p} \sim \mathrm{PPP}(\Lambda(x)p/(1-p)^2)$ 。
3. 对每个 $\mathrm{X} \in \mathbf{X}_{\Lambda,\,p}$，以概率 p 纳入 $\mathrm{X} \in \mathrm{dom}(\pi^*_{\Lambda,\,p})$。
4. 对每个 $\mathrm{X} \in \mathrm{dom}(\pi^*_{\Lambda,\,p})$，抽样 $\varepsilon \sim q(\cdot)$，并令 $\mathrm{Y} = \mathrm{X} + \varepsilon$。设 $\pi^*_{\Lambda,\,p}(\mathrm{X}) = \mathrm{Y}$，且 $\mathrm{range}(\pi^*_{\Lambda,\,p}) = \{\pi^*_\Lambda(\mathrm{X}) : \mathrm{X} \in \mathrm{dom}(\pi^*_{\Lambda,\,p})\}$。
5. 抽样 $\mathbf{Y}_{\Lambda,\,p} \sim \mathrm{PPP}(\Lambda(x)p/(1-p))$，并设 $\mathbf{Y}_{\Lambda,\,p} = \mathbf{Y}_{\Lambda,\,p} \cup \mathrm{range}(\pi^*_{\Lambda,\,p})$。

注意，点 $0 \in \mathbf{X}_{\Lambda,\,p}$ 对应于中心点 $X_I \in \{X_1, \ldots, X_{N_X}\}$。

我们将通过以下测试函数类 $f(P, j^*)$，来描述后验边缘分布与真实匹配 $\{(P_i, \pi^*(i))\}_{i\in[N_X]}$ 的经验分布的一种弱极限概念，其中 P 表示给定点的匹配概率向量，j^* 表示其真实匹配。

定义 2.6（匹配代价函数）。设 $\mathbb{N} = \{1, 2, 3, \ldots\}$，且令 $\mathcal{P}(\mathbb{N})$ 表示 $\mathbb{N}$ 上概率分布的空间。置换不变且全变差连续的匹配代价 是一个函数 $f\colon \mathcal{P}(\mathbb{N}) \times \mathbb{N} \to \mathbb{R}$，满足以下条件：

1.（置换不变性）任取 $P \in \mathcal{P}(\mathbb{N})$、$j^* \in \mathbb{N}$ 和双射 $\sigma : \mathbb{N} \to \mathbb{N}$，并令 $P^\sigma \in \mathcal{P}(\mathbb{N})$ 由 $P^\sigma(j) = P(\sigma^{-1}(j))$ 给出。那么 $f(P, j^*) = f(P^\sigma, \sigma(j^*))$。[3]
2.（总变差连续性）存在一个连续函数 $\iota\colon [0,1] \to [0,\infty)$，满足 $\lim_{x\to 0} \iota(x) = 0$，使得对于任意 $P, P' \in \mathcal{P}(\mathbb{N})$ 和 $j^* \in \mathbb{N}$,

$$|f(P, j^*) - f(P', j^*)| \le \iota(\mathrm{TV}(P, P')) = \iota\left(\frac{1}{2}\sum_{j\in\mathbb{N}} |P(j) - P'(j)|\right).$$

[2]对于任意 $D > 0$，以至少 $1 - \exp(-(\log n)^\kappa)$ 的概率，对于 $n \ge n_0(\kappa, K, L, D)$ 也成立：对于满足 $[x_-^{(i)}, x_+^{(i)}]$ 中观测数据值超过 D 的指标 $i \in [N_X]$，其比例至多为一个常数 $o_D(1)$，并且该常数在 $D \to \infty$ 时趋于零。对于此类指标，跳过（可能难以处理的）$\widehat{P}_i$ 计算 $i \in [N_X]$，会在 (10) 中额外引入一个 $o_D(1)$ 误差。因此，对于任意固定的 $\iota > 0$，我们也可以将 $D \equiv D(\iota, K, L)$ 选得足够大，从而得到一个线性时间局部算法，使其达到 $\frac{1}{N_X}\sum_{i=1}^{N_X}$ 全变差$(P_i, \widehat{P}_i) < \iota$。[3]也就是说，$f(P, j^*)$ 仅取决于真实匹配的概率 $P(j^*)$ 以及其余概率值的集合 $\{P(i) : i \ne j^*\}$ 与它们的排列顺序无关。

于是，对于有限或可数集 S 上的任意概率向量 $P \in \mathcal{P}(S)$ 和元素 $Y^* \in S$，我们将应用 f 所得的结果记作 $f(P, Y^*)$；若 $|S| < \infty$，则通过补零将 $P \in \mathcal{P}(S)$ 扩展。此类匹配代价函数的例子包括：

- P 赋予真实匹配 j^* 的概率：$f(P, j^*) = P(j^*)$。
- 贝叶斯可信集 $f(P, j^*)$ 的期望基数 $= \mathbb{E}|C(P)|$：将 P 的概率值按 $P_{s(1)} \geq P_{s(2)} \geq \cdots$ 排序（相同值的顺序均匀随机确定），并固定目标覆盖水平 $1 - \alpha \in (0,1)$，

$$C(P) = \begin{cases} \{s(1), \ldots, s(k)\} \text{ with probability } \xi \\ \{s(1), \ldots, s(k-1)\} \text{ with probability } 1 - \xi \end{cases}$$

  其中 $k \geq 1$ 和 $\xi \in (0, 1]$ 满足 $P_{s(1)} + \ldots + P_{s(k-1)} + \xi \cdot P_{s(k)} = 1 - \alpha$。（这里，$\mathbb{E}$ 是关于 $C(P)$ 随机构造的概率，并且在此例中，$f(P, j^*)$ 不依赖于 j^*。）
- 此集合 $C(P)$ 对真实匹配的期望覆盖率：$f(P, j^*) = \mathbb{P}[j^* \in C(P)]$（其中 $\mathbb{P}$ 再次表示关于 $C(P)$ 随机构造的概率）。

下述定理表明，此类函数在 $\{(P_i, \pi^*(i))\}_{i \in [N_X]}$ 处取值的平均值——例如真实匹配 $P_i(\pi^*(i))$ 的平均后验概率、贝叶斯可信集 $\mathbb{E}|C(P_i)|$ 的平均大小等——收敛于由上述泊松点过程所定义的极限。

定理 2.7. 设假设 2.3 成立。令 $\mathsf{Q} \equiv \mathsf{Q}(\mathbf{X}_{\Lambda,p}, \mathbf{Y}_{\Lambda,p}, \pi^*_{\Lambda,p}) \in \mathcal{P}(\{\varnothing\} \cup \mathbf{Y}_{\Lambda,p})$ 为命题 2.8 下文所给出的概率向量。

令 $f\colon \mathcal{P}(\mathbb{N}) \times \mathbb{N} \to \mathbb{R}$ 为任意有界、置换不变且全变差连续的匹配代价。则几乎必然

$$\lim_{n \to \infty} \frac{1}{N_X} \sum_{i=1}^{N_X} f(P_i, \pi^*(i)) = \mathbb{E}_{\Lambda,p}[f(\mathsf{Q}, \pi^*_{\Lambda,p}(0))]$$

其中 $\mathbb{E}_{\Lambda,p}$ 是对 $(\mathbf{X}_{\Lambda,p}, \mathbf{Y}_{\Lambda,p}, \pi^*_{\Lambda,p})$ 在定义 2.5 中的分布律所取的期望。

定理 2.7 中的概率向量 $\mathsf{Q} \in \mathcal{P}(\{\varnothing\} \cup \mathbf{Y}_{\Lambda,\,p})$ 可理解为在渐近后验分布的局部限制下，$0 \in \mathbf{X}_{\Lambda,\,p}$ 的边缘匹配概率向量的无限体积极限，该渐近后验分布定义如下：对于任意 $K > 0$，记作 $\mathbf{X}_K = \mathbf{X}_{\Lambda,\,p} \cap [-K, K]$ 以及 $\mathbf{Y}_K = \mathbf{Y}_{\Lambda,\,p} \cap [-K, K]$。令 $\mathbb{S}_K$ 为从 $\mathbf{X}_K$ 到 $\mathbf{Y}_K$ 的所有部分双射构成的空间。设 $U(x) = \sqrt{\Lambda(x)}$，并定义 $\pi \in \mathbb{S}_K$ 上的局部吉布斯测度为

$$Q_K(\pi) \propto \exp\left(- \sum_{\mathsf{X} \in \mathbf{X}_K : \pi(\mathsf{X}) \neq \varnothing} V(\mathsf{X} - \pi(\mathsf{Y})) + \sum_{\mathsf{X} \in \mathbf{X}_K : \pi(\mathsf{X}) = \varnothing} U(\mathsf{X}) + \sum_{\mathsf{Y} \in \mathbf{Y}_K : \pi^{-1}(\mathsf{Y}) = \varnothing} U(\mathsf{Y}) \right).$$

定义概率向量 $Q_K^0 = \{Q_K(\pi(0) = \mathsf{Y})\}_{\mathsf{Y} \in \{\varnothing\} \cup \mathbf{Y}_K} \in \mathcal{P}(\{\varnothing\} \cup \mathbf{Y}_K)$，表示点 $0 \in \mathbf{X}_K$ 在 Q_K 下的匹配。

命题 2.8. Q_K^0 在 $\mathbb{P}_{\Lambda,\,p}$-概率意义下收敛到唯一极限 $\mathsf{Q} \in \mathcal{P}(\{\varnothing\} \cup \mathbf{Y}_{\Lambda,\,p})$，且在 全变差 意义下收敛，当 $K \to \infty$ 时；也就是说，对于任意 $\varepsilon > 0$，

$$\lim_{K \to \infty} \mathbb{P}_{\Lambda,p}[\mathsf{TV}(\mathsf{Q}, Q_K^0) \geq \varepsilon] = 0$$

其中 $\mathbb{P}_{\Lambda,\,p}$ 是对 $(\mathbf{X}_{\Lambda,\,p}, \mathbf{Y}_{\Lambda,\,p}, \pi^*_{\Lambda,\,p})$ 的概率。

2.3 精确匹配的结果

2.3.1 局部计算后验边缘分布

在精确匹配模型中，令 s，t：$[n] \to [n]$ 为对 X 和 Y进行排序的置换，即

$$X_{s(1)} < X_{s(2)} < \ldots < X_{s(n)}, \qquad Y_{t(1)} < Y_{t(2)} < \ldots < Y_{t(n)}.$$

在给定点 X、Y的条件下，对于任意匹配 π，令 $\pi_{st} := t^{-1} \circ \pi \circ s$ 表示已排序坐标下对应的匹配，即 $\pi_{st}(i) = j$ 若 $\pi(s(i)) = t(j)$（X匹配中的第 i^{th} 个有序值与 Y中的第 j^{th} 个有序值相匹配）。对于每个索引 $i \in [n]$，我们用以下简写表示

$$P_i(j) = P(\pi(i) = j \mid X, Y), \quad P_i^{\text{sort}}(j) = P_{s(i)}(t(j)) = P(\pi_{st}(i) = j \mid X, Y). \tag{11}$$

因此 $P_i^{\text{sort}}(j)$ 是后验分布（4）下，X 中的第 i^{th} 个有序值与 j^{th}中的第 Y 个有序值相匹配的边缘概率。固定一个（整数）局部性参数 $M \geq 1$，算法 2 描述了我们对这些边缘分布进行局部近似的步骤 P_1、……、P_n，在对数据值进行上述全局排序之后。该算法通过计算 P_i^{sort}在以下对象之间双射的局部后验分布，来近似 O个围绕 $X_{s(i)}$ 的已排序值与对应的O个围绕 $Y_{t(i)}$的已排序值。

算法 2： 精确匹配后验的局部近似

---

for $i \leftarrow 1$ **to** n **do**

Let $i_- = \max(i - M, 1)$ and $i_+ = \min(i + M, n)$.

Let $\mathbb{S}^{(i)}$ be the set of bijections $\pi : \{i_-, \ldots, i_+\} \to \{i_-, \ldots, i_+\}$.

Compute the local posterior for matchings $\pi \in \mathbb{S}^{(i)}$ given by

$$P^{(i)}(\pi \mid X, Y) \propto \exp\left(- \textstyle\sum_{k=i_-}^{i_+} V(n(X_{s(k)} - Y_{t(\pi(k))}))\right).$$

Let $\widehat{P}_i^{\text{sort}}(j) = \begin{cases} P^{(i)}(\pi(i) = j \mid X, Y) & \text{if } j \in \{i_-, \ldots, i_+\}, \\ 0 & \text{otherwise.} \end{cases}$

Let $\widehat{P}_i(j) = \widehat{P}^{\text{sort}}_{s^{-1}(i)}(t^{-1}(j))$.

end

返回 $\widehat{P}_1$, ……, $\widehat{P}_n$.

下述定理为该算法建立了全变差保证。

^ ^定理 *2.9*。假设假设 *2.3*成立。设 P_1，……，P_n 表示在 $M = KL$ 条件下，对某些 K、$L \geq 1$运行算法 *2*所得的输出。那么存在一个常数 $C > 0$，它依赖于$\Lambda(\cdot)$ $V(\cdot)$，但不依赖于 K、L；使得对任意 $\kappa > 1$，以至少 $1 - \exp(-(\log n)^{\kappa})$的概率，在$(X, Y)$上对所有 $n \geq n_0(\kappa, K, L)$，

$$\frac{1}{n} \sum_{i=1}^{n} \mathsf{TV}(P_i, \widehat{P}_i) \leq CL^{-\delta} + C\left(1 - e^{-CLg(4L/\Lambda_{\min})}\right)^{K/3}. \tag{12}$$

同样，这意味着对于任意 $\iota > 0$，都可以选取足够大的 $L \equiv L(\iota)$，再选取足够大的 $K \equiv K(L, \iota)$，以确保当 n 足够大时，该平均全变差误差至多为 ι。

2.3.2 后验分布的渐近行为

与部分匹配情形一样，在给定 $(X,\ Y)$ 的条件下，对索引$I \in [n]$ 作均匀随机选择时，重缩放点过程 $\{n(X_i - X_I)\}_{i\in[n]}$ 和 $\{n(Y_i - X_I)\}_{i\in[n]}$ 几乎必然弱收敛到以下耦合泊松点过程的分布律（参见引理B.14）。

定义 2.10. 给定 $[0,1]$, 上的概率密度 $\Lambda(x)$，令（$\mathbf{X}_\Lambda$，$\mathbf{Y}_\Lambda$）为 $\mathbb{R}$ 与π^*_Λ 上的两个点过程：$\mathbf{X}_\Lambda \mapsto \mathbf{Y}_\Lambda$ 是$\mathbf{X}_\Lambda$ 与$\mathbf{Y}_\Lambda$之间的一个双射，其构造如下：

1. 抽样 $x \sim \Lambda$。

2. 抽样$\tilde{\mathbf{X}}_\Lambda = \{0\} \cup \mathbf{X}_\Lambda$，其中$\mathbf{X}_\Lambda \sim \mathrm{PPP}(\Lambda(x))$。

3. 对每个$\mathrm{X} \in \mathbf{X}_\Lambda$独立地抽样 $\varepsilon \sim q(\cdot)$，并令$\mathrm{Y} = \mathrm{X} + \varepsilon$。令$\pi^*_\Lambda(\mathrm{X}) = \mathrm{Y}$且$\mathbf{Y}_\Lambda = \{\pi^*_\Lambda(\mathrm{X}) \colon \mathrm{X} \in \mathbf{X}_\Lambda\}$。

关于后验边缘分布的经验分布的下述结果 $\{(P_i,\ \pi^*(i))\}_{i\in[n]}$ 类似于定理2.7 在部分匹配设定中。然而，该极限的构造Q 更为复杂，并在下文加以描述。

定理 2.11。 假设 *2.3* 成立。设 $\mathsf{Q} \equiv \mathsf{Q}(\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda, \pi^*_\Lambda) \in \mathcal{P}(\mathbf{Y}_\Lambda)$ 为下文命题 *2.14*给出的概率向量。

设 $f\colon\ \mathcal{P}(\mathbb{N}) \times \mathbb{N} \to \mathbb{R}$ 为任意有界、置换不变且全变差连续的匹配代价。则几乎必然

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(P_i, \pi^*(i)) = \mathbb{E}_\Lambda[f(\mathsf{Q}, \pi^*_\Lambda(0))],$$

其中 $\mathbb{E}_\Lambda$ 是关于定义 $(\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda, \pi^*_\Lambda)$ 中 *2.10*的分布律的期望。

概率向量 $\mathsf{Q} \in \mathcal{P}(\mathbf{Y}_\Lambda)$ 定义如下：限制在 $\mathbf{X}_\Lambda$ 和$\mathbf{Y}_\Lambda$ 均为局部有限的概率为 1 的事件上，我们将$\mathbf{X}_\Lambda$, $\mathbf{Y}_\Lambda$ 中的点标记为

$$\ldots < \mathsf{X}_{-2} < \mathsf{X}_{-1} < \mathsf{X}_0 = 0 < \mathsf{X}_1 < \mathsf{X}_2 < \ldots$$
$$\ldots < \mathsf{Y}_{-2} < \mathsf{Y}_{-1} < \mathsf{Y}_0 = \pi^*_\Lambda(0) < \mathsf{Y}_1 < \mathsf{Y}_2 < \ldots$$

其中，特定点 $0 \in \mathbf{X}_\Lambda$ 和 $\pi^*_\Lambda(0) \in \mathbf{Y}_\Lambda$ 的索引为 0。我们将双射$\pi^*_\Lambda\colon\ \mathbf{X}_\Lambda \to \mathbf{Y}_\Lambda$ 与其索引之间的双射 $\pi^*_\Lambda\colon\ \mathbb{Z} \to \mathbb{Z}$ 等同起来。

命题 2.12. 对任意 $a \in \mathbb{Z}$，定义

$$\mathsf{L}_a(\pi) = \sum_{i\in\mathbb{Z}:i\le a} \sum_{j\in\mathbb{Z}:j>a} \mathbf{1}\{\pi(i) = j\}, \qquad \mathsf{R}_a(\pi) = \sum_{i\in\mathbb{Z}:i>a} \sum_{j\in\mathbb{Z}:j\le a} \mathbf{1}\{\pi(i) = j\}.$$

则以概率 *1*，关于$(\mathbf{X}_\Lambda,\ \mathbf{Y}_\Lambda,\ \pi^*_\Lambda)$，我们有$\mathsf{L}_a(\pi^*_\Lambda)$，实数集$a(\pi^*_\Lambda) < \infty$ 对每一个 $a \in \mathbb{Z}$，且 d

$$\mathsf{F}(\pi^*_\Lambda) := \mathsf{L}_a(\pi^*_\Lambda) - \mathsf{R}_a(\pi^*_\Lambda) \in \mathbb{Z} \tag{13}$$

在所有 $a \in \mathbb{Z}$ 处取相同的值。

定义 2.13. 双射 $\pi\colon\ \mathbb{Z} \to \mathbb{Z}$ 若 $\mathsf{L}_a(\pi)$，$\mathsf{R}_a(\pi) < \infty$ 且 $\mathsf{L}_a(\pi) - \mathsf{R}_a(\pi) = k + \mathsf{F}(\pi^*_\Lambda)$ 对每个 $a \in \mathbb{Z}$ 都成立，则称其相对于 π^*_Λ 具有流 k 。

对于任意整数 $K \geq 1$，令 $\mathbb{S}_K$为双射空间 $\pi\colon \{-K,\dots,K\} \to \{-K+\mathsf{F}(\pi^*_\Lambda)\dots, K+\mathsf{F}(\pi^*_\Lambda)\}$，并定义 $\mathbb{S}_K$上的局部吉布斯测度

$$Q_K(\pi) \propto \exp\left(-\sum_{i\in\{-K,\dots,K\}} V(\mathsf{X}_i - \mathsf{Y}_{\pi(i)})\right). \tag{14}$$

Q_K 当 $K\to\infty$ 时的极限可理解为与形式哈密顿量 $H(\pi \mid \mathbf{X}_\Lambda, \mathbf{Y}_\Lambda) = \sum_{i\in\mathbb{Z}} V(\mathsf{X}_i - \mathsf{Y}^{\pi(i)})$ 对应的无限体积极后验分布，其中限制在相对于 π^*_Λ流为 0 的双射 $\pi\colon \mathbb{Z}\to\mathbb{Z}$ 上。令 $Q^0_K \in \mathcal{P}(\mathbf{Y}_\Lambda) \equiv \mathcal{P}(\mathbb{Z})$ 为概率向量，其定义为：对每个 $j\in\mathbb{Z}$，有 $Q^0_K(j) = Q_K(\pi(0) = j) \equiv Q_K(\pi(\mathsf{X}_0) = \mathsf{Y}j)$。

命题 2.14。Q^0_K 依概率收敛到唯一极限 $\mathsf{Q} \in \mathcal{P}(\mathbf{Y}_\Lambda)$ 当 $K\to\infty$时，即对任意 $\varepsilon > 0$,

$$\lim_{K\to\infty} \mathbb{P}_\Lambda[\mathsf{TV}(\mathsf{Q}, Q^0_K) \geq \varepsilon] = 0$$

其中 $\mathbb{P}_\Lambda$ 是关于($\mathbf{X}_\Lambda$、$\mathbf{Y}_\Lambda$、π^*_Λ)的概率。

定理2.9和2.11的证明概述 3

我们阐述证明精确匹配定理 2.9和 2.11 的主要思路，这些思路借鉴了 [8]的见解：对每个索引 $l\in[n]$，定义边界变量 $\Gamma_{l+\frac12}$(π在 π处的 $l+\frac12$如下：

$$\Gamma_{l+\frac12}(\pi) := \{(k,\pi_{st}(k))\colon k\leq l, \pi_{st}(k) > l\} \cup \{(k,\pi_{st}(k))\colon k > l, \pi_{st}(k)\leq l\}.$$

因此 $\Gamma_{l+\frac12}(\pi)$ 是排序后数据值对应的指标对（a，b），其中 $a\leq l$且$b > l$，反之亦然，并且 $X_{s(a)}$ 在 $Y_{t(b)}$ 下与 π匹配。固定（较大的）局部性参数 K，$L > 0$，算法 2 为 $M = KL$ 计算出的匹配概率，恰好等于条件后验分布 $P(\pi \mid X, Y, \Gamma_{i\pm(KL+\frac12)}(\pi) = \varnothing)$ 下的匹配概率。

吉布斯测度的局部性。回顾下界 $\Lambda(x) \geq \Lambda_{\min}$，对每个 $l\in[n]$，我们定义依赖于（X，Y）的事件

$$\mathcal{A}_l = \left\{ \sup_{j:|j-(l+\frac12)|\leq L} \{n|X_{s(j)} - X_{s(l)}|, n|Y_{t(j)} - Y_{t(l)}|\} \leq \frac{3L}{2\Lambda_{\min}},\ n\left|X_{s(l)} - Y_{t(l)}\right| \leq L \right\}$$

以及依赖于 π 的事件

$$\mathcal{C}_l(\pi) = \left\{|k-m|\leq L \text{ for all } (k,m)\in\Gamma_{l+\frac12}\right\}, \qquad \mathcal{L}_l(\pi) = \mathcal{A}_l \cap \mathcal{C}_l(\pi),$$

$$\mathcal{G}_i(\pi) = \left\{\sum_{k=1}^{KL} \mathbf{1}\{\mathcal{L}_{i-k}(\pi)\} \geq \frac{2KL}{3}\right\} \cap \left\{\sum_{k=0}^{KL-1} \mathbf{1}\{\mathcal{L}_{i+k}(\pi)\} \geq \frac{2KL}{3}\right\}$$

即，$\mathcal{G}_i(\pi)$ 表示这样的事件：存在足够多的位置 $l\in[i-KL, i+KL]$，使得 $X_{s(l)}$ 附近的点和 $Y_{t(l)}$ 附近的点足够规则，并且 π 在边界 $\Gamma_{l+\frac12}(\pi)$ 处不存在（在排序索引中大小为 $> L$ 的）长程跳跃。

一个关键引理（引理 B.3）指出，当 L 足够大时，在 $P(\pi \mid X,Y)$ 和局部后验分布 $P(\pi \mid X,Y,\Gamma_{i\pm(KL+\frac12)}=\varnothing)$ 下，对于 (X,Y)，高概率下大多数位置 $i\in[n]$ 都会满足 $\mathcal{G}_i(\pi)$。具体来说，我们将证明

$$\frac1n\sum_{i=1}^n P(\mathcal{G}_i^c(\pi) \mid X,Y) \leq CL^{-\delta}, \qquad \frac1n\sum_{i=1}^n P(\mathcal{G}_i^c(\pi) \mid X,Y,\Gamma_{i\pm(KL+\frac12)} = \varnothing) \leq CL^{-\delta} \tag{15}$$

其中 $\delta > 0$ 是 $V(\varepsilon)$ 下包络的指数，见假设 2.3。

相关性衰减。固定 $i \in [n]$，令 $\mathcal{Q}_i(\pi)$ 为满足 $\mathcal{L}_l(\pi)$ 的位置 $l \in [i-KL, i+KL]$ 列表。对于每个位置 $m \in [i-KL, i+KL]$，令 $\mathcal{F}_m(\pi)$ 为信息集，由 $\mathcal{Q}_i(\pi)$ 以及边界变量 $\{\Gamma_{l+\frac{1}{2}}(\pi)$ 的取值 $l \in \mathcal{Q}_i(\pi) \setminus [m-L, m+L]\}$ 构成。我们将为 $\Gamma_{m+\frac{1}{2}}(\pi)$ 为空的概率建立一个定量下界：

$$P(\Gamma_{m+\frac{1}{2}}(\pi) = \varnothing \mid X, Y, \mathcal{L}_m(\pi), \mathcal{F}_m(\pi)) \geq \iota(L) := e^{-CLg(4L/\Lambda_{\min})}$$

其中 $g(\cdot)$ 是假设 2.3中的单调上包络。我们将通过构造一个映射来证明这一点：将满足 $\Gamma_{m+\frac{1}{2}}(\pi) \neq \varnothing$的双射 π映射到满足 $\Gamma_{m+\frac{1}{2}}(\pi) = \varnothing$的双射；这一构造的关键在于$\Gamma_{m+\frac{1}{2}}(\pi)$ 是通过排序索引定义的。（部分匹配模型中的类似论证更简单：我们只需将边界变量 $\Gamma_x(\pi)$ 定义为索引对 $(k, \pi(k))$，其中 $X_k \leq x$ 且 $Y_{\pi(k)} > x$，或反之，无需排序。）

然后，应用 [8, 引理 2.14] 中独立耦合两个后验样本的思路$\pi, \pi' \sim P(\cdot \mid X, Y)$，可以证明，在事件 $\mathcal{G}_i(\pi) \cap \mathcal{G}_i(\pi')$ 发生时，在该耦合下，以至少$1-(1-\iota(L)^2)^{K/3}$ 的概率，存在一个索引 $m \in [i-KL, i)$，以及另一个索引$m \in [i, i+KL)$，使得二者均满足 $\Gamma_{m+\frac{1}{2}}(\pi) = \Gamma_{m+\frac{1}{2}}(\pi') = \varnothing$。这足以推出相关性衰减

$$\begin{aligned}
&\mathsf{TV}\left((P(\pi_{st}(i)=j \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}=\varnothing))_{j=1}^n, (P(\pi_{st}(i)=j \mid X, Y)_{j=1}^n\right) \\
&\leq C\Big((1-\iota(L)^2)^{K/3} + P(\mathcal{G}_i^c(\pi) \mid X, Y) + P(\mathcal{G}_i^c(\pi) \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}=\varnothing)\Big)
\end{aligned}$$

结合（15），可得定理 2.9。

吉布斯测度的渐近性。为证明定理 2.11，我们应用定理 2.9来近似$\widehat{}$每个后验边缘向量 P_i，使用其估计值 $\widehat{P}_i$，该估计值来自算法 2，

$$\frac{1}{n}\sum_{i=1}^n f(P_i, \pi^*(i)) \approx \frac{1}{n}\sum_{i=1}^n f(\widehat{P}_i, \pi^*(i)).$$

$\widehat{}$这一估计量 $\widehat{P}_i$ 并非仅依赖于周围点的局部函数 X_i，因为其计算需要知道哪个点 Y_j 与 X_i具有相同的排序秩。我们将证明，当 $D > 0$较大时，以高概率 $\widehat{P}_i = \widetilde{P}_i^D$，对于一个替代量 $\widetilde{P}_i^D$，其仅依赖于 π^* 以及点过程 $\{X_1, \ldots, X_n\}$ 和 $\{Y_1, \ldots, Y_n\}$ 在 $[X_i - \frac{D}{n}, X_i + \frac{D}{n}]$上的限制，该结论成立。这个替代量由算法2的一个类似版本给出，该算法计算真实双射的流的局部近似π^* 在位置 $i \in [n]$ 处的值（参见算法3）。然后，进一步作出近似

$$\frac{1}{n}\sum_{i=1}^n f(\widehat{P}_i, \pi^*(i)) \approx \frac{1}{n}\sum_{i=1}^n f(\widetilde{P}_i^D, \pi^*(i)),$$

我们可以近似 应用弱收敛到泊松点过程极限来证明定理 2.11 .

命题证明2.1 和 2.2

命题的证明 *2.1*。我们首先证明（c）。利用变量代换 $\varepsilon = n^{1/d}(y-x)$ 代入 $\mathbb{R}^d$ 和$V(\varepsilon) = V(-\varepsilon)$，由此可得

$$\begin{aligned} n\mathcal{Z}_n &= n\int_{[0,1]^d}\int_{[0,1]^d}\sqrt{\Lambda(x)\Lambda(y)}\exp\left(-V(n^{1/d}(x-y))\right)\mathrm{d}x\,\mathrm{d}y \\ &= \int_{[0,1]^d}\mathrm{d}x\int_{\mathbb{R}^d}\mathrm{d}\varepsilon\sqrt{\Lambda(x)\Lambda\left(x+n^{-1/d}\varepsilon\right)}\exp\left(-V(\varepsilon)\right)\mathbf{1}\left\{x+n^{-1/d}\varepsilon\in[0,1]^d\right\}. \end{aligned}$$

当 $n\to\infty$ 时，被积函数逐点收敛到 $\Lambda(x)e^{-V(\varepsilon)}$，且 $\int_{[0,1]^d}\Lambda(x)\mathrm{d}x\int_{\mathbb{R}^d}e^{-V(\varepsilon)}\mathrm{d}\varepsilon = 1$。因此，由控制收敛定理，$\lim_{n\to\infty} n\cdot\mathcal{Z}_n = 1$ 成立。

命题 (a) 类似可得，只需应用 (c)：对于任意固定的 $x\in(0,1)^d$，

$$\begin{aligned} p_n(x) &= \frac{1}{\mathcal{Z}_n}\int_{[0,1]^d}\sqrt{\Lambda(x)\Lambda(y)}\exp\left(-V(n^{1/d}(x-y))\right)\mathrm{d}y \\ &= \frac{1}{n\mathcal{Z}_n}\int_{\mathbb{R}^d}\sqrt{\Lambda(x)\Lambda(x+n^{-1/d}\varepsilon)}\exp\left(-V(\varepsilon)\right)\mathbf{1}\left\{x+n^{-1/d}\varepsilon\in[0,1]^d\right\}\mathrm{d}\varepsilon \\ &\to \Lambda(x)\int_{\mathbb{R}^d}e^{-V(\varepsilon)} = \Lambda(x). \end{aligned}$$

对于 (b)，根据定义，

$$p_n(\varepsilon\mid x) = \frac{(n\mathcal{Z}_n)^{-1}\sqrt{\Lambda(x)\Lambda(x+n^{-1/d}\varepsilon)}e^{-V(\varepsilon)}\mathbf{1}\{x+n^{-1/d}\varepsilon\in[0,1]^d\}}{p_n(x)}.$$

应用 (a) 和 (c)，当 $n\to\infty$ 时，对于任意固定的 $x\in(0,1)^d$ 以及有界区域 $S\subset\mathbb{R}^d$，该式关于 $\varepsilon\in S$ 一致收敛到 $q(\varepsilon) = e^{-V(\varepsilon)}$。 □

命题的证明 *2.2*。我们首先计算给定观测值（6）后所有潜变量的完整后验分布。我们记 $\propto$以吸收仅依赖于观测数据的任意数值常数（N_X, N_Y, X, Y）。于是

$$\begin{aligned} &P\left(\pi_X,\pi_Y,\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,N\mid N_X,N_Y,X,Y\right) \\ &\propto P\left(\pi_X,\pi_Y,\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,N,N_X,N_Y,X,Y\right) \\ &= \int P\left(\pi_X,\pi_Y,\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,N,N_X,N_Y,X,Y,\bar{X},\bar{Y}\right)\mathrm{d}\bar{X}\,\mathrm{d}\bar{Y} \\ &= \int P\left(X,Y\mid\pi_X,\pi_Y,\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,N,N_X,N_Y,\bar{X},\bar{Y}\right) \\ &\qquad\times P\left(\pi_X,\pi_Y\mid\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,N,N_X,N_Y,\bar{X},\bar{Y}\right) \\ &\qquad\times P\left(N_X,N_Y\mid N,\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing,\bar{X},\bar{Y}\right) \\ &\qquad\times P\left(\mathcal{S}_{XY},\mathcal{S}_X,\mathcal{S}_Y,\mathcal{S}_\varnothing\mid N,\bar{X},\bar{Y}\right)P\left(\bar{X},\bar{Y}\mid N\right)P(N)\,\mathrm{d}\bar{X}\,d\bar{Y} \end{aligned}$$

由此可得

$$
\begin{aligned}
&P\left(\pi_X, \pi_Y, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing, N \mid N_X, N_Y, X, Y\right) \\
&\propto \int \Bigg\{ \prod_{i=1}^{N_X} \mathbf{1}\left\{X_i = \bar{X}_{\pi_X(i)}\right\} \prod_{i=1}^{N_Y} \mathbf{1}\left\{Y_i = \bar{Y}_{\pi_Y(i)}\right\} \times \frac{1}{N_X! N_Y!} \\
&\quad \times \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y\right\} \\
&\quad \times \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} p^{2\left|\mathcal{S}_{XY}\right|} \cdot (p(1-p))^{\left|S_X\right|} \cdot (p(1-p))^{\left|S_Y\right|} \cdot (1-p)^{2\left|S_\varnothing\right|} \\
&\quad \times \prod_{i=1}^{N} p_n\left(\bar{X}_i, \bar{Y}_i\right) \times P(N) \Bigg\} \mathrm{d}\bar{X}\, \mathrm{d}\bar{Y} \\
&\propto \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} \\
&\quad \times p^{2\left|\mathcal{S}_{XY}\right| + \left|\mathcal{S}_X\right| + \left|\mathcal{S}_Y\right|} (1-p)^{2\left|\mathcal{S}_\varnothing\right| + \left|\mathcal{S}_X\right| + \left|\mathcal{S}_Y\right|} P(N) \\
&\quad \times \prod_{k \in \mathcal{S}_{XY}} p_n\left(X_{\pi_X^{-1}(k)}, Y_{\pi_Y^{-1}(k)}\right) \prod_{k \in \mathcal{S}_X} p_n\left(X_{\pi_X^{-1}(k)}\right) \prod_{k \in \mathcal{S}_Y} p_n\left(Y_{\pi_Y^{-1}(k)}\right) \\
&\propto \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} \\
&\quad \times p^{N_X + N_Y} (1-p)^{2N - (N_X + N_Y)} P(N) \\
&\quad \times \prod_{k \in \mathcal{S}_{XY}} \frac{p_n\left(X_{\pi_X^{-1}(k)}, Y_{\pi_Y^{-1}(k)}\right)}{p_n\left(X_{\pi_X^{-1}(k)}\right) \cdot p_n\left(Y_{\pi_Y^{-1}(k)}\right)} \prod_{k \in \mathcal{S}_{XY} \cup \mathcal{S}_X} p_n\left(X_{\pi_X^{-1}(k)}\right) \prod_{k \in \mathcal{S}_{XY} \cup \mathcal{S}_Y} p_n\left(Y_{\pi_Y^{-1}(k)}\right).
\end{aligned}
$$

注意 $N_X + N_Y$ 在此计算中为常数，且最后两个因子可写成 $\prod_{i=1}^{N_X} p_n(X_i)$与 $\prod_{i=1}^{N_Y} p_n(Y_i)$，因而在此计算中也是常数（仅依赖于$N_X$， N_Y， X， Y）。将联合密度（2）代入 $p_n(x,\ y)$，于是我们有

$$
\begin{aligned}
&P\left(\pi_X, \pi_Y, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing, N \mid N_X, N_Y, X, Y\right) \\
&\propto \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} (1-p)^{2N} P(N) \\
&\times \mathcal{Z}_n^{-\left|\mathcal{S}_{XY}\right|} \exp\left(\sum_{k \in \mathcal{S}_{XY}} -V\left(n^{1/d}\left(X_{\pi_X^{-1}(k)} - Y_{\pi_Y^{-1}(k)}\right)\right) + \log \frac{\sqrt{\Lambda\left(X_{\pi_X^{-1}(k)}\right)}}{p_n\left(X_{\pi_X^{-1}(k)}\right)} + \log \frac{\sqrt{\Lambda\left(Y_{\pi_Y^{-1}(k)}\right)}}{p_n\left(Y_{\pi_Y^{-1}(k)}\right)} \right) \\
&\propto \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} (1-p)^{2N} \mathcal{Z}_n^{-\left|\mathcal{S}_{XY}\right|} P(N) \\
&\times \exp\left(\sum_{k \in \mathcal{S}_{XY}} -V\left(n^{1/d}\left(X_{\pi_X^{-1}(k)} - Y_{\pi_Y^{-1}(k)}\right)\right) - \sum_{k \in \mathcal{S}_X} \log \frac{\sqrt{\Lambda\left(X_{\pi_X^{-1}(k)}\right)}}{p_n\left(X_{\pi_X^{-1}(k)}\right)} - \sum_{k \in \mathcal{S}_Y} \log \frac{\sqrt{\Lambda\left(Y_{\pi_Y^{-1}(k)}\right)}}{p_n\left(Y_{\pi_Y^{-1}(k)}\right)} \right),
\end{aligned}
$$

其中最后一行利用了如下事实：在此计算中，$\sum_{i=1}^{N_X} \log\left(\sqrt{\Lambda(X_i)}/p_n(X_i)\right)$ 与 $\sum_{i=1}^{N_Y} \log\left(\sqrt{\Lambda(Y_i)}/p_n(Y_i)\right)$也是常数。记 $\pi = \pi_Y^{-1} \circ \pi_X$，按约定它携带关于$\mathsf{dom}(\pi) = \pi_X^{-1}(\mathcal{S}_{XY})$与$\mathsf{range}(\pi) = \pi_Y^{-1}(\mathcal{S}_{XY})$的信息，并回顾(8)中哈密顿量 $H_n(\pi \mid X, Y)$的定义，这等价于

$$
\begin{aligned}
&P\left(\pi_X, \pi_Y, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing, N \mid N_X, N_Y, X, Y\right) \\
&\propto \mathbf{1}\left\{\left|\mathcal{S}_{XY} \cup \mathcal{S}_X\right| = N_X,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_Y\right| = N_Y,\, \left|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing\right| = N\right\} \\
&\quad \times (1-p)^{2N} \mathcal{Z}_n^{-\left|\mathcal{S}_{XY}\right|} P(N) \exp\left(-H_n(\pi \mid X, Y)\right).
\end{aligned}
$$

边际后验概率 $P(\pi \mid X, Y) \equiv P(\pi \mid N_X, N_Y, X, Y)$由对所有满足 $|\mathcal{S}_{XY} \cup \mathcal{S}_X| = N_X$、$|\mathcal{S}_{XY} \cup \mathcal{S}_Y| = N_Y$、$|\mathcal{S}_{XY} \cup \mathcal{S}_X \cup \mathcal{S}_Y \cup \mathcal{S}_\varnothing| = N$和$\pi_Y^{-1} \circ \pi_X = \pi$的$(\pi_X, \pi_Y, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing, N)$进一步求和给出。注意，固定 π 也会固定 $|\mathsf{dom}(\pi)| = |\mathsf{range}(\pi)| = |\mathcal{S}_{XY}|$。因此，对于固定的$(N, N_X, N_Y, \pi)$，所有基数 $|\mathcal{S}_{XY}|$、$|\mathcal{S}_X|$、$|\mathcal{S}_Y|$、$|\mathcal{S}_\varnothing|$均已确定。固定$(N, N_X, N_Y, \pi)$的元组$(\pi_X, \pi_Y, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing)$的个数为

$$\binom{N}{|\mathcal{S}_{XY}|, |\mathcal{S}_X|, |\mathcal{S}_Y|, N - |\mathcal{S}_{XY}| - |\mathcal{S}_X| - |\mathcal{S}_Y|} |\mathcal{S}_{XY}|! \, |\mathcal{S}_X|! \, |\mathcal{S}_Y|! = \frac{N!}{(N - N_X - N_Y + |\mathcal{S}_{XY}|)!}.$$

（多项式系数是选择 $\mathcal{S}_{XY}$，$\mathcal{S}_X$，$\mathcal{S}_Y$，$\mathcal{S}_\varnothing$的方式数。于是，有$|\mathcal{S}_X|!$种方式选择将 π_X限制为双射 $[N_X] \setminus \mathsf{dom}(\pi) \to \mathcal{S}_X$，有 $|\mathcal{S}_Y|!$种方式选择将 π_Y限制为双射 $[N_Y] \setminus \mathsf{range}(\pi) \to \mathcal{S}_Y$，有 $|\mathcal{S}_{XY}|!$种方式选择将 π_X 限制为双射$\mathsf{dom}(\pi) \to \mathcal{S}_{XY}$，并且这通过$\pi = \pi_Y^{-1} \circ \pi_X$ 确定了将π_Y限制为双射 $\mathsf{range}(\pi) \to \mathcal{S}_{XY}$的最后一项限制。）然后，对于固定的（$N_X$，$N_Y$，$\pi$），再对取可能值 $N \geq N_X + N_Y - |\mathcal{S}_{XY}| = |\mathcal{S}_X| + |\mathcal{S}_Y| + |\mathcal{S}_{XY}|$的 N作边缘化，这表明

$$\begin{aligned} &P(\pi \mid N_X, N_Y, X, Y) \\ &\propto \sum_{N \geq N_X + N_Y - |\mathcal{S}_{XY}|} (1-p)^{2N} \mathcal{Z}_n^{-|\mathcal{S}_{XY}|} \cdot \frac{N!}{(N - N_X - N_Y + |\mathcal{S}_{XY}|)!} \cdot P(N) \exp\big(-H_n(\pi \mid X, Y)\big). \end{aligned}$$

最后，在选取泊松先验（7）用于 N时，该和

$$\begin{aligned} &\sum_{N \geq N_X + N_Y - |\mathcal{S}_{XY}|} (1-p)^{2N} \mathcal{Z}_n^{-|\mathcal{S}_{XY}|} \cdot \frac{N!}{(N - N_X - N_Y + |\mathcal{S}_{XY}|)!} \cdot P(N) \\ &= \mathcal{Z}_n^{-|\mathcal{S}_{XY}|} \exp\left(-\frac{1}{(1-p)^2 \mathcal{Z}_n}\right) \sum_{N \geq N_X + N_Y - |\mathcal{S}_{XY}|} \frac{\mathcal{Z}_n^{-N}}{(N - N_X - N_Y + |\mathcal{S}_{XY}|)!} \\ &= \mathcal{Z}_n^{-N_X - N_Y} \exp\left(-\frac{1}{(1-p)^2 \mathcal{Z}_n}\right) \sum_{N' \geq 0} \frac{\mathcal{Z}_n^{-N'}}{N'!} \end{aligned}$$

对所有 π都相同，不依赖于 $|\mathcal{S}_{XY}| = |\mathsf{dom}(\pi)| = |\mathsf{range}(\pi)|$。因此 $P(\pi \mid X, Y) \equiv P(\pi \mid N_X, N_Y, X, Y) \propto \exp(-H_n(\pi \mid X, Y))$。 □

B 精确匹配的证明

在这些附录的其余部分，我们将范围限定为 $d = 1$。在全文中，C、C'、c、$c' > 0$ 等表示其取值可能随行而变的常数。对于数据 (X, Y)，我们将用 $\mathbb{P}[\cdot]$、$\mathbb{E}[\cdot]$ 表示概率和期望；在给定 X、Y 的条件下，则用 $P(\cdot \mid X, Y)$、$E(\cdot \mid X, Y)$表示后验分布 $P(\pi \mid X, Y)$ 下的概率和期望。

对于一个 (X, Y)-依赖事件 $\mathcal{E}$，我们称其

$$\mathcal{E} \text{ holds w.h.p.}$$

如果对于任意常数 $\kappa > 1$,

$$\mathbb{P}[\mathcal{E}] \geq 1 - \exp(-(\log n)^\kappa) \text{ for all } n \geq n_0(\kappa).$$

这里，$n_0(\kappa) > 0$ 是一个可能依赖于 κ、函数 $\Lambda(\cdot)$、$V(\cdot)$ 及其在假设2.3中的相关界限，以及可能定义事件 $\mathcal{E}$的局部性常数 K、L、D等的常数。

在本节的引理和分析中，我们始终假设条件 2.3 成立，并采用精确匹配模型（3）的设定。

B.1 定理2.9的证明

B.1.1 平均势

我们首先为（ P（$\pi \mid X$， Y）的后验分布下势函数的均值建立一个界（4）。

引理 B.1。 存在一个常数 $C > 0$ 使得

$$E\left[\frac{1}{n}\sum_{i=1}^{n} V(n(X_i - Y_{\pi(i)})) \,\middle|\, X, Y\right] \le C \ w.h.p.$$

证明。考虑以下模型（维数为 $d = 1$），其中 βV替代 V（2） ：

$$p_{n,\beta}(x,y) = \frac{1}{\mathcal{Z}_n(\beta)}\sqrt{\Lambda(x)\Lambda(y)}\exp\left(-\beta V(n(x-y))\right) \tag{16}$$
$$\mathcal{Z}_n(\beta) = \int_0^1\int_0^1 \sqrt{\Lambda(x)\Lambda(y)}\exp\left(-\beta V(n(x-y))\right)\,\mathrm{d}x\,\mathrm{d}y$$

在此设定下，后验分布（4）变为

$$P_\beta(\pi \mid X, Y) = \frac{1}{\mathcal{Z}_{\text{exact}}(\beta)}\exp\left(-\beta\sum_{i=1}^{n} V(n(X_i - Y_{\pi(i)}))\right) \tag{17}$$

其中 $\mathcal{Z}_{\text{exact}}(\beta) = \sum_{\pi\in\mathbb{S}_n}\exp(-\beta\sum_{i=1}^{n} V(n(X_i - Y_{\pi(i)})))$。我们用 $E_\beta[\cdot \mid X, Y]$表示在 P_β 下的期望。定义相应的自由能和平均势

$$\mathcal{F}_n(\beta) := \frac{1}{n}\log\mathcal{Z}_{\text{exact}}(\beta), \qquad \mathcal{D}_n(\beta) := E_\beta\left[\frac{1}{n}\sum_{i=1}^{n} V(n(X_i - Y_{\pi(i)})) \,\middle|\, X, Y\right]. \tag{18}$$

随后可按标准方法验证

$$\mathcal{D}_n(\beta) = -\mathcal{F}_n'(\beta), \qquad \mathcal{F}_n''(\beta) = \frac{1}{n}\mathsf{Var}_\beta\left[\sum_{i=1}^{n} V(n(X_i - Y_{\pi(i)})) \,\middle|\, X, Y\right] \ge 0.$$

特别地， $\mathcal{F}_n$ 是凸的，因此我们有

$$\mathcal{D}_n(1) \le 2\left(\mathcal{F}_n(1/2) - \mathcal{F}_n(1)\right).$$

通过将对 $\pi \in \mathbb{S}_n$ 求和下界估计为单个求和项 $\pi = \pi^*$，可得到 $\mathcal{F}_n(1)$ 的一个下界。这给出

$$\mathcal{F}_n(1) \ge \frac{1}{n}\log\exp\left(-\sum_{i=1}^{n} V\left(n\left(\bar{X}_i - \bar{Y}_i\right)\right)\right) = -\frac{1}{n}\sum_{i=1}^{n} V\left(n\left(\bar{X}_i - \bar{Y}_i\right)\right)$$

对于任意 $\mu > 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} V\left(n\left(\bar{X}_i - \bar{Y}_i\right)\right) \ge t\right) &\le e^{-\mu t n}\mathbb{E}\exp\left(\mu\sum_{i=1}^{n} V\left(n\left(\bar{X}_i - \bar{Y}_i\right)\right)\right) \\ &= e^{-\mu t n}\left[\mathbb{E}\exp\left(\mu V\left(\varepsilon\right)\right)\right]^n = e^{-\mu t n}\left[\int_{-\infty}^{\infty} e^{(\mu-1)V(\varepsilon)}\mathrm{d}\varepsilon\right]^n\end{aligned}$$

选择足够大的 $\mu = \frac{1}{2}$ 和 $t > 0$（取决于 V）可得

$$\mathcal{F}_n(1) \geq -C \text{ w.h.p.} \tag{19}$$

如下得到 $\mathcal{F}_n(1/2)$ 的一个上界：令 A 为一个 $n \times n$ 矩阵，其 (i,j) 元素为

$$A_{ij} := \exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right).$$

注意，$\mathcal{F}_n(1/2) = n^{-1}\log \operatorname{perm}(A)$。由于 A 的元素非负，其积和式可以用其行和的乘积作上界：

$$\mathcal{F}_n(1/2) \leq \frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{n}\exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right). \tag{20}$$

任取一个 $\kappa > 1$。对于每个 $i \in [n]$，令 $\mathcal{N}_\kappa(\bar{Y}_i) \subset [n]$ 表示 Y_i 在 $\{Y_1, \ldots, Y_n\}$中 $\lfloor (\log n)^\kappa \rfloor$ 个最近邻居的指标集合。注意，

$$\begin{aligned}\sum_{j\notin\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right) &\leq \sum_{j\notin\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{n}{2}\left|\bar{X}_i - \bar{Y}_j\right| + C'\right)\\ &\leq \exp\left(C' + \frac{n}{2}\left|\bar{X}_i - \bar{Y}_i\right|\right)\sum_{j\notin\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{n}{2}\left|\bar{Y}_i - \bar{Y}_j\right|\right)\end{aligned} \tag{21}$$

其中，第一个不等式利用了对某个常数 $C' > 0$，$V(\varepsilon) \geq c|\varepsilon|^{1+\delta} + V_{\min} \geq |\varepsilon| - C'$；第二个不等式利用了 $\left|\bar{X}_i - \bar{Y}_j\right| \geq \left|\bar{Y}_i - \bar{Y}_j\right| - \left|\bar{X}_i - \bar{Y}_i\right|$。令

$$B_i = \sum_{j:j\neq i}\mathbf{1}\left\{n\left|\bar{Y}_i - \bar{Y}_j\right| \leq 2\log n\right\}.$$

在 Y_i 条件下，B_i 服从二项分布 $n-1, p_{n,i}$，其中 $p_{n,i} \leq (C\log n)/n$ 是一个常数 $C > 0$。然后根据切尔诺夫界，对于任意 $t > 0$

$$\begin{aligned}\mathbb{P}\left(B_i \geq (\log n)^\kappa \mid \bar{Y}_i\right) &\leq \exp\left(-t(\log n)^\kappa\right)\mathbb{E}[\exp\left(tB_i\right) \mid \bar{Y}_i]\\ &\leq \exp\left(-t(\log n)^\kappa + np_n(e^t - 1)\right)\\ &\leq \exp\left(-t(\log n)^\kappa + C(e^t - 1)\log n\right).\end{aligned} \tag{22}$$

取 $t = \log[(\log n)^{\kappa-1}] = (\kappa - 1)\log\log n$，得到

$$\mathbb{P}\left(B_i \geq (\log n)^\kappa \mid \bar{Y}_i\right) \leq \exp(-c(\log n)^\kappa \log\log n). \tag{23}$$

若 $B_i < (\log n)^\kappa$，那么 $\bar{Y}_i$ 到每个点 $\{\bar{Y}_j : j \notin \mathcal{N}_\kappa(\bar{Y}_i)\}$ 的距离都大于 $(2\log n)/n$。因此，根据 (21)，其概率至少为 $1 - \exp\left[-c(\log n)^\kappa \log\log n\right]$ 除以 $(Y_j)_{j\neq i}$，

$$\begin{aligned}\sum_{j\notin\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right) &\leq \exp\left(C' + \frac{n}{2}\left|\bar{X}_i - \bar{Y}_i\right|\right)n\exp\left(-\frac{n}{2}\cdot\frac{2\log n}{n}\right)\\ &\leq \exp\left(C' + \frac{n}{2}\left|\bar{X}_i - \bar{Y}_i\right|\right)\end{aligned}$$

利用并集界，概率为 $1-n\exp\left[-c(\log n)^{\kappa}\log\log n\right]$,

$$\mathcal{F}_n(1/2)\le\frac{1}{n}\sum_{i=1}^{n}\log\left[\exp\left(C'+\frac{n}{2}\left|\bar{X}_i-\bar{Y}_i\right|\right)+\sum_{j\in\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right)\right] \tag{24}$$

No注意到给定条件 $q(\varepsilon)=e^{-V(\varepsilon)}\le e^{-c|\varepsilon|^{1+\delta}-V_{\min}}$ 蕴含尾界

$$\mathbb{P}[|\varepsilon|\ge t]\le Ce^{-ct^{1+\delta}}\ \text{for all}\ t\ge 0. \tag{25}$$

Then

$$\mathbb{P}\Big(\max_{i\in[n]}|\varepsilon_i|\ge(\log n)^{\kappa}\Big)\le\sum_{i=1}^{n}\mathbb{P}(|\varepsilon_i|\ge(\log n)^{\kappa})\le Cn\exp\left(-c(\log n)^{\kappa(1+\delta)}\right).$$

由此可知，以概率 $1-n\exp\left[-c(\log n)^{\kappa}\log\log n\right]-Cn\exp\left(-c(\log n)^{\kappa(1+\delta)}\right)$,

$$\mathcal{F}_n(1/2)\le\underbrace{\frac{1}{n}\sum_{i=1}^{n}\log\left[\exp\left(C'+\frac{1}{2}\left(n\left|\bar{X}_i-\bar{Y}_i\right|\wedge(\log n)^{\kappa}\right)\right)+\sum_{j\in\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right)\right]}_{:=\mathscr{F}\left((\bar{X}_1,\bar{Y}_1),\ldots,(\bar{X}_n,\bar{Y}_n)\right)} \tag{26}$$

该函数 $\mathscr{F}:([0,1]^2)^n\to\mathbb{R}$ 满足如下有界差分性质。对于任意$k\in[n]$，$(\bar{X}_1,\bar{Y}_1)\ldots,(\bar{X}_k,\bar{Y}_k)\ldots,(\bar{X}_n,\bar{Y}_n)$ 和 $(\bar{X}_k',\bar{Y}_k')\in[0,1]^2$，注意到至多有 $C(\log n)^{\kappa}$ 个定义 $\mathscr{F}$ 的求和项 $i\in[n]$ 可以发生变化，并对每个求和项应用统一的上界和下界，我们有

$$\left|\mathscr{F}\left((\bar{X}_1,\bar{Y}_1),\ldots,(\bar{X}_k,\bar{Y}_k),\ldots,(\bar{X}_n,\bar{Y}_n)\right)-\mathscr{F}\left((\bar{X}_1,\bar{Y}_1),\ldots,(\bar{X}_k',\bar{Y}_k'),\ldots,(\bar{X}_n,\bar{Y}_n)\right)\right|\le C'(\log n)^{2\kappa},$$

$\bar{}$ $\bar{}$几乎必然成立（排除 Y_k 或 Y_k' 拥有多于 $(\log n)^{\kappa}$ 个邻居的零测事件）。因此，对于任意常数 $t>0$，根据有界差分不等式，

$$\mathbb{P}\left(\frac{\mathscr{F}-\mathbb{E}\mathscr{F}}{n}>t\right)\le\exp\left(-\frac{cnt^2}{(\log n)^{4\kappa}}\right). \tag{27}$$

我们可以将$\frac{1}{n}\mathbb{E}\mathscr{F}$ 上界估计如下:

$$\begin{aligned}\frac{1}{n}\mathbb{E}\mathscr{F}&\le\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\log\left[\exp\left(C'+\frac{n}{2}\left|\bar{X}_i-\bar{Y}_i\right|\right)+\sum_{j=1}^{n}\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right)\right]\\&\le\frac{1}{n}\sum_{i=1}^{n}\log\left[\mathbb{E}\exp\left(C'+\frac{n}{2}\left|\bar{X}_i-\bar{Y}_i\right|\right)+\sum_{j=1}^{n}\mathbb{E}\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right)\right]\end{aligned}$$

对于 $i\ne j$，应用命题2.1 可验证

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right)\,\middle|\,\bar{X}_i\right]=\int_0^1 p_n(y)\exp\left(-\frac{1}{2}V(n(\bar{X}_i-y))\right)\mathrm{d}y\le\frac{C}{n}.$$

另外，对 $\mathbb{E}[\exp(|\varepsilon|/2)]$ 和 $\mathbb{E}[\exp(-V(\varepsilon)/2)]$（$i=j$）应用一个常数界，可得$\frac{1}{n}\mathbb{E}\mathscr{F}\le C$，其中 $C>0$ 为常数。结合 (26) 和 (27)，这表明 $\mathcal{F}_n(1/2)\le C'$ 以高概率成立。再结合 (19)，即可得到所需的 $\mathcal{D}_n(1)\le 2(\mathcal{F}_n(1/2)-\mathcal{F}_n(1))\le C''$。 □

接下来，我们证明在对后验测度施加局部限制的情况下也有类似的界。条件于数据 X， Y，回顾一下在排序索引上将 π 与 π_{st} 识别的过程。对每个 $l \in \mathbb{Z}$，定义边界变量

$$\Gamma_{l+\frac{1}{2}}(\pi) = \{(k, \pi_{st}(k)) \colon k \le l, \pi_{st}(k) > l\} \cup \{(k, \pi_{st}(k)) \colon k > l, \pi_{st}(k) \le l\} \tag{28}$$

（按定义，$\Gamma_{l+\frac{1}{2}} = \varnothing$ 如果 $l \le 0$ 或 $l \ge n$）。注意， $\widehat{P}_{\mathrm{sort}i}(j)$ 在算法 2中计算的恰好是后验概率 $P(\pi_{st}(i) = j \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing)$，条件是左右两侧的边界 $KL + \frac{1}{2}$ 都为空。

引理 B.2。 固定整数 $K, L > 0$。对于每个 $i \in [n]$，记 $i_- = \max(i - KL, 1)$ 和 $i_+ = \min(i + KL, n)$。那么存在一个常数 $C > 0$，它不依赖于 K, L，使得

$$\frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{KL}\sum_{j=i_-}^{i_+} V(n(X_{s(j)} - Y_{\pi(s(j))})) \,\middle|\, X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing\right] \le C \ \ w.h.p.$$

证明。 在给定 X, Y 的条件下，定义

$$P_\beta^i(\pi \mid X, Y) = \frac{1}{\mathcal{Z}_i(\beta)} \exp\left(-\beta \sum_{j=i_-}^{i_+} V(n(X_{s(j)} - Y_{\pi(s(j))}))\right), \quad \mathcal{F}_i(\beta) = \frac{1}{KL}\log \mathcal{Z}_i(\beta)$$

将其对应的期望记为 $E_\beta^i[\cdot \mid X, Y]$。那么待界定的量恰好是 $\frac{1}{n}\sum_{i=1}^n \mathcal{D}_n^i$，其中

$$\mathcal{D}_n^i := E_{\beta=1}^i\left[\frac{1}{KL}\sum_{j=i_-}^{i_+} V(n(X_{s(j)} - Y_{\pi(s(j))})) \,\middle|\, X, Y\right] = -\mathcal{F}_i'(1) \le 2(\mathcal{F}_i(1/2) - \mathcal{F}_i(1)).$$

这里 $\frac{1}{n}\sum_{i=1}^n \mathcal{F}_i(1/2)$ 可以按照引理 B.1 的证明进行上界估计：令 A^i 为行/列索引为 $\{i_-, \ldots, i_+\}$、元素为 $A_{jl}^i = \exp\left(-\frac{1}{2}V(n(X_{s(j)} - Y_{t(l)}))\right)$ 的矩阵。那么

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathcal{F}_i(1/2) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{KL}\log \operatorname{perm}(A^i) \\
&\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+}\log\sum_{l=i_-}^{i_+}\exp\left(-\frac{1}{2}V(n(X_{s(j)} - Y_{t(l)}))\right) \\
&\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+}\log\sum_{l=1}^{n}\exp\left(-\frac{1}{2}V(n(X_{s(j)} - Y_{t(l)}))\right) \\
&\le \frac{C}{n}\sum_{i=1}^{n}\log\sum_{l=1}^{n}\exp\left(-\frac{1}{2}V(n(X_{s(i)} - Y_{t(l)}))\right).
\end{aligned}$$

Th右侧与 (20) 相同，因此应用引理B.1可得

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{F}_i(1/2) \le C \text{ w.h.p.} \tag{29}$$

$\frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_i(1)$ 的一个下界可通过将对双射 π_{st} 的求和 $[i_-, i_+] \to [i_-, i_+]$ 下界化为单个求和项 $\pi_{st} = \mathrm{id}$ 得到：

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_i(1) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\log\exp\left(-\sum_{j=i_-}^{i_+}V\left(n(X_{s(j)}-Y_{t(j)})\right)\right)$$
$$\geq -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+}V\left(n(X_{s(j)}-Y_{t(j)})\right) \geq -\frac{C}{n}\sum_{i=1}^{n}V\left(n(X_{s(i)}-Y_{t(i)})\right). \tag{30}$$

根据假设2.3(ii)，存在一个常数 $C \geq 1+\delta$ 使得

$$\frac{1}{n}\sum_{i=1}^{n}V\left(n(X_{s(i)}-Y_{t(i)})\right) \leq C + \frac{C}{n}\sum_{i=1}^{n}|n(X_{s(i)}-Y_{t(i)})|^C$$
$$\overset{(*)}{=} C + \frac{C}{n}\min_{\pi\in\mathbb{S}_n}\sum_{i=1}^{n}|n(X_i-Y_{\pi(i)})|^C \leq C + \frac{C}{n}\sum_{i=1}^{n}|n(\bar{X}_i-\bar{Y}_i)|^C. \tag{31}$$

这里，$(*)$ 用到了凸函数 $g(x) = (n|x|)^C$ 的重排不等式。（对于任意 $x_1 \leq x_2$和 $y_1 \leq y_2$，令 $r = x_1 - y_2$，$s = x_2 - y_1$，$p = x_1 - y_1$，$q = x_2 - y_2$。于是 $p + q = r + s$ 且$p, q \in [r, s]$，这意味着存在 $\alpha \in [0,1]$，使得 $p = \alpha r + (1-\alpha)s$, $q = (1-\alpha)r + \alpha s$。因此$g(x_1 - y_1) + g(x_2 - y_2) = g(p) + g(q) \leq g(r) + g(s) = g(x_1 - y_2) + g(x_2 - y_1)$。）回忆一下，$\varepsilon_i := n(Y_i - X_i)$ 满足尾界 (25)，因此 (31) 的右侧以高概率有一个常数$C' > 0$ 作为上界。将此结果用于 (30) 可得

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_i(1) \geq -C \text{ w.h.p.}$$

结合（29），这意味着所需的界 $\frac{1}{n}\sum_{i=1}^{n}\mathcal{D}_n^i \leq C'$ 以高概率 □

B.1.2 点过程的正则性与 Gibbs 测度的局部性

固定（较大的）局部性参数 K、$L > 0$ 并回顾下界 $\Lambda(x) \geq \Lambda_{\min}$，对每个 $l \in [n]$成立。为方便起见，我们重复第3 节中的事件定义。定义（X、Y）相关事件

$$\mathcal{A}_l = \left\{\sup_{j:|j-(l+\frac{1}{2})|\leq L}\{n|X_{s(j)}-X_{s(l)}|, n|Y_{t(j)}-Y_{t(l)}|\} \leq \frac{3L}{2\Lambda_{\min}},\ n\left|X_{s(l)}-Y_{t(l)}\right| \leq L\right\}.$$

在给定（X、Y）的条件下，定义与 π 相关的事件

$$\mathcal{C}_l \equiv \mathcal{C}_l(\pi) = \left\{|k-m| \leq L \text{ for all } (k,m) \in \Gamma_{l+\frac{1}{2}}(\pi)\right\},$$
$$\mathcal{L}_l \equiv \mathcal{L}_l(\pi) \equiv \mathcal{A}_l \cap \mathcal{C}_l(\pi).$$

这里，$\mathcal{C}_l(\pi)$ 是一个“局部性”事件，即边界集合 $\Gamma_{l+\frac{1}{2}}(\pi)$ 中的所有匹配，根据数据值的排序索引，均属于短程（$\leq L$）。

对每个 $i \in [n]$，定义事件

$$\mathcal{G}_i(\pi) = \left\{\sum_{k=1}^{KL}\mathbf{1}\{\mathcal{L}_{i-k}(\pi)\} \geq \frac{2KL}{3}\right\} \cap \left\{\sum_{k=0}^{KL-1}\mathbf{1}\{\mathcal{L}_{i+k}(\pi)\} \geq \frac{2KL}{3}\right\}.$$

我们用 $\mathcal{G}_i^c(\pi)$ 表示其补集，并约定若 $i \leq KL$ or$i > n - KL$，则 $\mathcal{G}_i(\pi)$ 不成立。本节将证明以下引理。

引理 B.3。对于一个不依赖于 K, L 的常数 $C > 0$，以高概率成立。

$$\frac{1}{n}\sum_{i=1}^{n} P(\mathcal{G}_i^c(\pi) \mid X, Y) \le CL^{-\delta}, \qquad \frac{1}{n}\sum_{i=1}^{n} P(\mathcal{G}_i^c(\pi) \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing) \le CL^{-\delta}.$$

首先，为了证明对于大多数位置 $i \in [n]$，$\mathcal{A}_i$ 以高概率成立，我们使用有界密度下顺序统计量的以下尾界。

引理 B.4。对于任意 $1 \le i < j \le n$,

$$\mathbb{P}\left(X_{s(j)} - X_{s(i)} \le \frac{j-i}{2\Lambda_{\max} n}\right) \le \exp\left(-\frac{j-i}{8}\right), \quad \mathbb{P}\left(X_{s(j)} - X_{s(i)} \ge \frac{3(j-i)}{2\Lambda_{\min} n}\right) \le \exp\left(-\frac{j-i}{8}\right).$$

证明。令 F 为 X 的累积分布函数，并设 $U_i := F(X_i)$。则 $U_i \sim$服从 $\mathrm{Unif}[0,1]$，且 $U_{s(1)} < \cdots < U_{s(n)}$。此外，

$$U_{s(j)} - U_{s(i)} \sim \mathrm{Beta}(j-i, n-(j-i)+1).$$

注意到 $X_{s(i)} = F^{-1}\left(U_{s(i)}\right)$，并且由于 $\Lambda(x)$ 有上下界，$\frac{1}{\Lambda_{\max}} \le \left(F^{-1}\right)'(u) \le \frac{1}{\Lambda_{\min}}$。因此，对于任意 ~~$u,$~~ $v \in [0,1]$,

$$\frac{1}{\Lambda_{\max}}|u-v| \le \left|F^{-1}(u) - F^{-1}(v)\right| \le \frac{1}{\Lambda_{\min}}|u-v|.$$

由此可得

$$\begin{aligned}\mathbb{P}\left(X_{s(j)} - X_{s(i)} \le \frac{j-i}{2\Lambda_{\max} n}\right) &\le \mathbb{P}\left(U_{s(j)} - U_{s(i)} \le \frac{j-i}{2n}\right) \\ &\le \exp\left(-\frac{(i-j)^2/4n^2}{2(j-i)(n-(j-i)+1)/n^3}\right) \le \exp\left(-\frac{j-i}{8}\right)\end{aligned}$$

其中第二个不等式是 Beta 分布的尾界 [34, 定理 1]。类似地,

$$\mathbb{P}\left(X_{s(j)} - X_{s(i)} \ge \frac{3(j-i)}{2\Lambda_{\min} n}\right) \le \mathbb{P}\left(U_{s(j)} - U_{s(i)} \ge \frac{3(j-i)}{2n}\right) \le \exp\left(-\frac{j-i}{8}\right). \qquad \square$$

以下两个引理接下来将控制不满足 $\mathcal{A}_i$ 的位置所占比例 $i \in [n]$.

引理 B.5. 定义

$$\mathcal{B}_1 \equiv \mathcal{B}_1(L) = \left\{i \in [n-L] \colon X_{s(i+L)} - X_{s(i)} > \frac{3L}{2\Lambda_{\min} n}\right\},$$
$$\mathcal{B}_1' \equiv \mathcal{B}_1'(L) = \left\{i \in [n-L] \colon Y_{t(i+L)} - Y_{t(i)} > \frac{3L}{2\Lambda_{\min} n}\right\}.$$

Then

$$\frac{1}{n}\max(|\mathcal{B}_1|, |\mathcal{B}_1'|) \le 2e^{-L/8} \ \textit{w.h.p.}$$

证明。注意，基数 $|\mathcal{B}_1|$ 作为无序点 $X_1, \ldots, X_n$ 的函数满足有界差分性质:

$$\left||\mathcal{B}_1(\bar{X}_1, \cdots, \bar{X}_i, \cdots, \bar{X}_n)| - |\mathcal{B}_1(\bar{X}_1, \cdots, \bar{X}_i', \cdots, \bar{X}_n)|\right| \le 2L$$

对任意 $X_1,\cdots,X_i,X_i'\cdots,X_n\in[0,1]$。因此，由有界差分不等式，对任意$t>0$,

$$\mathbb{P}\left(\frac{|\mathcal{B}_1|-\mathbb{E}|\mathcal{B}_1|}{n}\geq t\right)\leq\exp\left(-\frac{nt^2}{2L^2}\right).$$

此外，根据引理 B.4,

$$\frac{1}{n}\mathbb{E}|\mathcal{B}_1|=\frac{1}{n}\sum_{i=1}^{n-L}\mathbb{P}\left(X_{s(i+L)}-X_{s(i)}>\frac{3L}{2\Lambda_{\min}n}\right)\leq e^{-L/8}.$$

取 $t=e^{-L/8}$ 即可得到 $\mathcal{B}_1$ 的界，而 $\mathcal{B}_1'$ 的界同理。 □

引理 B.6。定义

$$\mathcal{B}_2\equiv\mathcal{B}_2(L)=\left\{i\in[n]:n\left|X_{s(i)}-Y_{t(i)}\right|>L\right\}.$$

对于某些与 L 无关的常数 C、$c>0$，

$$\frac{1}{n}|\mathcal{B}_2|\leq Ce^{-cL}\ w.h.p.$$

证明。注意到 $\left\{n\left|X_{s(i)}-Y_{t(i)}\right|\geq L\right\}=\mathcal{E}_1^i\cup\mathcal{E}_2^i$，其中

$$\mathcal{E}_1^i=\left\{\left|\left\{j:Y_j<X_{s(i)}-\frac{L}{n}\right\}\right|\geq i\right\},\qquad\mathcal{E}_2^i=\left\{\left|\left\{j:Y_j>X_{s(i)}+\frac{L}{n}\right\}\right|\geq n-i+1\right\}\tag{32}$$

对于 $\mathcal{E}_1^i$，我们有

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_1^i\}&=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\sum_{j=1}^{n}\mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(j)}<X_{s(i)}-\frac{L}{n}\right\}\geq i\right\}\\&\leq\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\sum_{j=i}^{n}\mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(j)}<X_{s(i)}-\frac{L}{n}\right\}\geq 1\right\}\\&\leq\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(j)}<X_{s(i)}-\frac{L}{n}\right\}\\&=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(j)}-X_{s(j)}<\underbrace{X_{s(i)}-X_{s(j)}}_{\leq 0}-\frac{L}{n}\right\}\\&\leq(A)+(B)+(C)\end{aligned}\tag{33}$$

其中

$$\begin{aligned}(A)&:=\frac{1}{n}\sum_{i=1}^{n}\sum_{j:\,i\leq j<i+L}\mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(j)}-X_{s(j)}<-\frac{L}{n}\right\},\\(B)&:=\frac{1}{n}\sum_{i=1}^{n}\sum_{j:\,j\geq i+L}\mathbf{1}\left\{X_{s(j)}-X_{s(i)}\leq\frac{j-i}{2\Lambda_{\max}n}\right\},\\(C)&:=\frac{1}{n}\sum_{i=1}^{n}\sum_{j:\,j\geq i+L}\mathbf{1}\left\{n\left(Y_{(\pi^*)^{-1}\circ s(j)}-X_{s(j)}\right)\leq-\frac{j-i}{2\Lambda_{\max}}-L\right\}.\end{aligned}$$

令 $\varepsilon_i = n(Y_i - X_i) = n(Y_i - X_{\pi^*(i)})$，注意到

$$(A) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, i\le j<i+L} \mathbf{1}\left\{\varepsilon_{(\pi^*)^{-1}\circ s(j)} < -L\right\} \le \frac{L}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\varepsilon_i < -L\right\}.$$

根据尾界（25）和二项分布尾不等式，有

$$(A) \le CLe^{-cL^{1+\delta}} \le C'e^{-c'L} \quad \text{w.h.p.} \tag{34}$$

对于（B），固定任意常数 $\kappa > 1$，由引理 B.4 可知

$$\sum_{i=1}^{n}\sum_{j:\, j-i\ge(\log n)^\kappa} \mathbb{P}\left(X_{s(j)} - X_{s(i)} \le \frac{j-i}{2\Lambda_{\max} n}\right) \le n^2\exp\left(-(\log n)^\kappa/8\right) < \exp(-c(\log n)^\kappa). \tag{35}$$

因此，以 $1-\exp(-c(\log n)^\kappa)$ 的概率，

$$(B) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, j-i\in[L,(\log n)^\kappa)} \mathbf{1}\left\{X_{s(j)} - X_{s(i)} \le \frac{j-i}{2\Lambda_{\max} n}\right\}.$$

该量在 $X_1, \ldots, X_n$ 中满足有界差分性质，其中改变任意 X_i 至多会使其值改变 $2(\log n)^\kappa/n$。由引理 B.4，该量的期望至多为

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j-i\in[L,(\log n)^\kappa)} \exp\left(-\frac{j-i}{8}\right) \le Ce^{-cL}.$$

因此，结合有界差分不等式，可得

$$(B) \le Ce^{-cL} \text{ w.h.p.} \tag{36}$$

类似地，再次记 $\varepsilon_i = n\left(Y_i - X_{\pi^*(i)}\right)$，并注意到尾界（25）意味着，对于任意$\kappa > 1$,

$$\sum_{i=1}^{n}\sum_{j:\, j-i\ge(\log n)^\kappa} \mathbb{P}\left(\varepsilon_{(\pi^*)^{-1}\circ s(j)} \le -\frac{j-i}{2\Lambda_{\max}} - L\right) \le n^2\exp\left(-c(\log n)^{\kappa(1+\delta)}\right) \le \exp\left(-c'(\log n)^{\kappa(1+\delta)}\right),$$

我们以概率 $1-\exp(-c'(\log n)^{\kappa(1+\delta)})$ 有

$$(C) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j:\, j-i\in[L,(\log n)^\kappa)} \mathbf{1}\left\{\varepsilon_{(\pi^*)^{-1}\circ s(j)} \le -\frac{j-i}{2\Lambda_{\max}} - L\right\}.$$

在条件 $X_1, \ldots, X_n$（因此也在排序 $s\colon [n] \to [n]$ 的条件下，该量在 $\varepsilon_1, \ldots, \varepsilon_n$ 中满足有界差分性质，因为改变任意 ε_i 至多会使其值变化 $(\log n)^\kappa/n$。再次利用尾界 (25) 对其期望（在 $\bar{X}$ 的条件下）进行界定，并应用有界差分不等式，可得

$$(C) \le Ce^{-cL^{1+\delta}} \text{ w.h.p.} \tag{37}$$

结合 (34)、(36) 和 (37) 可得 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_1^i\} \le Ce^{-cL}$ 以高概率成立。对于 $\mathcal{E}_2^i$，注意到

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_2^i\} \le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\sum_{j=1}^{i}\left\{Y_{(\pi^*)^{-1}\circ s(j)} > X_{s(i)} + \frac{L}{n}\right\} \ge 1\right\}.$$

然后，同样的论证可用于界定 $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_2^i\}$。 □

接下来，我们分析在 $l \in [n]$ 下不满足 $\mathcal{C}_l(\pi)$ 的$P(\pi \mid X, Y)$ 位置的期望比例，或在后验分布的局部限制下的位置期望比例。这需要用到此前引理B.1 和B.2中得到的平均势界。

引理 B.7。 固定常数 K、$L > 0$，并令 $i_- = \max(i - KL, 1)$ 且 $i_+ = \min(i + KL, n)$。那么，对于一个不依赖于 K、L 的常数 $C > 0$，以高概率成立。

$$\frac{1}{n}\sum_{i=1}^{n} P\left(\mathcal{C}_i^c(\pi) \mid X, Y\right) \le CL^{-\delta}, \quad \frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l=i_-}^{i_+-1} P(\mathcal{C}_l^c(\pi) \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing) \le CL^{-\delta}.$$

证明。 对于全局吉布斯测度，注意到

$$\mathcal{C}_i^c(\pi) = \{\text{there exists } j : |\pi_{st}(j) - j| > L,\ j \le i < \pi_{st}(j) \text{ or } \pi_{st}(j) \le i < j\}.$$

Then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{C}_i^c(\pi)\} &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\{|\pi_{st}(j) - j| > L,\ j \le i < \pi_{st}(j) \text{ or } \pi_{st}(j) \le i < j\} \\ &\le \frac{1}{n}\sum_{i,j=1}^{n}(A_{ij}) + (B_{ij}) + (C_{ij})\end{aligned}$$

其中

$$(A_{ij}) = \mathbf{1}\{|\pi_{st}(j) - j| > L,\ j \le i < \pi_{st}(j) \text{ or } \pi_{st}(j) \le i < j\}\mathbf{1}\left\{|Y_{t\circ\pi_{st}(j)} - Y_{t(j)}| \le \frac{|\pi_{st}(j) - j|}{2\Lambda_{\max} n}\right\}$$

$$(B_{ij}) = \mathbf{1}\{|\pi_{st}(j) - j| > L,\ j \le i < \pi_{st}(j) \text{ or } \pi_{st}(j) \le i < j\}\mathbf{1}\left\{|X_{s(j)} - Y_{t(j)}| > \frac{|\pi_{st}(j) - j|}{4\Lambda_{\max} n}\right\}$$

$$(C_{ij}) = \mathbf{1}\{|\pi_{st}(j) - j| > L,\ j \le i < \pi_{st}(j) \text{ or } \pi_{st}(j) \le i < j\}\mathbf{1}\left\{|Y_{t\circ\pi_{st}(j)} - X_{s(j)}| > \frac{|\pi_{st}(j) - j|}{4\Lambda_{\max} n}\right\}$$

记 $(A) = \frac{1}{n}\sum_{i,j=1}^{n}(A_{ij})$、$(B) = \frac{1}{n}\sum_{i,j=1}^{n}(B_{ij})$ 和 $(C) = \frac{1}{n}\sum_{i,j=1}^{n}(C_{ij})$。

任取 $\kappa > 1$，我们可以用一个与 π 无关的量来界定 (A)

$$\begin{aligned}(A) &\le \frac{2}{n}\sum_{i=1}^{n}\sum_{\substack{j:j\le i \\ |k-j|>L}}\sum_{k:k>i}\mathbf{1}\left\{\left|Y_{t(k)} - Y_{t(j)}\right| \le \frac{|k-j|}{2\Lambda_{\max} n}\right\} \\ &= \frac{2}{n}\sum_{i=1}^{n}\sum_{\substack{j:j\le i \\ (\log n)^\kappa \ge |k-j|>L}}\sum_{k:k>i}\mathbf{1}\left\{\left|Y_{t(k)} - Y_{t(j)}\right| \le \frac{|k-j|}{2\Lambda_{\max} n}\right\} \\ &\quad + \frac{2}{n}\sum_{i=1}^{n}\sum_{\substack{j:j\le i \\ |k-j|>(\log n)^\kappa}}\sum_{k:k>i}\mathbf{1}\left\{\left|Y_{t(k)} - Y_{t(j)}\right| \le \frac{|k-j|}{2\Lambda_{\max} n}\right\}.\end{aligned}$$

应用引理 B.4 并对所有指标元组进行并集界（i、j、k），第二项以概率 $1 - Ce^{-c(\log n)^\kappa}$ 为 0。第一项在 $\bar{Y}_1$、.....、$\bar{Y}_n$ 上满足有界差分性质，

即改变任意一个 $\bar{Y}_i$，其值的变化至多为 $\frac{C}{(\log n)^{2\kappa}}$。此外，根据引理 B.4，第一项的期望至多为

$$\frac{2}{n}\sum_{i=1}^{n}\sum_{\substack{j:j\le i\\ |k-j|>L}}\sum_{k:k>i} Ce^{-c|k-j|} \le \frac{C'}{n}\sum_{i=1}^{n}\Big(\sum_{j\in[i-L,i]} e^{-cL} + \sum_{j:j<i-L} e^{-c|i-j|}\Big) \le C''e^{-c'L}. \tag{38}$$

然后，应用有界差分不等式，我们得到

$$(A) \le Ce^{-cL} \text{ w.h.p.} \tag{39}$$

类似地，我们也可以用一个与 π无关的量来界定（B）

$$(B) \le \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j:j\le i\\ |k-j|>L}}\sum_{k:k>i} \mathbf{1}\left\{\left|X_{s(j)}-Y_{t(j)}\right| > \frac{|k-j|}{4\Lambda_{\max}n}\right\} + \mathbf{1}\left\{\left|X_{s(k)}-Y_{t(k)}\right| > \frac{|k-j|}{4\Lambda_{\max}n}\right\}$$

与引理 B.6 中相同的论证表明，对于任意固定的 $\kappa > 1$，以至少为$1 - Ce^{-c(\log n)^{\kappa}}$的概率，所有满足 $|k-j| \ge (\log n)^{\kappa}$ 的求和项在上述右侧均为 0。于是在这一高概率事件上，

$$\begin{aligned}(B) &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}} \mathbf{1}\left\{\left|X_{s(j)}-Y_{t(j)}\right| > \frac{|k-j|}{4\Lambda_{\max}n}\right\}\\ &\le \underbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}} \mathbf{1}\left\{\left|\left\{l : Y_l < X_{s(j)} - \frac{|k-j|}{4\Lambda_{\max}n}\right\}\right| \ge j\right\}}_{:=(B_1)}\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}} \underbrace{\mathbf{1}\left\{\left|\left\{l : Y_l > X_{s(j)} + \frac{|k-j|}{4\Lambda_{\max}n}\right\}\right| \ge n-j+1\right\}}_{:=(B_2)},\end{aligned}$$

其中最后一个不等式以类似于（32）的方式分解该事件。应用（33）的界，我们有

$$(B_1) \le \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}}\sum_{l=j}^{n} \mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(l)} \le X_{s(j)} - \frac{|k-j|}{4\Lambda_{\max}n}\right\} \le (B_1') + (B_1'')$$

其中

$$(B_1') := \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}}\sum_{l=j}^{n} \mathbf{1}\left\{X_{s(l)} - X_{s(j)} \le \frac{l-j}{2\Lambda_{\max}n},\ Y_{(\pi^*)^{-1}\circ s(l)} - X_{s(l)} \le -\frac{|k-j|}{4\Lambda_{\max}n}\right\}$$

$$(B_1'') := \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{i-(\log n)^{\kappa}\le j,k\le i+(\log n)^{\kappa}\\ |k-j|>L}}\sum_{l=j}^{n} \mathbf{1}\left\{Y_{(\pi^*)^{-1}\circ s(l)} - X_{s(l)} \le -\frac{l-j}{2\Lambda_{\max}n} - \frac{|k-j|}{4\Lambda_{\max}n}\right\}.$$

此处，$n(Y_{(\pi^*)^{-1}\circ s(l)} - X_{s(l)}) = \varepsilon_{(\pi^*)^{-1}\circ s(l)}$。同样，以至少 $1 - Ce^{-c(\log n)^\kappa}$ 的概率，根据并集界，定义 (B_1') 且满足 $l - j \geq (\log n)^\kappa$ 的求和项均为 0。在这个高概率事件上，

$$(B_1') = \frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{i-(\log n)^\kappa \leq j,k \leq i+(\log n)^\kappa \\ |k-j|>L}} \sum_{l\in[j,j+(\log n)^\kappa]} \mathbf{1}\left\{X_{s(l)} - X_{s(j)} \leq \frac{l-j}{2\Lambda_{\max} n},\ \varepsilon_{(\pi^*)^{-1}\circ s(l)} \leq -\frac{|k-j|}{4\Lambda_{\max}}\right\}.$$

在条件 $(X_1, \ldots, X_n)$（因而也在排序 $s\colon [n] \to [n]$）下，它作为 $\varepsilon_1, \ldots, \varepsilon_n$ 的函数满足有界差分性质，其中改变 ε_i 至多可使该值改变 $C(\log n)^{3\kappa}/n$。由尾界 (25)，它对 $\varepsilon_1, \ldots, \varepsilon_n$ 的条件期望上界为

$$\frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{i-(\log n)^\kappa \leq j,k \leq i+(\log n)^\kappa \\ |k-j|>L}} \sum_{l\in[j,j+(\log n)^\kappa]} \mathbf{1}\left\{X_{s(l)} - X_{s(j)} \leq \frac{l-j}{2\Lambda_{\max} n}\right\} \cdot Ce^{-c|k-j|^{1+\delta}}.$$

这进而作为 $X_1, \ldots, X_n$ 的函数满足有界差分性质，其中改变 $\bar{X}_i$ 至多可使该值改变 $C'(\log n)^{3\kappa}/n$。由引理 B.4 以及与 (38) 类似的论证，它对 $\bar{X}_1, \ldots, \bar{X}_n$ 的期望上界为 $Ce^{-cL^{1+\delta}}$。然后，两次应用有界差分不等式，先证明关于 $\varepsilon_1, \ldots, \varepsilon_n$ 的集中性，再证明关于 $X_1, \ldots, X_n$ 的集中性，我们得到

$$(B_1') \leq Ce^{-cL^{1+\delta}} \text{ with probability } 1 - Ce^{-c(\log n)^\kappa}.$$

同样，对于 (B_1'')，以至少 $1 - Ce^{-c(\log n)^\kappa}$ 的概率，当 $l - j \geq (\log n)^\kappa$ 时，求和项均为 0，因此

$$(B_1'') = \frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{i-(\log n)^\kappa \leq j,k \leq i+(\log n)^\kappa \\ |k-j|>L}} \sum_{l\in[j,j+(\log n)^\kappa]} \mathbf{1}\left\{\varepsilon_{\pi_*^{-1}\circ s(l)} \leq -\frac{l-j}{2\Lambda_{\max}} - \frac{|k-j|}{4\Lambda_{\max}}\right\}$$

在 $\varepsilon_1, \ldots, \varepsilon_n$ 的条件下，关于 $X_1, \ldots, X_n$，这满足有界差分性质。由尾界（25）及类似于（38）的论证，其条件均值至多为 $Ce^{-cL^{1+\delta}}$，因此

$$(B_1'') \leq Ce^{-cL^{1+\delta}} \text{ with probability } 1 - Ce^{-c(\log n)^\kappa}.$$

量（B_2）可用类似方法加以界定，结合这些界可得

$$(B) \leq Ce^{-cL^{1+\delta}} \text{ w.h.p.} \tag{40}$$

为了界定最后一项（C），注意到

$$\begin{aligned}(C) &= \frac{1}{n}\sum_{j=1}^{n} |\pi_{st}(j) - j| \cdot \mathbf{1}\{|\pi_{st}(j) - j| > L\}\mathbf{1}\left\{|Y_{\pi\circ s(j)} - X_{s(j)}| > \frac{|\pi_{st}(j) - j|}{4\Lambda_{\max} n}\right\} \\ &\leq 4\Lambda_{\max} \cdot \frac{1}{n}\sum_{j=1}^{n} n|Y_{\pi\circ s(j)} - X_{s(j)}| \cdot \mathbf{1}\left\{n|Y_{\pi\circ s(j)} - X_{s(j)}| > \frac{L}{4\Lambda_{\max}}\right\}\end{aligned}$$

注意，对于任意非负随机变量 Z，根据 H¨older 不等式

$$\mathbb{E}[Z\mathbf{1}\{Z>a\}] \le (\mathbb{E}Z^{1+\delta})^{\frac{1}{1+\delta}}\mathbb{P}[Z>a]^{\frac{\delta}{1+\delta}} \le (\mathbb{E}Z^{1+\delta})^{\frac{1}{1+\delta}}\left(\frac{1}{a^{1+\delta}}\mathbb{E}Z^{1+\delta}\right)^{\frac{\delta}{1+\delta}} \le \frac{\mathbb{E}Z^{1+\delta}}{a^{\delta}}. \tag{41}$$

引理 B.1 与下界 $V(\varepsilon) - V_{\min} \ge c|\varepsilon|^{1+\delta}$ 蕴含

$$\frac{1}{n}\sum_{j=1}^{n} E\left[\left(n|Y_{\pi\circ s(j)} - X_{s(j)}|\right)^{1+\delta} \;\middle|\; X, Y\right] \le C \text{ w.h.p.}$$

因此，应用（41），

$$E[(C) \mid X, Y] \le 4\Lambda_{\max}\cdot\frac{C}{(L/4\Lambda_{\max})^{\delta}} \le C'L^{-\delta} \text{ w.h.p.} \tag{42}$$

结合（39）、（40）和（42），

$$\frac{1}{n}\sum_{i=1}^{n} P(\mathcal{C}_i^c(\pi) \mid X, Y) \le Ce^{-cL} + Ce^{-cL^{1+\delta}} + CL^{-\delta} \le C'L^{-\delta} \text{ w.h.p.}$$

局部吉布斯测度的论证类似：我们希望界定

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l=i_-}^{i_+-1} P(\mathcal{C}_l^c(\pi) \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing)$$

对每个固定的 $i \in [n]$，我们有如上所述

$$\frac{1}{KL}\sum_{l=i_-}^{i_+-1}\mathbf{1}\{\mathcal{C}_i^c(\pi)\} \le \frac{1}{KL}\sum_{l,j=i_-}^{i_+}(A_{lj}) + (B_{lj}) + (C_{lj})$$

注意，我们有与 π 无关的界

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l,j=i_-}^{i_+}(A_{lj}) + (B_{lj}) \le C[(A) + (B)] \le C'e^{-cL} \text{ w.h.p.}$$

根据（39）和（40）。为了界定涉及（C_{lj}）的最后一项，采用全局吉布斯测度下与上述论证类似的方法，我们注意到

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l,j=i_-}^{i_+} E\left[(C)_{lj} \;\middle|\; X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing\right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+} E\Bigg[|\pi_{st}(j) - j|\,\mathbf{1}\{|\pi_{st}(j) - j| > L\} \\
&\qquad\qquad \mathbf{1}\left\{|Y_{\pi\circ s(j)} - X_{s(j)}| > \frac{|\pi_{st}(j) - j|}{4\Lambda_{\max}n}\right\} \Bigg| X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing\Bigg] \\
&\le 4\Lambda_{\max}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+} E\left[n|Y_{\pi\circ s(j)} - X_{s(j)}|\mathbf{1}\left\{n|Y_{\pi\circ s(j)} - X_{s(j)}| > \frac{L}{4\Lambda_{\max}}\right\} \Bigg| X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing\right] \\
&\le 4\Lambda_{\max}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{j=i_-}^{i_+}\frac{E[(n|Y_{\pi\circ s(j)} - X_{s(j)}|)^{1+\delta} \mid X, Y, \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing]}{(L/4\Lambda_{\max})^{\delta}}.
\end{aligned}$$

应用引理B.2中的局部平均势界，可得

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l,j=i_-}^{i_+}E\left[(C)_{lj}\;\middle|\;X,Y,\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing\right]\le CL^{-\delta}\text{ w.h.p.}$$

并结合上文可得

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{KL}\sum_{l=i_-}^{i_+-1}P(\mathcal{C}_l^c(\pi)\mid X,Y,\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)\le Ce^{-cL}+CL^{-\delta}\le C'L^{-\delta}\text{ w.h.p.}$$

□

我们现在完成引理 B.3 的证明。

引理 *B.3* 的证明。由马尔可夫不等式，对每个 $i\in[KL+1,n-KL]$,

$$\begin{aligned}\mathcal{G}_i^c(\pi)&=\left\{\sum_{k=1}^{KL}\mathbf{1}\left\{\mathcal{L}_{i-k}^c(\pi)\right\}\ge\frac{KL}{3}\right\}\cup\left\{\sum_{k=0}^{KL-1}\mathbf{1}\left\{\mathcal{L}_{i+k}^c(\pi)\right\}\ge\frac{KL}{3}\right\}\\&\le\frac{3}{KL}\sum_{j=i_-}^{i_+-1}\mathbf{1}\left\{\mathcal{L}_j^c(\pi)\right\}\le\frac{3}{KL}\sum_{j=i_-}^{i_+-1}\left(\mathbf{1}\left\{\mathcal{A}_j^c\right\}+\mathbf{1}\left\{\mathcal{C}_j^c(\pi)\right\}\right)\end{aligned}$$

于是，对于全局吉布斯测度，由引理 B.5、B.6 和 B.7,

$$\frac{1}{n}\sum_{i=1}^{n}P(\mathcal{G}_i^c(\pi)\mid X,Y)\le\frac{2KL}{n}+\frac{6}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{A}_i^c\}+\frac{6}{n}\sum_{i=1}^{n}P(\mathcal{C}_i^c(\pi)\mid X,Y)\le CL^{-\delta}\text{ w.h.p.}$$

类似地，对于局部吉布斯测度,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}P(\mathcal{G}_i^c(\pi)\mid X,Y,\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)\\&\le\frac{2KL}{n}+\frac{6}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{A}_i^c\}+\frac{1}{n}\sum_{i=1}^{n}\frac{3}{KL}\sum_{j=i_-}^{i_+-1}P(\mathcal{C}_j^c(\pi)\mid X,Y,\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)\le CL^{-\delta}\text{ w.h.p.}\end{aligned}$$ □

B.1.3 相关性衰减

现在，我们利用前述局部性估计证明后验分布中的相关性衰减，从而完成定理 2.9的证明。我们需要以下初等观察。

引理 B.8. 固定 $i\in[n]$。在后验分布 $P(\pi\mid X,Y)$ 下，给定 $\Gamma_{i+\frac{1}{2}}(\pi)=\varnothing$ 后，$(\pi_{st}(k)\colon\ k\le i)$ 中的取值与$(\pi_{st}(k):\ k>i)$相互独立。

证明。 任何满足 $\Gamma_m(\pi)=\varnothing$ 的 π 都可以分解为 $\pi=\pi_{\le i}\oplus\pi_{>i}$，其中 $\pi_{\le i}$ 是 $\{1,\dots,i\}$ 上的双射，而 $\pi_{>i}$ 是 $\{i+1,\dots,n\}$ 上的双射。对于此类 π，哈密顿量可分解为

$$H(\pi)=H(\pi_{\le i})+H(\pi_{>i}):=\sum_{j=1}^{i}V\left(n\left(X_j-Y_{\pi(j)}\right)\right)+\sum_{j=i+1}^{n}V\left(n\left(X_j-Y_{\pi(j)}\right)\right)$$

因此，

$$P\left(\pi \mid \Gamma_{i+\frac{1}{2}}(\pi)=\varnothing, X, Y\right)=\frac{\exp\left(-H\left(\pi_{\leq i}\right)\right)}{\sum_{\pi_{\leq i}} \exp\left(-H\left(\pi_{\leq i}\right)\right)} \frac{\exp\left(-H\left(\pi_{>i}\right)\right)}{\sum_{\pi_{>i}} \exp\left(-H\left(\pi_{>i}\right)\right)},$$

这意味着所需的条件独立性。 □

定理 2.9 的证明。 固定 $i \in [n]$，令 π, π' 为从后验分布$P(\pi \mid X, Y)$中独立抽取的两个样本。在整个证明中，我们以 X, Y为条件，并将 $P(\cdot)$记作 $P(\cdot \mid X, Y)$。定义位置列表

$$\begin{aligned}
\mathcal{Q}_i^-(\pi) &= \{k \in \{i-KL, \ldots, i-1\} : \mathcal{L}_k(\pi) \text{ holds}\}, \\
\mathcal{Q}_i^+(\pi) &= \{k \in \{i, \ldots, i+KL-1\} : \mathcal{L}_k(\pi) \text{ holds}\}, \\
\mathcal{Q}_i(\pi) &= \mathcal{Q}_i^-(\pi) \cup \mathcal{Q}_i^+(\pi),
\end{aligned}$$

并定义 si　类似地，$\mathcal{Q}_i^-(\pi')\mathcal{Q}_i^+(\pi')$ $\mathcal{Q}_i(\pi')$ 对于 π'。在事件 $\mathcal{G}_i(\pi) \cap \mathcal{G}_i(\pi')$ 发生时，我们必有

$$|\mathcal{Q}_i^-(\pi) \cap \mathcal{Q}_i^-(\pi')| \geq \frac{KL}{3}, \qquad |\mathcal{Q}_i^+(\pi) \cap \mathcal{Q}_i^+(\pi')| \geq \frac{KL}{3}.$$

对于任意 $m \in \{i-KL, \ldots, i+KL-1\}$，令 $\mathcal{F}_m(\pi)$ 表示由 $\mathcal{Q}_i(\pi)$ 和边界变量列表给出的信息集

$$\{\Gamma_{l+\frac{1}{2}}(\pi) : l \in \mathcal{Q}_i(\pi) \setminus [m-L, m+L]\}.$$

我们给出一个下界

$$P(\Gamma_{m+\frac{1}{2}}(\pi)=\varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi)).$$

考虑任意满足 $\mathcal{L}_m(\pi)$ 的 π——也就是说， $\mathcal{A}_m$ 和 $\mathcal{C}_m(\pi)$ 均成立——且 $\Gamma_{m+\frac{1}{2}}(\pi) = \Gamma \neq \varnothing$ 的情形。回忆一下，每个 $(k,l) \in \Gamma$ 都表示一个索引对，其中 $X_{s(k)}$ 与 $Y_{t(l)}$ 匹配，并且 k, l 中恰有一个属于集合 $\{1, \ldots, m\}$，也恰有一个属于集合 $\{m+1, \ldots, n\}$。记作

$$\Gamma = \{(k_{-N}, l_{-N}), \ldots, (k_{-1}, l_{-1}), (k_1, l_1), \ldots, (k_N, l_N)\} \tag{43}$$

其中我们按递增顺序对 $k_{-N} < \ldots < k_{-1} < k_1 < \ldots < k_N$ 排序。重要的是，满足 $k \leq m$ 和 $l > m$ 的元组 $(k,l) \in \Gamma$ 的数量，等于满足 $k > m$ 和$l \leq m$ 的元组数量，因为 m 表示 X 值和 Y 值中相同的排名索引，而 π 是一个双射。因此，恰有 N 个索引 $k_{-N} < \ldots < k_{-1} \leq m$，且恰有 N 个索引 $m < k_1 < \ldots < k_N$。我们从 π 构造出匹配 $\tilde{\pi}$ ： 除 (43) 中的匹配外，其余匹配均保持不变，并按排序顺序将 (43) 中的索引 $k_{-N} < \ldots < k_N$ 与 $\{l_{-N}, \ldots, l_N\}$ 匹配。于是根据构造，$\Gamma_{m+\frac{1}{2}}(\tilde{\pi}) = \varnothing$，所以 $\mathcal{L}_m(\tilde{\pi})$ 也成立。此外，由于 $\mathcal{L}_m(\pi)$ 成立， $\tilde{\pi}$ 与 π 在值域 $\leq L$ 上具有相同的匹配，也就是说， $|\pi_{st}(j) - j| \leq L$ 当且仅当 $|\tilde{\pi}_{st}(j) - j| \leq L$。因此，对于所有 $l \in [n]$， $\mathcal{L}_l(\pi) = \mathcal{L}_l(\tilde{\pi})$，于是 $\mathcal{Q}_i(\pi) = \mathcal{Q}_i(\tilde{\pi})$ 也成立。此外，对于每个 $l \in \mathcal{Q}_i(\pi) \setminus [m-L, m+L]$， $\Gamma_{l+\frac{1}{2}}(\pi) = \Gamma_{l+\frac{1}{2}}(\tilde{\pi})$，因此

$$\mathcal{F}_m(\pi) = \mathcal{F}_m(\tilde{\pi}).$$

对每个 $(k, \pi_{st}(k)) \in \Gamma_{m+\frac{1}{2}}$应用该条件，其中

$$\begin{aligned}
n|X_{s(k)} - Y_{t(\tilde{\pi}_{st}(k))}| &\leq n|X_{s(k)} - X_{s(m)}| + n|X_{s(m)} - Y_{t(m)}| + n|Y_{t(m)} - Y_{t(\tilde{\pi}_{st}(k))}| \\
&\leq \frac{3L}{2\Lambda_{\min}} + L + \frac{3L}{2\Lambda_{\min}} \leq \frac{4L}{\Lambda_{\min}}
\end{aligned}$$

根据定义， $\mathcal{A}_m \cap \mathcal{C}_m(\pi) \cap \mathcal{C}_m(\tilde{\pi}) = \mathcal{L}_m(\pi) \cap \mathcal{L}_m(\tilde{\pi})$成立；再结合界 $g(|\varepsilon|) + V_{\min} \geq V(\varepsilon) \geq V_{\min}$，可得

$$\begin{aligned}\frac{\exp(-H(\pi))}{\exp(-H(\tilde{\pi}))} &= \exp\left(\sum_{(k,\pi_{st}(k))\in\Gamma} V(n(X_{s(k)} - Y_{t(\tilde{\pi}_{st}(k))})) - V(n(X_{s(k)} - Y_{t(\pi_{st}(k))}))\right) \\ &\leq \exp\left(|\Gamma| g(4L/\Lambda_{\min})\right) \end{aligned} \tag{44}$$

对于每个固定的此类边界条件 Γ，从 π 到 $\tilde{\pi}$ 的这个映射是单射，因为我们可以通过重新指定匹配来恢复π （43）中定义的 Γ。因此，这表明

$$\frac{P(\Gamma_m(\pi) = \Gamma \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi))}{P(\Gamma_m(\pi) = \varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi))} \leq \exp\left(|\Gamma| g(4L/\Lambda_{\min})\right).$$

对所有与 $\mathcal{L}_m(\pi)$ 兼容的可能 $\Gamma \neq \varnothing$ 求和，其中满足 $|\Gamma| = 2N$ 的情况至多有$\binom{L}{N}^4 (N!)^2$ 种，进而得到

$$\begin{aligned}&\frac{1 - P(\Gamma_m(\pi) = \varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi))}{P(\Gamma_m(\pi) = \varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi))} \leq \sum_{\Gamma\neq\varnothing} \exp\left(|\Gamma| g(4L/\Lambda_{\min})\right) \\ &\leq \sum_{N=1}^{L} \binom{L}{N}^4 (N!)^2 \exp(2N[g(4L/\Lambda_{\min})]) \leq \exp(CL[g(4L/\Lambda_{\min}) + \log L]) \leq \exp(C' L\, g(4L/\Lambda_{\min})),\end{aligned}$$

最后两个界分别由斯特林近似得出常数 $C > 0$ ，以及由下界 $g(|\varepsilon|) \geq c|\varepsilon|^{\delta}$得出常数 $C' > 0$ 。因此

$$\begin{aligned}P(\Gamma_m(\pi) = \varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi)) &\geq \frac{1}{1 + \exp(C' L\, g(4L/\Lambda_{\min}))} \\ &\geq \exp(-C'' L\, g(4L/\Lambda_{\min})) =: \iota(L).\end{aligned} \tag{45}$$

注意，上述论证同样适用于在给定$\Gamma_{i\pm(KL+\frac{1}{2})} = \varnothing$条件下 π 的后验分布，因此

$$P(\Gamma_m(\pi) = \varnothing \mid \mathcal{L}_m(\pi), \mathcal{F}_m(\pi), \Gamma_{i\pm(KL+\frac{1}{2})}(\pi) = \varnothing) \geq \iota(L). \tag{46}$$

现在，在给定 $\mathcal{Q}_i(\pi)$ 和 $\mathcal{Q}_i(\pi')$ 的条件下，令 $J^+(\pi, \pi')$ 表示公共位点 $\{k_1, \ldots, k_{J^+(\pi,\pi')}\} \subset \mathcal{Q}_i^+(\pi) \cap \mathcal{Q}_i^+(\pi')$ 中最长序列的长度，其中每个增量 $k_{j+1} - k_j \geq L$。定义事件 $\mathcal{E}_j \equiv \mathcal{E}_j(\pi, \pi') = \{\Gamma_{k_j}(\pi) = \Gamma_{k_j}(\pi') = \varnothing\}$。于是

$$\begin{aligned}&P\Big(\exists\, m \in \mathcal{Q}_i^+(\pi) \cap \mathcal{Q}_i^+(\pi') : \Gamma_m(\pi) = \Gamma_m(\pi') = \varnothing \;\Big|\; \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi')\Big) \\ &\geq \underbrace{P(\mathcal{E}_1 \mid \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi'))}_{:=\iota_1} + \underbrace{P(\mathcal{E}_2 \mid \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi'), \mathcal{E}_1^c)}_{:=\iota_2} P(\mathcal{E}_1^c \mid \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi')) \\ &\qquad + \underbrace{P(\mathcal{E}_3 \mid \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi'), \mathcal{E}_1^c, \mathcal{E}_2^c)}_{:=\iota_3} P(\mathcal{E}_1^c, \mathcal{E}_2^c \mid \mathcal{Q}_i^+(\pi), \mathcal{Q}_i^+(\pi')) + \ldots \\ &= \sum_{j=1}^{J^+(\pi,\pi')} \iota_j (1-\iota_1)\ldots(1-\iota_{j-1}) = 1 - \prod_{j=1}^{J^+(\pi,\pi')} (1-\iota_j) \geq 1 - (1-\iota(L)^2)^{J^+(\pi,\pi')},\end{aligned}$$

最后一个不等式对每个 $j=1,\ \ldots\ldots,\ J^+(\pi,\pi')$ 均适用，这是由 (45) 推出的。注意，在事件 $\mathcal{G}_i(\pi)\cap\mathcal{G}_i(\pi')$ 发生且 $|\mathcal{Q}_i^+(\pi)\cap\mathcal{Q}_i^+(\pi')|\geq KL/3$ 时，我们有 $J^+(\pi,\pi')\geq K/3$。因此

$$P\Big(\exists\, m\in\mathcal{Q}_i^+(\pi)\cap\mathcal{Q}_i^+(\pi'):\Gamma_m(\pi)=\Gamma_m(\pi')=\varnothing\ \Big|\ \mathcal{G}_i(\pi)\cap\mathcal{G}_i(\pi')\Big)\geq 1-(1-\iota(L)^2)^{K/3}. \tag{47}$$

用（46）替代（45），同样

$$\begin{aligned}&P\Big(\exists\, m\in\mathcal{Q}_i^+(\pi)\cap\mathcal{Q}_i^+(\pi'):\Gamma_m(\pi)=\Gamma_m(\pi')=\varnothing\ \Big|\ \mathcal{G}_i(\pi)\cap\mathcal{G}_i(\pi'),\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing\Big)\\&\geq 1-(1-\iota(L)^2)^{K/3},\end{aligned}$$

对于 $\mathcal{Q}_i^-(\pi)\cap\mathcal{Q}_i^-(\pi')$，类似的界也成立。令

$$\begin{aligned}M_-\equiv M_-(\pi,\pi')&=\min\{m:m\in\mathcal{Q}_i^-(\pi)\cap\mathcal{Q}_i^-(\pi),\,\Gamma_m(\pi)=\Gamma_m(\pi')=\varnothing\},\\M_+\equiv M_+(\pi,\pi')&=\max\{m:m\in\mathcal{Q}_i^+(\pi)\cap\mathcal{Q}_i^+(\pi),\,\Gamma_m(\pi)=\Gamma_m(\pi')=\varnothing\},\end{aligned}$$

约定：如果不存在这样的索引，则令 $M_-=0$ or $M_+=0$。于是可得

$$P\Big(M_-\neq 0\text{ and }M_+\neq 0\ \Big|\ \mathcal{G}_i(\pi)\cap\mathcal{G}_i(\pi')\Big)\geq 1-2(1-\iota(L)^2)^{K/3}, \tag{48}$$

$$P\Big(M_-\neq 0\text{ and }M_+\neq 0\ \Big|\ \mathcal{G}_i(\pi)\cap\mathcal{G}_i(\pi'),\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing\Big)\geq 1-2(1-\iota(L)^2)^{K/3}. \tag{49}$$

现在令 $P_{\text{排序}i}$，并令 $\widehat{P}_i^{\text{sort}}$分别表示在实际后验分布和算法 2 的局部后验下，$X_{s(i)}$的匹配的后验概率。记 π、π' 为从 $P(\pi\mid X,\ Y)$中进行的两次独立同分布抽样，

$$\begin{aligned}\mathsf{TV}\left(\widehat{P}_i^{\text{sort}},P_i^{\text{sort}}\right)&=\frac{1}{2}\sum_{j=1}^{n}\Big|P(\pi_{st}(i)=j\mid\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)-P(\pi'_{st}(i)=j)\Big|\\&\leq\frac{1}{2}\sum_{j=1}^{n}\Bigg|\sum_{m\neq 0}\sum_{m'\neq 0}P(\pi_{st}(i)=j,M_-=m,M_+=m'\mid\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)\\&\qquad\qquad-P(\pi'_{st}(i)=j,M_-=m,M_+=m')\Bigg|\\&\qquad+P(M_-=0\text{ or }M_+=0)+P(M_-=0\text{ or }M_+=0\mid\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing).\end{aligned}$$

对于任意固定的 $m,\ m'\neq 0$，注意到

$$\begin{aligned}&P(\pi_{st}(i)=j\mid M_-=m,M_+=m',\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing)=P(\pi_{st}(i)=j\mid M_-=m,M_+=m')\\&=P(\pi'_{st}(i)=j\mid M_-=m,M_+=m')=P(\pi'_{st}(i)=j\mid M_-=m,M_+=m',\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing),\end{aligned}$$

其中，第一个和最后一个等号成立，是因为 $\pi(i)$ 和 $\pi'(i)$ 在给定 $\Gamma_{i\pm(KL+\frac{1}{2})}(\pi)$ 的条件下相互独立，且 $M_-=m\neq 0$ 和 $M_+=m'\neq 0$ (根据引理 B.8)；第二个等号则由

(π, π') 在此条件律下的可交换性。于是

$$\begin{aligned}
\mathsf{TV}\left(\widehat{P}_i^{\mathrm{sort}}, P_i^{\mathrm{sort}}\right) \le \frac{1}{2}\sum_{j=1}^{n}\Bigg|\sum_{m\neq 0}\sum_{m'\neq 0} P(\pi'_{st}(i)=j, M_-=m, M_+=m' \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) \\
- P(\pi'_{st}(i)=j, M_-=m, M_+=m')\Bigg| \\
+ P(M_-=0 \text{ or } M_+=0) + P(M_-=0 \text{ or } M_+=0 \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) \\
\le \frac{1}{2}\sum_{j=1}^{n}\left|P(\pi'_{st}(i)=j \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) - P(\pi'_{st}(i)=j)\right| \\
+ 2P(M_-=0 \text{ or } M_+=0) + 2P(M_-=0 \text{ or } M_+=0 \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) \\
= 2P(M_-=0 \text{ or } M_+=0) + 2P(M_-=0 \text{ or } M_+=0 \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing),
\end{aligned}$$

最后一个等号成立，是因为 π' 与 π相互独立。最后，界限 (48–49) 表明

$$\begin{aligned}
P(M_-=0 \text{ or } M_+=0) &\le 2(1-\iota(L)^2)^{K/3} + P(\mathcal{G}_i(\pi)^c) + P(\mathcal{G}_i(\pi')^c) \\
P(M_-=0 \text{ or } M_+=0 \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) &\le 2(1-\iota(L)^2)^{K/3} \\
&\quad + P(\mathcal{G}_i(\pi)^c \mid \Gamma_{i\pm(KL+\frac{1}{2})}(\pi)=\varnothing) + P(\mathcal{G}_i(\pi')^c).
\end{aligned}$$

然后应用引理 B.3 对位置取平均，以高概率

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}\left(\widehat{P}_i^{\mathrm{sort}}, P_i^{\mathrm{sort}}\right) \le C(1-\iota(L)^2)^{K/3} + CL^{-\delta}$$

这与定理的陈述等价。 □

B.2 定理 2.11 的证明

B.2.1 局部弱收敛到一个极限点过程

我们回顾点过程弱收敛的一些定义；更全面的讨论请参见 [31, 第 3], 或 [24, 15, 16]。

考虑一个度量空间 E 及其Borel σ-代数 $\mathcal{E}$。 E 上的点测度是一个测度 m，使得存在可数点集 $\{x_i \in E:\ i \in I \subseteq \mathbb{N}\}$，且 $m = \sum_{i\in I}\delta_{x_i}$。令M $=$ M(E)为 E上所有点测度的集合。令 $\mathcal{M} = \mathcal{M}$($E$)为使所有求值映射 $m \mapsto m(A)$对所有 $A \in \mathcal{E}$均可测的最小 σ-代数。点过程 η 是(M $\mathcal{M}$)中的随机元素（即从某个底层概率空间(Ω $\mathcal{F}$)到(M $\mathcal{M}$)的可测映射）。我们可以将 η 等同于 E中的可数随机点集 $\{X_i:\ i \in I\}$，从而$\eta = \sum_{i\in I}\delta_{X_i}$。点过程 η 的强度测度 μ 是 E 上由 $\mu(A) = \mathbb{E}\eta(A)$对所有$A \in \mathcal{E}$给出的测度。若对每个有界的 $A \in \mathcal{E}$都有 $\mathbb{P}(\eta(A) < \infty) = 1$，则称 η 是局部有限的。

定义 B.9。 设 μ 是 E上的无原子测度。强度测度为 μ的泊松点过程，是指定义在 E 上的点过程 η，满足对于任意一组两两不相交的集合 A_1、……、 $A_k \subseteq E$，随机变量 $\eta(A_1)$、……、 $\eta(A_k)$相互独立，且 $\eta(A_\ell) \sim$服从参数为 $\mu(A_\ell)$的泊松分布。记作 $\eta \sim$PPP(μ)。

在 E为欧几里得空间且强度测度具有形式 $\mu(\cdot) = \lambda\cdot$Leb$(\cdot)$的情形下，其中$\lambda \ge 0$ 为速率参数，Leb为勒贝格测度，我们称 η 为速率为 λ 的齐次泊松点过程，并记作 $\eta \sim$PPP(λ)。

定义 B.10. 设 η 为 E上的点过程，并设 $\mathbb{R}_+(E)$表示 E上非负实值可测函数的空间。 η 的拉普拉斯泛函是由 $\Psi_\eta(u) = \mathbb{E}\exp\left\{-\int_E u(x)\ \eta(\mathrm{d}x)\right\}$给出的映射 $\Psi_\eta\colon \mathbb{R}_+(E) \to [0,1]$，其中期望是对$\eta$的随机性取的。

这是一个标准结果：点过程 η 的分布律 $\eta\sim$为 PPP(μ)，当且仅当

$$\Psi_\eta(u) = \exp\left\{-\int\left[1-e^{-u(x)}\right]\mu(\mathrm{d}x)\right\},\quad u\in\mathbb{R}_+(E). \tag{50}$$

当 $\eta\sim$PPP(μ)，且 μ 本身是一个随机测度时， η 的分布律是一个 混合泊松点过程（更一般地说，是一个 Cox 过程）。其拉普拉斯泛函满足

$$\Psi_\eta(u) = \mathbb{E}_\mu\exp\left\{-\int\left[1-e^{-u(x)}\right]\mu(\mathrm{d}x)\right\},\quad u\in\mathbb{R}_+(E), \tag{51}$$

其中期望是对随机测度 μ 取的。

令 $C_c^+(E)$ 表示 E 上连续、紧支撑且取非负实值的函数空间。任何点过程都是一个随机拉东测度，因此 $\eta(u) = \int u(x)\eta(\mathrm{d}x)$ 几乎必然有定义，对任意 $u\in C_c^+(E)$ 均如此。对于 $\{\eta_n\}, \eta\in$ M，当 $\eta_n(u)\to\eta(u)$ 对所有 $u\in C_c^+(E)$ 都成立时，我们记作 $\eta_n \overset{\mathrm{v}}{\longrightarrow}\eta$（模糊收敛）。这定义了模糊收敛拓扑；根据命题 3.17], 该拓扑可度量化，使 M 成为一个波兰空间 [31, 。这使得我们能够定义点过程的弱收敛概念。令 C_b(M) 表示 M 上连续（相对于该模糊收敛度量）且取有界实值的函数空间。

定义 B.11. 设 $\{\eta_n\}$和 η 是 E上的点过程。若序列 $\{\eta_n\}$ 弱收敛到 η，记作 $\eta_n \overset{\mathrm{w}}{\longrightarrow}\eta$，则对所有 $\phi\in C_b$(M)， $\mathbb{E}\phi(\eta_n)\to\mathbb{E}\phi(\eta)$。

命题 B.12. *[31,* 命题 *3.19]* 设 $\{\eta_n\}$ 和 η 是 E上的点过程。以下两个断言等价：*(i)* $\eta_n \xrightarrow{w}\eta$ 和 *(ii)* $\Psi_{\eta_n}(u)\to\Psi_\eta(u)$ 对所有 $u\in C_c^+(E)$成立。

对我们的讨论而言，以下关于欧几里得空间上的点过程弱收敛的判据很有用。

命题 B.13. 设 $\{\eta_n\}$ 和 η 是 $E=\mathbb{R}^2$上的点过程。假设

(i) 对所有紧集 $U\subset\mathbb{R}^2$，$\sup_n\mathbb{E}[\eta_n(U)]<\infty$，并且

(ii) 对于每个有理矩形简单函数 $s=\sum_{j=1}^m a_j\mathbf{1}_{R_j}$，其中 $a_j\geq 0$、 $m\in\mathbb{N}$ 和 R_j 是端点为有理数且两两不相交的矩形，我们有 $\Psi_{\eta_n}(s)\to\Psi_\eta(s)$。

那么，对于所有 $u\in C_c^+(\mathbb{R}^2)$， $\Psi_{\eta_n}(u)\to\Psi_\eta(u)$。

证明。固定一个紧集 $U\subset\mathbb{R}^2$。设 u 和 $\widetilde{u}$ 是两个非负有界可测函数，其支集的并包含于 U 中。我们先证明 Ψ_n 和 Ψ 满足利普希茨性质。对于每个 n， $|\exp(-\eta_n(u)) - \exp(-\eta_n(\widetilde{u}))|\leq\eta_n(|u-\widetilde{u}|)$。取期望后，得到 $\sup_n|\Psi_{\eta_n}(u)-\Psi_{\eta_n}(\widetilde{u})|\leq\|u-\widetilde{u}\|_\infty\sup_n\mathbb{E}\eta_n(U)$。类似地，我们有 $|\Psi_\eta(u)-\Psi_\eta(\widetilde{u})|\leq\|u-\widetilde{u}\|_\infty\mathbb{E}\eta(U)$。

固定 $u\in C_c^+(\mathbb{R}^2)$，并令 U 表示 u的紧支集。由于 u 在 U上连续，因此它一致连续且有界。于是，对每个 $\epsilon>0$，都存在一个有理矩形简单函数 s，使得 $s\geq 0$， $\mathrm{supp}(s)\subset U$，且 $\|u-s\|_\infty<\min\{\epsilon/(3C_U)\ \epsilon/(3\mathbb{E}\eta(U))\}$，其中$C_U=\sup_n\mathbb{E}\eta_n(U)$。存在 $n_0(\epsilon)$，使得对所有 $n\geq n_0(\epsilon)$， $|\Psi_{\eta_n}(s)-\Psi_\eta(s)|\leq\epsilon/3$。由前述讨论可知，对于 $n\geq n_0$，

$$|\Psi_{\eta_n}(u)-\Psi_\eta(u)|\leq|\Psi_{\eta_n}(u)-\Psi_{\eta_n}(s)|+|\Psi_{\eta_n}(s)-\Psi_\eta(s)|+|\Psi_\eta(u)-\Psi_\eta(s)|\leq\epsilon.\qquad\square$$

利用上述刻画，我们建立点过程局部弱收敛的概念

$\{n(X_i - X_I,\ Y_i - X_I)\}_{i=1}^n$，以 $\{(X_i,\ Y_i)\}_{i=1}^n$ 为条件，其随机性由索引的均匀随机选择所诱导 $I \in [n]$。（这类似于随机图的局部弱收敛 [1]，其中局部邻域的分布由在给定图实现的条件下随机选择一个顶点所诱导。）

局部弱极限由 $(\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda, \pi_\Lambda^*)$ 给出，如定义 2.10 所构造，我们将其隐式等同于 $\mathbb{R}^2$ 上的点过程，其中包含随机点 $\{(\mathsf{X}\pi_\Lambda^*(\mathsf{X})) : \mathsf{X} \in \mathbf{X}_\Lambda\}$。

引理 B.14. 在 $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ 的条件下，设 $I \in [n]$ 为均匀随机索引。则当 $n \to \infty$ 时几乎必然，

$$\{n(\bar{X}_i - \bar{X}_I, \bar{Y}_i - \bar{X}_I)\}_{i=1}^n \xrightarrow{w} (\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda, \pi_\Lambda^*)\,.$$

证明。设 R_1，……，R_m 为 $\mathbb{R}^2$ 中满足 $R_\ell = B_\ell \times C_\ell$ 的互不相交有界矩形，设 a_1，……，$a_m \geq 0$，并设 $s = \sum_{\ell=1}^m a_\ell \mathbf{1}_{R_\ell}$。设 η 表示 $\mathbb{R}^2$ 上代表 $(\mathbf{X}_\Lambda, \mathbf{Y}_\Lambda, \pi_\Lambda^*)$ 的点过程。我们首先计算拉普拉斯泛函 $\Psi_\eta(s)$：记 $\eta = \eta_0 + \bar{\eta}$，其中 η_0 由单个点 $(0, \pi_\Lambda^*(0)) \equiv (0, \varepsilon)$ 组成，$\bar{\eta}$ 由所有其余点组成，且它们相互独立。则

$$\begin{aligned}
\Psi_\eta(s) &= \mathbb{E}\exp\left(-\int s(x,y)\eta_0(\mathrm{d}x\mathrm{d}y) - \int s(x,y)\bar{\eta}(\mathrm{d}x\mathrm{d}y)\right) \\
&= \mathbb{E}_{\varepsilon\sim q}[e^{-s(0,\varepsilon)}] \cdot \underbrace{\mathbb{E}\exp\left(-\sum_{\ell=1}^m a_\ell \sum_{\mathsf{X}\in\mathbf{X}_\Lambda\setminus\{0\}} \mathbf{1}_{\{\mathsf{X}\in B_\ell\}}\mathbf{1}_{\{\pi_\Lambda^*(\mathsf{X})\in C_\ell\}}\right)}_{=\Psi_{\bar{\eta}}(s)}.
\end{aligned}$$

在$\mathbf{X}_\Lambda$的条件下，由于 $\{\pi_\Lambda^*(\mathsf{X}) : \mathsf{X} \in \mathbf{X}_\Lambda \setminus\{0\}\}$ 相互独立，且 $\pi_\Lambda^*(\mathsf{X}) \sim q(\cdot - \mathsf{X})$，我们有

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\sum_{\ell=1}^m a_\ell \sum_{\mathsf{X}\in\mathbf{X}_\Lambda\setminus\{0\}} \mathbf{1}_{\{\mathsf{X}\in B_\ell\}}\mathbf{1}_{\{\pi_\Lambda^*(\mathsf{X})\in C_\ell\}}\right) \,\middle|\, \mathbf{X}_\Lambda\right] \\
&= \prod_{\mathsf{X}\in\mathbf{X}_\Lambda\setminus\{0\}} \int \exp\left(-\sum_{\ell=1}^m a_\ell \mathbf{1}_{\{\mathsf{X}\in B_\ell\}}\mathbf{1}_{\{y\in C_\ell\}}\right) q(y-\mathsf{X})\mathrm{d}y.
\end{aligned}$$

那么，由于 $\mathbf{X}_\Lambda \setminus\{0\}$ 的分布律是混合泊松点过程，对51)应用（$\mathbf{X}_\Lambda$ 取期望，得到

$$\begin{aligned}
\Psi_{\bar{\eta}}(s) &= \mathbb{E}_{X\sim\Lambda}\exp\left\{-\int\left[1 - \int\exp\left(-\sum_{\ell=1}^m a_\ell \mathbf{1}_{\{x\in B_\ell\}}\mathbf{1}_{\{y\in C_\ell\}}\right) q(y-x)\mathrm{d}y\right]\Lambda(X)\mathrm{d}x\right\} \\
&= \mathbb{E}_{X\sim\Lambda}\exp\left\{-\Lambda(X)\iint\left[1 - \exp\left(-\sum_{\ell=1}^m a_\ell \mathbf{1}_{\{x\in B_\ell\}}\mathbf{1}_{\{y\in C_\ell\}}\right)\right] q(y-x)\mathrm{d}x\,\mathrm{d}y\right\} \\
&= \mathbb{E}_{X\sim\Lambda}\exp\left\{\Lambda(X)\sum_{\ell=1}^m (e^{-a_\ell} - 1)\iint_{R_\ell} q(y-x)\mathrm{d}x\,\mathrm{d}y\right\},
\end{aligned}$$

最后一个等式利用了 R_1、…、R_m 是 $\mathbb{R}^2$ 的互不相交子集这一事实。根据 (51) 和命题 B.13，注意到这表明二元过程 $(\mathbf{X}_\Lambda \setminus\{0\}, \mathbf{Y}_\Lambda \setminus\{\mathsf{Y}_0\}, \pi_\Lambda^*)$ 也是一个混合泊松点过程，其随机强度测度为 $\mu_X(\mathrm{d}x\,\mathrm{d}y) = \Lambda(X)q(y-x)\mathrm{d}x\,\mathrm{d}y$ 和 $X \sim \Lambda$。于是

$$\Psi_\eta(s) = \mathbb{E}_{\varepsilon\sim q}[e^{-s(0,\varepsilon)}] \cdot \mathbb{E}_{X\sim\Lambda}\exp\left\{\Lambda(X)\sum_{\ell=1}^m (e^{-a_\ell} - 1)\iint_{R_\ell} q(y-x)\mathrm{d}x\,\mathrm{d}y\right\}.$$

现在令 η_n 为点集为 $\{n(\bar{X}_i - \bar{X}_I,\ \bar{Y}_i - \bar{X}_I)\}_{i=1}^n$ 的点过程。定义 $\mathcal{N}_{(i)}(R_\ell) = |\{j \in [n] : j \neq i,\ n(X_j - X_i,\ Y_j - X_i) \in R_\ell\}|$，并令

$$\mathcal{N}^{(i)} := \left(\mathcal{N}^{(i)}(R_1), \ldots, \mathcal{N}^{(i)}(R_m)\right), \qquad a = (a_1, \ldots, a_m), \qquad \varepsilon_i = n(\bar{Y}_i - \bar{X}_i). \tag{52}$$

Then

$$\Psi_{\eta_n}(s) := \mathbb{E}\left[\exp\left(-\int s(x,y)\eta_n(\mathrm{d}x\mathrm{d}y)\right) \,\middle|\, \mathcal{D}_n\right] = \frac{1}{n}\sum_{i=1}^n \exp\left(-s(0,\varepsilon_i) - a^\top \mathcal{N}^{(i)}\right)$$

我们首先证明 $\mathbb{E}\Psi_{\eta_n}(s)$（对 $\mathcal{D}_n$取期望）收敛到 $\Psi_\eta(s)$，然后证明$\Psi_{\eta_n}(s)$集中于其均值附近。注意，

$$\mathbb{E}\Psi_{\eta_n}(s) = \mathbb{E}\Big[e^{-s(0,\varepsilon_i)}\,\mathbb{E}[e^{-a^\top \xi} \mid \bar{X}_i]^{n-1}\Big]$$

其中右侧对所有 $i \in [n]$ 都相同，而 ξ 表示一个属于 $\{e_1, \ldots, e_m\} \cup \{0\}$ 的向量，其在给定 X_i 条件下的分布为

$$\mathbb{P}\left[\xi = 0 \mid \bar{X}_i\right] = 1 - \sum_{\ell=1}^m \rho_{\ell,n}(\bar{X}_i), \qquad \mathbb{P}\left[\xi = e_\ell \mid \bar{X}_i\right] = \rho_{\ell,n}(\bar{X}_i) \quad \text{for } 1 \le \ell \le m,$$

$$\begin{aligned}
\rho_{\ell,n}(\bar{X}_i) &= \mathbb{P}\left(n\left(\bar{X}_j - \bar{X}_i, \bar{Y}_j - \bar{X}_i\right) \in R_\ell \mid \bar{X}_i\right) \text{ for any } j \neq i \\
&= \frac{1}{\mathcal{Z}_n}\int_{(\bar{X}_i,\bar{X}_i)+\frac{R_\ell}{n}} p_n(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{n^2\mathcal{Z}_n}\int_{B_\ell}\mathrm{d}u\int_{C_\ell}\mathrm{d}\varepsilon\sqrt{\Lambda\left(\bar{X}_i + \frac{x}{n}\right)\Lambda\left(\bar{X}_i + \frac{y}{n}\right)}\exp\left(-V(y-x)\right).
\end{aligned} \tag{53}$$

根据命题2.1，对任意固定的 $X_i \in (0,1)$，当 $n \to \infty$时，

$$n\rho_{\ell,n}(\bar{X}_i) \to \Lambda(\bar{X}_i)\iint_{R_\ell} q(y-x)\mathrm{d}x\,\mathrm{d}y,$$

且 $(X_i,\ \varepsilon_i)$ 的分布律收敛到 $(X,\ \varepsilon)$ 的分布律 $\sim \Lambda \times q$。因此，根据控制收敛定理，

$$\begin{aligned}
\mathbb{E}\Psi_{\eta_n}(s) &= \mathbb{E}\Big[e^{-s(0,\varepsilon_i)}\,\mathbb{E}[e^{-a^\top\xi} \mid \bar{X}_i]^{n-1}\Big] \\
&= \mathbb{E}\left[e^{-s(0,\varepsilon_i)}\left(1 + \sum_{\ell=1}^m \left(e^{-t_\ell} - 1\right)\rho_{\ell,n}(\bar{X}_i)\right)^{n-1}\right] \\
&\to \mathbb{E}_{\varepsilon\sim q}[e^{-s(0,\varepsilon)}]\cdot\mathbb{E}_{X\sim\Lambda}\left[\exp\left(\Lambda(X)\sum_{\ell=1}^m(e^{-t_\ell}-1)\iint_{R_\ell} q(y-x)\,\mathrm{d}x\,\mathrm{d}y\right)\right] = \Psi_\eta(s).
\end{aligned} \tag{54}$$

接下来，我们证明 $\Psi_{\eta_n}(s) = n^{-1}\sum_{i=1}^n \exp(-s(0,\varepsilon_i) - a^\top\mathcal{N}^{(i)})$ 关于 $\mathcal{D}_n$ 的集中性。令 $\bigcup_{\ell=1}^m R_\ell \subset W := [-R,R]^2$，其中 $0 < R < \infty$。定义以下量

$$D_j := \left|\left\{i \in [n] : i \neq j,\ n(\bar{X}_j - \bar{X}_i, \bar{Y}_j - \bar{X}_i) \in W\right\}\right| \text{ for } j = 1, \cdots, n$$

以及事件 $G_n := \{\max_j D_j < 3\log n\}$。在给定 (X_j, Y_j) 的条件下，$D_j \sim$服从 Binomial$(n-1, p_j)$，其中 $p_j \le \frac{2R\Lambda_{\max}}{n}$。应用并集界和切尔诺夫界可得

$$\mathbb{P}(G_n^c) = \sum_{i=1}^{n} \mathbb{P}\left(D_i \ge 3\log n\right) \le n^{-3} \sum_{i=1}^{n} \mathbb{E}e^{D_i} \le n^{-2} e^{8R\Lambda_{\max}}.$$

如果 $\{(X_1, Y_1) \ldots, (X_n, Y_n)\}$ 和 $\{(X_1', Y_1') \ldots, (X_n', Y_n')\}$ 都满足 G_n，且仅有一对不同，设为 $(\bar{X}_j, \bar{Y}_j) \ne (\bar{X}_j', \bar{Y}_j')$，那么相应的计数向量 $\mathcal{N}^{(i)}$ 和 $\mathcal{N}^{(i)\prime}$ 仅当 $n(\bar{X}_j - \bar{X}_i, \bar{Y}_j - \bar{X}_i)$ 和/或 $n(\bar{X}_j' - \bar{X}_i', \bar{Y}_j' - \bar{X}_i')$ 属于 W 时才会不同。因此，$\Psi_{\eta_n}(s)$ 的差异至多为 $1+6\log n$。接着，对限制在 G_n 上的有界差分不等式应用条件版本（参见[12]），对于常数 $C, c > 0$,

$$\mathbb{P}\left(|\Psi_{\eta_n}(s) - \mathbb{E}[\Psi_{\eta_n}(s) \mid G_n]| \ge \varepsilon\right) \le \frac{C}{n^2} + \exp\left(-\frac{cn\left(\varepsilon - \frac{C\log n}{n^2}\right)_+^2}{(\log n)^2}\right)$$

这在 n 中可求和，因此根据 Borel-Cantelli 引理，$\Psi_{\eta_n}(s) - \mathbb{E}[\Psi_{\eta_n}(s) \mid G_n] \to 0$ 几乎必然成立。此外,

$$|\mathbb{E}\Psi_{\eta_n}(s) - \mathbb{E}[\Psi_{\eta_n}(s) \mid G_n]| = \mathbb{P}(G_n^c)\,|\mathbb{E}[\Psi_{\eta_n}(s) \mid G_n] - \mathbb{E}[\Psi_{\eta_n}(s) \mid G_n^c]| \le \frac{2e^{8R\Lambda_{\max}}}{n^2}.$$

结合式（54），我们得到 $\Psi_{\eta_n}(s) \to \Psi_{\eta}(s)$ 几乎必然成立。因此，在一个概率为 1 的事件上，$\Psi_{\eta_n}(s) \to \Psi_{\eta}(s)$，对每个由有理数 s 定义的此类简单函数 $a_1, \ldots\ldots, a_m \ge 0$ 以及 $R_1, \ldots\ldots, R_m$（端点为有理数）都成立。由此根据命题 B.12 和 B.13 可得 $\eta_n \to \eta$ 几乎必然弱收敛。 □

B.2.2 算法 2 的局部近似

算法 2 涉及对 X 和 Y 的值进行全局排序。本节考虑一种相关程序——算法 3，它利用真实双射 π^*进行后验近似。与算法 2 的主要区别在于，它不再进行排序，而是利用 π^*的流的局部近似，确定要匹配哪些局部窗口。因此，为每个 $\widetilde{P}_i$ 计算的后验概率向量 X_i 只取决于 $\{n(\bar{X}_j - X_i, \bar{Y}_j - X_i)\}_{j=1}^n \in [-D, D]^2$，因此适合使用上一节建立的局部弱收敛进行分析。我们纯粹将此算法用作理论证明的工具。

对于每个索引 $i \in [n]$，定义

$$\mathbf{L}_n(i) := \sum_{j=1}^{n} \mathbf{1}\{X_j \le X_i, Y_{\pi^*(j)} > Y_{\pi^*(i)}\}, \tag{55}$$

$$\mathbf{R}_n(i) := \sum_{j=1}^{n} \mathbf{1}\left\{X_j > X_i, Y_{\pi^*(j)} \le Y_{\pi^*(i)}\right\} \tag{56}$$

因此，$\mathbf{L}_n(i)$ 是在 π^* 下与某个点 $\{Y_j : Y_j > Y_{\pi^*(i)}\}$ 匹配的点 $\{X_j : X_j \le X_i\}$ 的数量，而 $\mathbf{R}_n(i)$ 是与某个点 $\{Y_j : Y_j \le Y_{\pi^*(i)}\}$ 匹配的点 $\{X_j : X_j > X_i\}$ 的数量。固定一个（较大的）局部性参数 $D > 0$，将它们的局部近似定义为

$$\mathbf{L}_n^D(i) := \sum_{j=1}^{n} \mathbf{1}\left\{X_j \in \left[X_i - \frac{D}{n}, X_i\right], Y_{\pi^*(j)} \in \left(Y_{\pi^*(i)}, X_i + \frac{D}{n}\right]\right\}, \tag{57}$$

$$\mathbf{R}_n^D(i) := \sum_{j=1}^{n} \mathbf{1}\left\{X_j \in \left(X_i, X_i + \frac{D}{n}\right], Y_{\pi^*(j)} \in \left[X_i - \frac{D}{n}, Y_{\pi^*(i)}\right]\right\}. \tag{58}$$

于是，我们可以将 π^* 在 i 处的流及其局部近似定义为

$$\mathbf{F}_n(i) \coloneqq \mathbf{L}_n(i) - \mathbf{R}_n(i), \qquad \mathbf{F}_n^D(i) \coloneqq \mathbf{L}_n^D(i) - \mathbf{R}_n^D(i).$$

算法3： 精确匹配中基于流与重排序的后验近似 g

对 $i \leftarrow 1$ **到** n **执行**

计算$\mathbf{F}_n^D(i) = \mathbf{L}_n^D(i) - \mathbf{R}_n^D(i) \in \mathbb{Z}$。

将 X_i 左侧和右侧的 M 个点索引为

$X_{\tilde{s}(-M)} < \ldots < X_{\tilde{s}(0)} \equiv X_i < \ldots < X_{\tilde{s}(M)}$。

将$Y_{\pi^*(i)}$ 左侧和右侧的 $M - \mathbf{F}_n^D(i)$与 $M + \mathbf{F}_n^D(i)$个点索引为

$Y_{\tilde{t}(-M+\mathbf{F}_n^D(i))} < \ldots < Y_{\tilde{t}(0)} \equiv Y_{\pi^*(i)} < \ldots < Y_{\tilde{t}(M+\mathbf{F}_n^D(i))}$。

如果任意 $X_{\tilde{s}(-M)}$, ..., $X_{\tilde{s}(M)}$, $Y_{\tilde{t}(-M+\mathbf{F}_n^D(i))} \cdots$, $Y_{\tilde{t}(M+\mathbf{F}_n^D(i))} \notin [X_i - \frac{D}{n}, X_i + \frac{D}{n}]$ **那么**

Let $\widetilde{P}_i$ 是任意的（例如，$\widetilde{P}_i(j) = 1$ 如果 $j = \pi^*(i)$，否则为 0）。

else

令 $\mathbb{S}$ 为所有双射映射的集合

$\pi : \{-M, \ldots, M\} \to \{-M + \mathbf{F}_n^D(i) \text{ .....}, \; M + \mathbf{F}_n^D(i)\}$。

计算给定于 $\pi \in \mathbb{S}$ 上的局部后验分布

$$P^{(i)}(\pi \mid X, Y) \propto \exp\left(- \sum_{k=-M}^{M} V(n(X_{\tilde{s}(k)} - Y_{\tilde{t}(\pi(k))})) \right).$$

Let

$$= \begin{cases} \widetilde{P}_i(j) \; P^{(i)}(\pi(0) = k) \mid X, Y) & \text{如果对于某个 } k \in \{-M + \mathbf{F}_n^D(i) \text{.....}, \; M + \mathbf{F}_n^D(i)\}, \text{ 有 } j = \tilde{t}(k) \\ 0 & \text{否则} \end{cases}$$

end

返回 $\widetilde{P}_1$, ..., $\widetilde{P}_n$.

引理 B.15。 存在常数 C、$c > 0$，使得对于任意 $D > 0$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\mathbf{L}_n^D(i) \neq \mathbf{L}_n(i)\} \leq Ce^{-cD^{1+\delta}}, \quad \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\mathbf{R}_n^D(i) \neq \mathbf{R}_n(i)\} \leq Ce^{-cD^{1+\delta}} \; w.h.p.$$

证明。我们有

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\mathbf{L}_n^D(i) \neq \mathbf{L}_n(i)\right\} \leq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\exists j : X_j < X_i - \frac{D}{n}, \, Y_{\pi^*(j)} > Y_{\pi^*(i)}\right\} \\ + \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{\exists j : Y_{\pi^*(j)} > X_i + \frac{D}{n}, \, X_j \leq X_i\right\}. \end{aligned} \tag{59}$$

记 $\varepsilon_i = n(Y_{\pi^*(i)} - X_i)$，(59) 中的第一项可进一步上界为

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\varepsilon_i \le -\frac{D}{2} \text{ or } \exists j\colon X_j - X_i < -\frac{D}{n},\, \varepsilon_j \ge -\frac{D}{2} + n(X_i - X_j)\right\} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\varepsilon_i \le -\frac{D}{2}\right\} + \sum_{j=1}^{n}\frac{1}{n}\sum_{i\colon X_i > X_j + \frac{D}{n}}\mathbf{1}\left\{\varepsilon_j \ge -\frac{D}{2} + n(X_i - X_j)\right\} \qquad (60)
\end{aligned}$$

任取 $\kappa > 1$，由尾界（25）和二项分布尾不等式可得，（60）中的第一项有如下界

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\varepsilon_i \le -\frac{D}{2}\right\} \le Ce^{-cD^{1+\delta}} \text{ w.h.p.}$$

对于（60）中的第二项，令 S_j 为 $\{X_1, \ldots, X_n\}$ 中大于且最接近 X_j 的 $\lfloor(\log n)^\kappa\rfloor$ 个值组成的集合。通过与（35）类似的论证，对于某些常数 C、c、$c_0 > 0$，以概率 $1 - Ce^{-c(\log n)^\kappa}$，所有不属于 S_j 的点 $X_i > X_j + \frac{D}{n}$ 都必须满足 $X_i > X_j + c_0(\log n)^\kappa/n$。此外，以至少 $1 - Ce^{-c(\log n)^{\kappa(1+\delta)}}$ 的概率，我们有 $\max_j |\varepsilon_j| < c_0(\log n)^\kappa$。因此，以高概率，（60）中的第二项等于

$$V := \sum_{j=1}^{n}\underbrace{\frac{1}{n}\sum_{i\colon X_i \in S_j}\mathbf{1}\left\{\varepsilon_j \ge -\frac{D}{2} + n(X_i - X_j)\right\}}_{:=V_j}$$

注意，由于对 ε_j的尾界（25），

$$p_j := \mathbb{E}[V_j \mid X] \le \frac{1}{n}\sum_{i\colon X_i > X_j + \frac{D}{n}} Ce^{-c[-D/2 + n(X_i - X_j)]^{1+\delta}}.$$

Thus

$$\begin{aligned}
\sum_{j=1}^{n} p_j &\le \frac{1}{n}\sum_{i,j=1}^{n}\mathbf{1}\left\{X_i > X_j + \frac{D}{n}\right\} Ce^{-c[-D/2 + n(X_i - X_j)]^{1+\delta}} \\
&= \frac{1}{n}\sum_{i<j}\underbrace{\left(\mathbf{1}\left\{\bar{X}_i > \bar{X}_j + \frac{D}{n}\right\} Ce^{-c[-D/2 + n(\bar{X}_i - \bar{X}_j)]^{1+\delta}} + \mathbf{1}\left\{\bar{X}_j > \bar{X}_i + \frac{D}{n}\right\} Ce^{-c[-D/2 + n(\bar{X}_j - \bar{X}_i)]^{1+\delta}}\right)}_{U_{ij}}.
\end{aligned}$$

考虑 Hoeffding 分解 $U_{ij} = \mathbb{E}U_{ij} + U_i + U_j + U_{ij}$，其中 $U_i = \mathbb{E}[U_{ij} \mid X_i] - \mathbb{E}U_{ij}$ 且 $\mathbb{E}[U_{ij} \mid X_i] = \mathbb{E}[U_{ij} \mid X_j] = 0$。不难验证，存在一些常数 C，$c > 0$,

$$\mathbb{E}U_{ij} \le \frac{C}{n}e^{-cD^{1+\delta}}, \quad |U_i| \le \frac{C}{n}e^{-cD^{1+\delta}} \text{ a.s.}, \quad \mathbb{E}|\bar{U}_{ij}| \le \frac{C}{n}e^{-cD^{1+\delta}}, \quad |\bar{U}_{ij}| \le Ce^{-cD^{1+\delta}} \text{ a.s.}.$$

则 $\frac{1}{n}\sum_{i<j}(\mathbb{E}U_{ij} + U_i + U_j) \le Ce^{-cD^{1+\delta}}$，且 [3, 命题 2.3] 关于退化 U 统计量的尾不等式表明

$$\mathbb{P}\left[\frac{1}{n}\sum_{i<j}\bar{U}_{ij} > t\right] \le C\exp\left(-\frac{ct}{e^{-cD^{1+\delta}}n^{-1/2} + t^{1/3}n^{-1/3}}\right),$$

因而也 $\frac{1}{n}\sum_{i<j}\bar{U}_{ij} \le C'e^{-c'D^{1+\delta}}$ 以高概率。因此，对某些常数 C， $c>0$,

$$\sum_{j=1}^{n} p_j \le Ce^{-cD^{1+\delta}} \text{ w.h.p.} \tag{61}$$

在给定 X 的条件下，注意到

$$\mathsf{Var}[V_j \mid X] = \frac{1}{n^2}\sum_{i:X_i\in S_j} \mathsf{Var}\left[\mathbf{1}\left\{\varepsilon_j \ge -\frac{D}{2} + n(X_i - X_j)\right\} \,\middle|\, X\right] \le \frac{p_j}{n},$$

并且 $|V_j| \le (\text{对数}\, n)^{\kappa}/n$ 根据 S_j的定义。然后，在给定 X 且事件（61）发生的条件下，伯恩斯坦不等式也表明

$$\sum_{j=1}^{n}(V_j - p_j) \le Ce^{-cD^{1+\delta}} \text{ w.h.p.}$$

结合这些界，我们得到存在某些常数 C， $c>0$， （60）以高概率至多为 $Ce^{-cD^{1+\delta}}$。（59）的第二项上界为

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\exists j\colon X_j \le X_i,\, \varepsilon_j > n(X_i - X_j) + D\right\},$$

类似的论证表明，该值以高概率至多为 $Ce^{-cD^{1+\delta}}$。因此 $n^{-1}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{L}_n^D(i) \ne \mathbf{L}_n(i)\} \le Ce^{-cD^{1+\delta}}$。最后，应用对称的论证来界定 $\mathbf{1}\{\text{实数集}_n^D \ne \mathbf{R}_n(i)\}$，从而证明该引理。 □

推论 *B.16*。设 $P_i(j)$ 为局部性参数为 $M \ge 1$ 时算法 *2* 的输出，且 $\sim P_i(j)$ 为局部性参数为 $M \ge 1$ 和 $D = D(M)$ 时算法*3* 的输出，其中常数 $D(M) > 0$ 足够大。则存在常数 $C, c > 0$，它们不依赖于 D, M，使得

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widehat{P}_i, \widetilde{P}_i) \le Ce^{-cD^{1+\delta}} + Ce^{-cM} \text{ w.h.p.}$$

证明。 假设 $i = s(a)$ 与 $\pi^*(i) = t(b)$，即 X_i, $Y_{\pi^*(i)}$ 是 a^{th} 和 b^{th} 已排序值。对任意双射 $\pi : [n] \to [n]$，注意

$$\mathbf{F}_n^{\pi}(i) := \sum_{j=1}^{n}\mathbf{1}\{X_j \le X_{(a)}, Y_{\pi(j)} > Y_{(b)}\} - \sum_{j=1}^{n}\mathbf{1}\{X_j > X_{(a)}, Y_{\pi(j)} \le Y_{(b)}\} = a - b,$$

因为这对按已排序顺序匹配各点的双射 $\pi(s(1)) = t(1) \ldots, \pi(s(n)) = t(n)$ 成立，并且容易验证 $\mathbf{F}_n^{\pi}$ 在任意两个指标的对换 $(\pi(i)\ \pi(j)) \leftrightarrow (\pi(j)\ \pi(i))$ 下不变。于是特别地，$\mathbf{F}_n(i) = \mathbf{F}_n^{\pi^*}(i) = a - b$。在算法3中我们有 $\tilde{s}(0) = i = s(a)$ 与 $\tilde{t}(0) = \pi^*(i) = t(b)$，因此在事件 $\mathbf{F}_n^D(i) = \mathbf{F}_n(i)$ 上，这蕴含$\{\tilde{t}(-M + \mathbf{F}_n^D(i)) \ldots \tilde{t}(M + \mathbf{F}_n^D(i))\} = \{t(a - M) \ldots, t(a + M)\}$。因此算法3将 $X_{s(a)}$ 左侧与右侧的M 个点同 $Y_{t(a)}$ 左侧与右侧的 M 个点相匹配，而 $Y_{t(a)}$ 与 $X_{s(a)}$具有相同的排序秩。

因此，对于每个 $i = s(a) \in [n]$，输出 P_i 和 $\sim P_i$ 在该事件上相同

$$\{\mathbf{F}_n(i) = \mathbf{F}_n^D(i)\} \cap \left\{X_{s(a-M)}, X_{s(a+M)} \in \left[X_i - \frac{D}{n}, X_i + \frac{D}{n}\right]\right\}$$
$$\cap \left\{Y_{s(a-M)}, Y_{s(a+M)} \in \left[X_i - \frac{D}{n}, X_i + \frac{D}{n}\right]\right\} =: \mathcal{E}_0(i) \cap \mathcal{E}_1(i) \cap \mathcal{E}_2(i)$$

根据引理 B.15， $n^{-1}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_0(i)^c\}\le Ce^{-cD^{1+\delta}}$ 在（X，Y）上以高概率成立。根据引理 B.5，将M代替 L 应用后，只要 $D>3M/2\Lambda_{\min}$，就有 $n^{-1}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_1(i)^c\}\le Ce^{-cM}$ 以高概率成立。类似地，由引理 B.5 和 B.6，对于 $D>3M/2\Lambda_{\min}+M$，有 $n^{-1}\sum_{i=1}^{n}\mathbf{1}\{\mathcal{E}_2(i)^c\}\le Ce^{-cM}$ 以高概率成立。结合这些陈述可得

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widehat{P}_i,\widetilde{P}_i)\le\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\widehat{P}_i\neq\widetilde{P}_i\}\le Ce^{-cD^{1+\delta}}+Ce^{-cM}.\qquad\square$$

B.2.3 命题2.12的证明

证明。 在$\mathbf{X}_\Lambda$ 和$\mathbf{Y}_\Lambda$ 局部有限这一概率为1的事件上，回顾我们对这些点的索引方式，其中 $\mathsf{X}_0=0$ 和 $\mathsf{Y}_0=\pi^*_\Lambda(0)$。固定某个常数 $D>0$，并定义以下量：

$$\mathsf{L}:=\sum_{i\in\mathbb{Z}:i\le0}\ \sum_{j\in\mathbb{Z}:j>0}\mathbf{1}\{\pi^*_\Lambda(\mathsf{X}_i)=\mathsf{Y}_j\},\quad \mathsf{L}^D:=\sum_{i\in\mathbb{Z}:\mathsf{X}_i\in[-D,0]}\ \sum_{j\in\mathbb{Z}:\mathsf{Y}_j\in(\mathsf{Y}_0,D]}\mathbf{1}\{\pi^*_\Lambda(\mathsf{X}_i)=\mathsf{Y}_j\}.\tag{62}$$

这里 $\mathsf{L}^D<\infty$ 几乎必然成立，因为$\mathbf{X}_\Lambda$ 和$\mathbf{Y}_\Lambda$ 几乎必然局部有限，并且 $\mathsf{L}=\lim_{D\to\infty}\mathsf{L}^D$。

我们先证明 $\mathsf{L}<\infty$ 几乎必然成立。回顾 (55) 和 (57) 中的$\mathbf{L}_n(i)$ 和$\mathbf{L}_n^D(i)$，它们可以用 $X_i=X_{\pi^*(i)}$ 和 $Y_i=Y_i$ 表示为等价形式：

$$\mathbf{L}_n(i)=\sum_{j=1}^{n}\mathbf{1}\{n(\bar{X}_j-\bar{X}_{\pi^*(i)})\le0,n(\bar{Y}_j-\bar{Y}_{\pi^*(i)})>0\},$$
$$\mathbf{L}_n^D(i)=\sum_{j=1}^{n}\mathbf{1}\left\{n(\bar{X}_j-\bar{X}_{\pi^*(i)})\in[-D,0],\,n(\bar{Y}_j-\bar{Y}_{\pi^*(i)})\in(0,D]\right\}.$$

根据点过程的弱收敛（命题 B.14），在给定 $\mathcal{D}_n=\{(X_i,Y_i)\}_{i=1}^n$的条件下，并对索引 $I\in\{1,\dots,n\}$作均匀随机选择，$\mathsf{L}_n^D(I)\to\mathsf{L}^D$ 几乎必然弱收敛。由于 $\mathsf{L}_n^D(I)$ 和 L^D 的分布律取整数值，这意味着 $\mathsf{TV}\left(\mathsf{L}_n^D(I),\mathsf{L}^D\mid\mathcal{D}_n\right)\to0$ 几乎必然成立。使用索引 I的同一实现进行耦合，引理 B.15 意味着

$$\mathsf{TV}\left(\mathbf{L}_n^D(I),\mathbf{L}_n(I)\mid\mathcal{D}_n\right)\le\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{L}_n^D(i)\neq\mathbf{L}_n(i)\}\le Ce^{-cD^{1+\delta}}$$

对所有充分大的 n 几乎必然。因此，对任意 $D_1,D_2>0$，

$$\begin{aligned}&\mathsf{TV}\left(\mathsf{L}^{D_1},\mathsf{L}^{D_2}\right)\\&\le\limsup_{n\to\infty}\mathsf{TV}\left(\mathsf{L}^{D_1},\mathbf{L}_n^{D_1}(I)\mid\mathcal{D}_n\right)+\mathsf{TV}\left(\mathsf{L}^{D_2}(I),\mathbf{L}_n^{D_2}(I)\mid\mathcal{D}_n\right)+\mathsf{TV}\left(\mathbf{L}_n^{D_1}(I),\mathbf{L}_n^{D_2}(I)\mid\mathcal{D}_n\right)\\&\le Ce^{-cD_1^{1+\delta}}+Ce^{-cD_2^{1+\delta}}.\end{aligned}$$

因此，$\{\mathsf{L}^D\}_{D\ge1}$构成了 $\{0,1,2,\dots\}$（除去 ∞）上的一列概率分布，且该序列在全变差意义下是柯西的。由这类分布空间在全变差度量下的完备性，存在支撑于 $\{0,1,2,\dots\}$（除去 ∞）上的极限概率分布 $\mathcal{L}$，使得 $\lim_{D\to\infty}$全变差 $\left(\mathsf{L}^D,\mathcal{L}\right)\to0$。由于 $\mathsf{L}^D\to\mathsf{L}$ 几乎必然，$\mathcal{L}$必为 L 的分布律，从而 $\mathsf{L}<\infty$几乎必然。定义 $\mathsf{R}:=\sum_{i\in\mathbb{Z}:i>0}\sum_{j\in\mathbb{Z}:j\le0}\mathbf{1}\{\pi^*_\Lambda(\mathsf{X}_i)=\mathsf{Y}_j\}$，同样的论证表明 $\mathsf{R}<\infty$几乎必然。

按命题2.12的记号，上述量L、R 分别为$\mathsf{L}=\mathsf{L}_0(\pi^*_\Lambda)$与$\mathsf{R}=\mathsf{R}_0(\pi^*_\Lambda)$。对每个 $a\in\mathbb{Z}$，定义

$$\mathsf{F}_a(\pi^*_\Lambda)=\mathsf{L}_a(\pi^*_\Lambda)-\mathsf{R}_a(\pi^*_\Lambda).$$

尚需证明：若$\mathsf{F}_a(\pi^*_\Lambda)$在 $a-1$ 或 a中任一处取有限值，则它在这两个值处均为有限，且$\mathsf{F}_{a-1}(\pi^*_\Lambda) = \mathsf{F}_a(\pi^*_\Lambda)$。在$\pi^*_\Lambda$ 下，唯一可能对$\mathsf{L}_{a-1}(\pi^*_\Lambda)$与$\mathsf{L}_a(\pi^*_\Lambda)$（或对$\mathsf{R}_{a-1}(\pi^*_\Lambda)$与$\mathsf{R}_a(\pi^*_\Lambda)$）贡献不同的匹配是$(a, \pi^*_\Lambda(a))$与$(\pi^{*-1}_\Lambda(a)\ a)$。只需考虑五种情形：

(i) $\pi^*_\Lambda(a) < a$ and $\pi^{*-1}_\Lambda(a) < a$,

(ii) $\pi^*_\Lambda(a) < a$ and $\pi^{*-1}_\Lambda(a) > a$,

(iii) $\pi^*_\Lambda(a) > a$ and $\pi^{*-1}_\Lambda(a) < a$,

(iv) $\pi^*_\Lambda(a) > a$ and $\pi^{*-1}_\Lambda(a) > a$,

(v) $\pi^*_\Lambda(a) = a$.

在情形 (i) 中，$(a,\ \pi^*_\Lambda(a))$ 对 L_{a-1} 的贡献为 $+1$，但这被 $(\pi^{*-1}_\Lambda(a)\ a)$ 对 R_{a-1} 的 $+1$ 贡献抵消了；这两个匹配对 L_a 和 R_a 均无贡献。因此 $\mathsf{F}_{a-1}(\pi^*_\Lambda) = \mathsf{F}_a(\pi^*_\Lambda)$。其余情形可类似验证，从而完成证明。

□

B.2.4 命题2.14的证明

证明。设 $M' \geq M > 0$为两个（较大的）局部性参数。设 $\widehat{P}_i^M$, $\widehat{P}_i^{M'}$分别表示采用参数 2 和 M 以及 M'运行算法 2.9 可知，几乎必然地，对所有充分大的 n,

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widehat{P}_i^M, \widehat{P}_i^{M'}) \leq \frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widehat{P}_i, \widehat{P}_i^M) + \mathsf{TV}(\widehat{P}_i, \widehat{P}_i^{M'}) \leq o_M(1)$$

其中 $o_M(1)$ 表示当 $M', M \to \infty$ 时趋于零的误差。取 $D(M, M') > 0$充分大，使得 $D(M, M') \to \infty$当 $M \to \infty$时，并令 $\widetilde{P}^M_{,\ i}{}^{D}, \widetilde{P}^{M'}_{,\ i}{}^{D}$ 表示采用局部性参数 M, M'和 $D = D(M, M')$ 运行算法3所得的输出。于是由推论 B.16 可知，同样几乎必然地，对所有充分大的 n,

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widetilde{P}_i^{M,D}, \widetilde{P}_i^{M',D}) \leq o_M(1).$$

根据算法3的定义，全变差$(\widetilde{P}_i^M,\ \widetilde{P}_i^{M'})$是点过程 $\{n(\bar{X}_j - \bar{X}_i,\ \bar{Y}_j - \bar{X}_i)\}_{j=1}^n \cap [-D, D]^2$ 的函数，并且在给定这类点的数量后，它连续依赖于这些点的位置。由引理B.14 的局部弱收敛以及点过程的连续映射定理 [31, pg 152],几乎必然

$$\frac{1}{n}\sum_{i=1}^{n}\mathsf{TV}(\widetilde{P}_i^{M,D}, \widetilde{P}_i^{M',D}) \to \mathbb{E}_\Lambda[\mathsf{TV}(Q^0_{M,D}, Q^0_{M',D})].$$

这里，$Q^0_{M,D}$,$Q^0_{M',D} \in \mathcal{P}(\mathsf{Y}_\Lambda)$ 是将算法3的步骤应用于 $\{n(\mathsf{X}\ \pi^*_\Lambda(\mathsf{X}))\}_{\mathsf{X}\in\mathbf{X}_\Lambda} \cap [-D, D]^2$ 后得到的概率向量。在以下事件下，它们分别与 $Q^0_M = (Q_M(\pi(0) = j))^{j\in\mathbb{Z}}$和$Q^0_{M'} = (Q_{M'}(\pi(0) = j))^{j\in\mathbb{Z}}$ 相同：回顾式 (62) 中为 $\mathsf{L} = \mathsf{L}_0(\pi^*_\Lambda)$ 定义的近似 L^D。令 R^D 表示 $\mathsf{R} = \mathsf{R}_0(\pi^*_\Lambda)$ 的类似近似，并令

$$\mathsf{F}^D = \mathsf{R}^D - \mathsf{L}^D$$

为相应的 $\mathsf{F}(\pi^*_\Lambda)$ 的 ng 近似。于是，在事件上有 $Q^0_{M,\ D} = Q^0_M$ 和 $Q^0_{M',\ D} = Q^0_{M'}$

$$\mathsf{F}^D = \mathsf{F}(\pi^*_\Lambda) \text{ and } \mathsf{X}_{-M'}, \ldots, \mathsf{X}_{M'}, \mathsf{Y}_{-M'+\mathsf{F}^D}, \ldots, \mathsf{Y}_{M'+\mathsf{F}^D} \in [-D, D].$$

由于 $\lim_{D\to\infty}\mathsf{F}^D \to \mathsf{F}(\pi^*_\Lambda)$ 几乎必然成立，我们有 $\lim_{D\to\infty}\mathbb{P}_\Lambda[\mathsf{F}^D \neq \mathsf{F}(\pi^*_\Lambda)] = 0$。由于 $\mathsf{F}(\pi^*_\Lambda)$ 的分布律不依赖于 D，这也意味着 $\lim_{D\to\infty}\mathbb{P}_\Lambda[|\mathsf{F}^D| \geq D/2] = 0$。接着，选取足够大的 $D = D(M, M')$——例如 $D \geq 3M'$——由标准的泊松尾界可知

$$\lim_{M'\to\infty}\mathbb{P}_\Lambda[\mathsf{X}_{-M'},\ldots,\mathsf{X}_{M'},\mathsf{Y}_{-M'+\mathsf{F}^D},\ldots,\mathsf{Y}_{M'+\mathsf{F}^D} \in [-D,D]] = 1.$$

Thus

$$\mathbb{P}_\Lambda[Q^0_{M,D} \neq Q^0_M \text{ or } Q^0_{M',D} \neq Q^0_{M'}] = o_M(1),$$

由上述结果可知

$$\mathbb{E}_\Lambda[\mathsf{TV}(Q^0_M, Q^0_{M'})] = o_M(1).$$

因此，序列 $\{Q^0_M\}_{M\geq 1}$ 在全变差意义下以及 $\mathbb{P}_\Lambda$-概率意义下是柯西序列；也就是说，对于任意固定的 $\varepsilon > 0$，存在 $M_0 \equiv M_0(\varepsilon) > 0$ 使得

$$\mathbb{P}_\Lambda[\mathsf{TV}(Q^0_M, Q^0_{M'}) \geq \varepsilon] \leq \frac{1}{\varepsilon}\mathbb{E}_\Lambda[\mathsf{TV}(Q^0_M, Q^0_{M'})] \leq \varepsilon \text{ for all } M, M' \geq M_0(\varepsilon). \tag{63}$$

存在　　于是存在一个子序列 M_1，M_2，……，使得 $\mathbb{P}_\Lambda[\mathsf{TV}(Q^0_{M_k}, Q^0_{M_{k+1}}) \geq 2^{-k}] \leq 2^{-k}$，s　　o

$$\lim_{k\to\infty}\sup_{j,j'>k}\mathsf{TV}(Q^0_{M_j}, Q^0_{M'_j}) = 0$$

由 Borel-Cantelli 引理，几乎必然成立。因此，在概率为1的事件上，$\{Q^0_{M_k}\}_{k\geq 1}$ 在全变差意义下是柯西序列，故在全变差意义下收敛到 $\mathsf{Q} \in \mathcal{P}(\mathbf{Y}_\Lambda)$。于是

$$\lim_{M\to\infty}\mathbb{P}_\Lambda[\mathsf{TV}(\mathsf{Q}, Q^0_M) \geq \varepsilon] = \lim_{M\to\infty}\lim_{k\to\infty}\mathbb{P}_\Lambda[\mathsf{TV}(Q^0_{M_k}, Q^0_M) \geq \varepsilon] = 0,$$

因此，Q 是 $\{Q^0_M\}_{M\geq 1}$ 依概率收敛的极限，正如所需。关于唯一性，只需注意，若 Q' 是另一个这样的依概率极限，则对于任意 $\varepsilon > 0$,

$$\mathbb{P}_\Lambda[\mathsf{TV}(\mathsf{Q}, \mathsf{Q}') \geq \varepsilon] \leq \lim_{M\to\infty}\mathbb{P}_\Lambda[\mathsf{TV}(\mathsf{Q}, Q^0_M) \geq \varepsilon/2] + \mathbb{P}_\Lambda[\mathsf{TV}(\mathsf{Q}', Q^0_M) \geq \varepsilon/2] = 0,$$

因此 $\mathbb{P}_\Lambda[\mathsf{Q} = \mathsf{Q}'] = 1$。 □

现在我们完成定理2.11的证明。

定理 *2.11*的证明。 固定任意的局部性参数 $M > 0$ 和 $D > 0$，并令 $\widehat{P}^M_i$, $\widetilde{P}^{M}_{,i}{}^{D}$，以及$Q^0_{M,D}$ 如上所定义。我们有

$$\left|\frac{1}{n}\sum_{i=1}^n f(P_i, \pi^*(i)) - \mathbb{E}_\Lambda[f(\mathsf{Q}, \pi^*_\Lambda(0))]\right| \leq (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}) + (\mathrm{IV}) + (\mathrm{V}),$$

其中

$$\begin{aligned}
(\mathrm{I}) &= \left|\frac{1}{n}\sum_{i=1}^n f(P_i, \pi^*(i)) - \frac{1}{n}\sum_{i=1}^n f(\widehat{P}^M_i, \pi^*(i))\right| \\
(\mathrm{II}) &= \left|\frac{1}{n}\sum_{i=1}^n f(\widehat{P}^M_i, \pi^*(i)) - \frac{1}{n}\sum_{i=1}^n f(\widetilde{P}^{M,D}_i, \pi^*(i))\right| \\
(\mathrm{III}) &= \left|\frac{1}{n}\sum_{i=1}^n f(\widetilde{P}^{M,D}_i, \pi^*(i)) - \mathbb{E}_\Lambda[f(Q^0_{M,D}, \pi^*_\Lambda(0))]\right| \\
(\mathrm{IV}) &= \left|\mathbb{E}_\Lambda[f(Q^0_{M,D}, \pi_{\Lambda^*}(0))] - \mathbb{E}_\Lambda[f(Q^0_M, \pi_{\Lambda^*}(0))]\right|, \\
(\mathrm{V}) &= \left|\mathbb{E}_\Lambda[f(Q^0_M, \pi_{\Lambda^*}(0))] - \mathbb{E}_\Lambda[f(\mathsf{Q}, \pi_{\Lambda^*}(0))]\right|.
\end{aligned}$$

由于 $f(\cdot)$ 是全变差连续的，根据前面的论证，对于足够大的 $M > 0$ 和 $D = D(M) > 0$，几乎必然地，对所有足够大的 n，都有 (I) $+$(II) $\leq o_M(1)$，并且 (IV) $+$(V) $\leq o_M(1)$ 也成立。由于 $f(\tilde{P}^M_{,i}{}^D, \pi^*(i))$ 是点过程 $\{n(\bar{X}_j - \bar{X}_i, \bar{Y}_j - \bar{X}_i)\}_{j=1}^n \cap [-D, D]^2$ 的有界函数，且随这些点的位置连续变化，根据引理 B.14 的局部弱收敛和连续映射定理，我们有 $\lim_{n\to\infty}$(III) $= 0$ 几乎必然成立。然后取 $M \to \infty$ 即得证。 □

C 部分匹配的证明

现在我们证明关于部分匹配模型的定理2.4 和2.7 。相关分析与精确匹配情形类似，且更为简单。在本节的引理和分析中，我们始终假设假设条件 2.3 成立，并采用部分匹配模型（6）的设定。

C.1 定理 2.4 的证明

C.1.1 平均势

引理 C.1。 存在一个常数 $C > 0$，使得

$$\frac{1}{n} E\left[\sum_{i\in[N_X]:\pi(i)\neq\varnothing} V\left(n\left(X_i - Y_{\pi(i)}\right)\right) \,\middle|\, X, Y\right] \leq C \ \ w.h.p.$$

证明。 与精确匹配情形类似，我们考虑模型（16），其中用 βV 替代 V。该模型中后验分布的哈密顿量为

$$H_n(\pi \mid X, Y; \beta) = \sum_{i\in[N_X]:\pi(i)\neq\varnothing} \beta V(n(X_i - Y_{\pi(i)})) - \sum_{i\in[N_X]:\pi(i)=\varnothing} U_n(X_i) - \sum_{j\in[N_Y]:\pi^{-1}(j)=\varnothing} U_n(Y_j).$$

将相应的自由能记为

$$\mathcal{F}_n(\beta) := \frac{1}{n}\log \mathcal{Z}(\beta), \qquad \mathcal{Z}(\beta) := \sum_{\pi\in\mathbb{S}_n} \exp\left(-H_n\left(\pi \mid X, Y; \beta\right)\right) \tag{64}$$

其中我们回顾一下， $\mathbb{S}_n$ 是从 $[N_X]$ 到 $[N_Y]$的所有部分双射构成的空间。于是

$$\mathcal{D}_n(\beta) := \frac{1}{n} E_\beta\left[\sum_{i\in[N_X]:\pi(i)\neq\varnothing} V\left(n\left(X_i - Y_{\pi(i)}\right)\right) \,\middle|\, X, Y\right] = -\mathcal{F}_n'(\beta), \tag{65}$$

其中 $\mathcal{F}_n(\beta)$是凸集，因此需要界定的量为

$$\mathcal{D}_n(1) = -\mathcal{F}_n'(1) \leq 2\left(\mathcal{F}_n(1/2) - \mathcal{F}_n(1)\right). \tag{66}$$

与精确匹配中引理 B.1 的证明类似， $\mathcal{F}_n(1)$ 的下界可通过将关于 $\pi \in \mathbb{S}_n$ 的求和下界估计为单个求和项 π^*得到:

$$\begin{aligned}\mathcal{F}_n(1) &\geq -\frac{1}{n} H_n(\pi^* \mid X, Y) \\ &= -\frac{1}{n}\sum_{i\in[N_X]:\pi^*(i)\neq\varnothing} V(n(X_i - Y_{\pi^*(i)})) + \frac{1}{n}\sum_{i\in[N_X]:\pi^*(i)=\varnothing} U_n(X_i) + \frac{1}{n}\sum_{j\in[N_Y]:\pi^{*-1}(j)=\varnothing} U_n(Y_j).\end{aligned}$$

根据命题 2.1的论证可以验证，对所有充分大的 n 以及某些常数C， $c>0$,

$$C \geq \sup_{x\in[0,1]} p_n(x)/\Lambda(x) \geq \inf_{x\in[0,1]} p_n(x)/\Lambda(x) \geq c. \tag{67}$$

于是，对于所有 $x\in[0,1]$, $\overline{U_n}(x) \geq \log c\sqrt{\Lambda(x)} \geq -C$，因此

$$\mathcal{F}_n(1) \geq -\frac{1}{n}\sum_{i\in\mathcal{S}_{XY}} V(n(\bar{X}_i-\bar{Y}_i)) - \frac{C(N_X+N_Y)}{n}$$

记 $\varepsilon_i = n(Y_i - X_i)$，对于任意 $t>0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\sum_{i\in\mathcal{S}_{XY}} V(n(\bar{X}_i-\bar{Y}_i)) \geq t\right) \leq e^{-tn/2}\mathbb{E}\exp\left(\frac{1}{2}\sum_{i\in\mathcal{S}_{XY}} V(\varepsilon_i)\right)\\
&= e^{-tn/2}\mathbb{E}\left[\left(\mathbb{E}\exp\left(V(\varepsilon)/2\right)\right)^{|\mathcal{S}_{XY}|}\right] \leq e^{-tn/2}\mathbb{E}\exp\left(CN\right) \leq \exp\left(-tn/2 + C'n\right).
\end{aligned}$$

取 $t>0$ 足够大即可得到

$$\frac{1}{n}\sum_{i\in\mathcal{S}_{XY}} V(n(\bar{X}_i-\bar{Y}_i)) \leq C \text{ w.h.p.}$$

由标准的泊松尾界， $N_X+N_Y \leq N \leq Cn$ 以高概率成立。因此，我们得到

$$\mathcal{F}_n(1) \geq -C \text{ w.h.p.} \tag{68}$$

为了给 $\mathcal{Z}_n(1/2)$ 一个上界，令

$$\mathcal{Q}_{X,Y} = \{(\mathcal{R}_X,\mathcal{R}_Y)\colon \mathcal{R}_X\subseteq[N_X],\mathcal{R}_Y\subseteq[N_Y], N_X-|\mathcal{R}_X| = N_Y-|\mathcal{R}_Y|\}$$

为所有可能的元组 $(\mathcal{R}_X,\ \mathcal{R}_Y)$ 构成的集合，其中 $\mathcal{R}_X$ 是 $[N_X]$的子集， $\mathcal{R}_Y$ 是 $[N_Y]$的子集，且 $[N_X]\backslash\mathcal{R}_X$ 和 $[N_Y]\backslash\mathcal{R}_Y$ 的基数相等。于是

$$\begin{aligned}
&\mathcal{Z}_n(1/2) = \sum_{\pi\in\mathbb{S}_n}\exp\left(-H_n\left(\pi \mid X,Y;1/2\right)\right)\\
&= \sum_{(\mathcal{R}_X,\mathcal{R}_Y)\in\mathcal{Q}_{X,Y}} \exp\left(\sum_{i\in\mathcal{R}_X} U_n(X_i) + \sum_{j\in\mathcal{R}_Y} U_n(Y_j)\right) \sum_{\pi:[N_X]\backslash\mathcal{R}_X\to[N_Y]\backslash\mathcal{R}_Y} \exp\left(-\frac{1}{2}\sum_{i\in[N_X]\backslash\mathcal{R}_X} V(n(X_i-Y_{\pi(i)}))\right)\\
&\leq \exp\left(C(N_X+N_Y)\right) \sum_{\mathcal{R}_X,\mathcal{R}_Y\in\mathcal{Q}_{X,Y}} \sum_{\pi:[N_X]\backslash\mathcal{R}_X\to[N_Y]\backslash\mathcal{R}_Y} \exp\left(-\frac{1}{2}\sum_{i\in[N_X]\backslash\mathcal{R}_X} V(n(X_i-Y_{\pi(i)}))\right)
\end{aligned}$$

其中最后一个不等式利用（67）对 $U_n(x)\leq C$进行上界估计。再次利用非负矩阵的积和式不超过其行和的乘积这一事实,

$$\begin{aligned}
&\sum_{\pi:[N_X]\backslash\mathcal{R}_X\to[N_Y]\backslash\mathcal{R}_Y} \exp\left(-\frac{1}{2}\sum_{i\in[N_X]\backslash\mathcal{R}_X} V(n(X_i-Y_{\pi(i)}))\right)\\
&\leq \prod_{i\in[N_X]\backslash\mathcal{R}_X}\sum_{j\in[N_Y]\backslash\mathcal{R}_Y} \exp\left(-\frac{1}{2}V(n(X_i-Y_j))\right) \leq \prod_{i\in\pi_X^*([N_X]\backslash\mathcal{R}_X)}\sum_{j=1}^{N}\exp\left(-\frac{1}{2}V(n(\bar{X}_i-\bar{Y}_j))\right).
\end{aligned}$$

Thus

$$\mathcal{F}_n(1/2) \le \frac{CN}{n} + \frac{1}{n}\log \sum_{\mathcal{R}_X,\mathcal{R}_Y \in \mathcal{Q}_{X,Y}} \prod_{i \in \pi_X^*([N_X]\setminus \mathcal{R}_X)} \sum_{j=1}^{N} \exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right). \tag{69}$$

为对右侧的第二项进行上界估计，我们沿用（21）和（23）中的论证：令 $\mathcal{N}_\kappa(\bar{Y}_j)$ 表示 $\lfloor (\log n)^\kappa \rfloor$ 个最近邻的索引集，这些最近邻属于 Y_j 在 $\{Y_1,\dots,Y_N\}$中。于是，以高概率，对每个 $i \in [N]$,

$$\sum_{j=1}^{N} \exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right) \le \exp\left(C + \frac{n}{2}\left|\bar{X}_i - \bar{Y}_i\right|\right) + \sum_{j \in \mathcal{N}_\kappa(\bar{Y}_i)} \exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right).$$

这个上界 至少为 1，因此我们可以进一步纳入指标 $i \notin \pi_X^*([N_X] \setminus \mathcal{R}_X)$，从而得到

$$\mathcal{F}_n(1/2) \le \frac{CN}{n} + \frac{1}{n}\log\left|\mathcal{Q}_{X,Y}\right| + \frac{1}{n}\sum_{i=1}^{N} \log\left[\exp\left(C + \frac{n}{2}\left|\bar{X}_i - \bar{Y}_i\right|\right) + \sum_{j \in \mathcal{N}_\kappa(\bar{Y}_i)} \exp\left(-\frac{1}{2}V(n(\bar{X}_i - \bar{Y}_j))\right)\right].$$

第三项以高概率至多为一个常数，论证与精确匹配中的（24）相同。$\mathcal{Q}_{X,Y}$的基数可由 $2^{N_X}2^{N_Y} \le 2^N$ 从上界控制。因此，再次应用泊松尾界 $N \le Cn$ ，以高概率,

$$\mathcal{F}_n(1/2) \le C \text{ w.h.p.} \tag{70}$$

结合（68）与（70）表明 $\mathcal{D}_n$（1） $\le C$ 以高概率 □

接下来我们证明在局部吉布斯测度下类似的平均势结果。对任意$x \in [0,1]$ 与部分匹配 π，将其在 x 处的边界变量定义为

$$\Gamma_x(\pi) \coloneqq \{(k,\pi(k))\colon k,\pi(k) \ne \varnothing,\ X_k \le x < Y_{\pi(k)} \text{ or } X_k > x \ge Y_{\pi(k)}\}. \tag{71}$$

引理 C.2. 固定任意 K、$L > 0$。存在不依赖于 K、L 的常数 $C > 0$ ，使得

$$\frac{1}{n}\sum_{l=1}^{n} E\left[\frac{1}{KL}\sum_{i \in [N_X]\colon \pi(i) \ne \varnothing,\ X_i \in (\frac{l-KL}{n}, \frac{l+KL}{n}]} V\left(n\left(X_i - Y_{\pi(i)}\right)\right) \,\middle|\, X, Y, \Gamma_{\frac{l \pm KL}{n}}(\pi) = \varnothing\right] \le C \ \ w.h.p.$$

证明。 固定 $l \in [n]$。令 $\mathbb{S}^{(l)}$ 为从 $S_X^l = \{i\colon X_i \in (\frac{l-KL}{n}, \frac{l+KL}{n}]\}$到 $S_Y^l = \{j\colon Y_j \in (\frac{l-KL}{n}, \frac{l+KL}{n}]\}$的所有部分双射 π 组成的集合。在 $\pi \in \mathbb{S}^{(l)}$ 上定义局部吉布斯测度及相应的自由能

$$P_\beta^l(\pi \mid X, Y) = \frac{1}{\mathcal{Z}_l(\beta)} \exp\left(-H_n(\pi \mid S_X^l, S_Y^l; \beta)\right), \quad \mathcal{F}_l(\beta) = \frac{1}{KL}\log \mathcal{Z}_l(\beta) \tag{72}$$

其中

$$H_n(\pi \mid S_X^l, S_Y^l; \beta) = \beta \sum_{i \in S_X^l\colon \pi(i) \in S_Y^l} V(n(X_i - Y_{\pi(i)})) - \sum_{i \in S_X^l\colon \pi(i) = \varnothing} U_n(X_i) - \sum_{j \in S_Y^l\colon \pi^{-1}(j) = \varnothing} U_n(Y_j).$$

那么需要界定的量为 $\frac{1}{n}\sum_{l=1}^{n}\mathcal{D}_n^l$，其中

$$\mathcal{D}_n^l := E_{\beta=1}^{l}\left[\frac{1}{KL}\sum_{i\in S_X^l:\pi(i)\in S_Y^l} V(n(X_i - Y_{\pi(i)}))\middle| X, Y\right] = -\mathcal{F}_l'(1) \le 2\left(\mathcal{F}_l(1/2) - \mathcal{F}_l(1)\right).$$

我们将 $\frac{1}{n}\sum_{i=1}^{n}\mathcal{F}_l(1/2)$ 上界估计如下：令

$$\mathcal{Q}^l = \left\{(\mathcal{R}_X, \mathcal{R}_Y)\colon \mathcal{R}_X \subseteq S_X^l, \mathcal{R}_Y \subseteq S_Y^l, \left|S_X^l \backslash \mathcal{R}_X\right| = \left|S_Y^l \backslash \mathcal{R}_Y\right|\right\},$$

其基数为 $|\mathcal{Q}^l| \le 2^{|S_X^l|}2^{|S_Y^l|}$。接着沿用前面对全局吉布斯测度给出的自由能上界，可得

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n}\mathcal{F}_l(1/2) &\le \frac{1}{n}\sum_{l=1}^{n}\frac{C(|S_X^l| + |S_Y^l|)}{KL} \\
&\quad + \frac{1}{n}\sum_{l=1}^{n}\frac{1}{KL}\sum_{i\in\pi_X^*(S_X^l)}\log\left[\exp\left(C + \frac{n\left|\bar{X}_i - \bar{Y}_i\right|}{2}\right) + \sum_{j\in\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V\left(n\left|\bar{X}_i - \bar{Y}_j\right|\right)\right)\right] \\
&\le \frac{CN}{n} + \frac{C}{n}\sum_{i=1}^{N}\log\left[\exp\left(C + \frac{n\left|\bar{X}_i - \bar{Y}_i\right|}{2}\right) + \sum_{j\in\mathcal{N}_\kappa(\bar{Y}_i)}\exp\left(-\frac{1}{2}V\left(n\left|\bar{X}_i - \bar{Y}_j\right|\right)\right)\right] \\
&\le C' \text{ w.h.p.}
\end{aligned}$$

我们可以通过将对 $\pi \in \mathbb{S}^{(l)}$ 的求和下界估计为限制在 $(S_X^l,\ S_Y^l)$ 上的单个求和项 π^*，从而得到 $\frac{1}{n}\sum_{l=1}^{n}\mathcal{F}_l(1)$ 的下界：

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{n}\mathcal{F}_l(1) \ge \frac{1}{n}\sum_{l=1}^{n}\frac{1}{KL}\sum_{i\in S_X^l:\pi^*(i)\in S_Y^l} &-V\left(n(X_i - Y_{\pi^*(i)})\right) \\
&+ \sum_{i\in N_X^l:\pi^*(i)=\varnothing \text{ or } \pi^*(i)\notin S_Y^l} U_n(X_i) + \sum_{j\in S_Y^l:\pi^{*-1}(j)=\varnothing \text{ or } \pi^{*-1}(j)\notin S_X^l} U_n(Y_j).
\end{aligned}$$

回顾下界 $U_n(x) \ge -C$，并对 $\varepsilon_i = n(Y_i - X_i)$ 应用与前面对全局吉布斯测度的论证中相同的集中性，我们得到

$$\frac{1}{n}\sum_{l=1}^{n}\mathcal{F}_l(1) \ge -C \text{ w.h.p.},$$

结合这些上界和下界即可得到该引理。 □

C.1.2 点过程的正则性与吉布斯测度的局部性

固定局部性参数 $K, L > 0$，并回顾上界 $\Lambda(x) \le \Lambda_{\max}$，对每个 $x \in [0,1]$，我们定义依赖于(X, Y)的事件

$$\mathcal{A}_x = \left\{\sum_{i\in[N_X]}\mathbf{1}\left\{|X_i - x| \le \frac{L}{n}\right\} \le \frac{9L\Lambda_{\max}}{(1-p)^2},\ \sum_{j\in[N_Y]}\mathbf{1}\left\{|Y_j - x| \le \frac{L}{n}\right\} \le \frac{9L\Lambda_{\max}}{(1-p)^2}\right\}$$

在(X,Y)条件下，我们定义依赖于 π的事件

$$\mathcal{C}_x \equiv \mathcal{C}_x(\pi) = \{|X_k - Y_m| \le L/n \text{ for all } (k,m) \in \Gamma_x(\pi)\},$$
$$\mathcal{L}_x \equiv \mathcal{L}_x(\pi) = \mathcal{A}_x \cap \mathcal{C}_x(\pi).$$

对于每个 $i \in [N_X]$ 和 $k \in \mathbb{Z}$，记

$$X_i^k = \frac{\lfloor nX_i \rfloor + k}{n}.$$

注意，$P_i(j)$ 在算法 1 中计算得到的结果恰好是 $P(\pi(i) = j \mid X, Y, \Gamma_{X_i^{\pm KL}} = \varnothing)$，对每个$j \in [N_Y] \cup \{\varnothing\}$。定义

$$\mathcal{G}_i(\pi) = \left\{ \sum_{k=0}^{KL-1} \mathbf{1}\{\mathcal{L}_{X_i^{-k}}(\pi)\} \ge \frac{2KL}{3} \right\} \cap \left\{ \sum_{k=1}^{KL} \mathbf{1}\{\mathcal{L}_{X_i^k}(\pi)\} \ge \frac{2KL}{3} \right\}.$$

本节证明以下引理。

引理 C.3. 固定局部性参数 $K, L > 0$。存在一个不依赖于 K, L 的常数 $C > 0$，使得关于 (X, Y)，以下结论以高概率成立：

$$\frac{1}{N_X} \sum_{i=1}^{N_X} P\left(\mathcal{G}_i^c(\pi) \mid X, Y\right) \le CL^{-\delta}, \qquad \frac{1}{N_X} \sum_{i=1}^{N_X} P\left(\mathcal{G}_i^c(\pi) \,\middle|\, X, Y, \Gamma_{X_i^{\pm KL}}(\pi) = \varnothing\right) \le CL^{-\delta}. \quad (73)$$

我们首先证明，对于大多数位置 X_i^k， $\mathcal{A}_x$ 以高概率成立。

引理 C.4. 存在不依赖于 K, L 的常数 $C, c > 0$，

$$\frac{1}{N_X} \sum_{i=1}^{N_X} \frac{1}{KL} \sum_{k=-KL}^{KL} \mathbf{1}\{\mathcal{A}_{X_i^k}^c\} \le Ce^{-cL}.$$

证明。 任取 $\kappa > 1$，并令 $\mathcal{N}_i^k$为 $\{X_1, \ldots, X_{N_X}\}$中距离 X_i^k最近的 $\lfloor (\log n)^\kappa \rfloor$ 个点的指标。根据二项尾界，以高概率，所有点 $X_j \in X_i^k \pm \frac{L}{n}$ 都属于 $\mathcal{N}_i^k$。在此事件下，

$$\begin{aligned} &\frac{1}{N_X} \sum_{i=1}^{N_X} \frac{1}{KL} \sum_{k=-KL}^{KL} \mathbf{1}\{\mathcal{A}_{X_i^k}^c\} \\ &= \frac{1}{N_X} \sum_{i=1}^{N_X} \frac{1}{KL} \sum_{k=-KL}^{KL} \mathbf{1}\left\{ \left| \left\{ j \in \mathcal{N}_i^k : X_j \in X_i^k \pm \frac{L}{n} \right\} \right| \ge \frac{9\Lambda_{\max} L}{(1-p)^2} \right\} \end{aligned} \quad (74)$$

在给定计数 N_X 的条件下，(74) 对独立同分布变量X_1、$\cdots$、X_{N_X}满足有界差分性质，因为改变任意一个 X_i，其值的变化至多为 $C(\log n^\kappa)$。对 $B \sim \text{Binomial}(n, p)$应用 Chernoff 界和矩母函数不等式 $\mathbb{E}[e^B] \le e^{4np}$，其在给定 N_X^p 条件下的期望也有如下上界

$$\begin{aligned} &\frac{1}{N_X} \sum_{i=1}^{N_X} \frac{1}{KL} \sum_{k=-KL}^{KL} \mathbb{P}\left[\left| \left\{ j \in [N_X] : X_j \in X_i^k \pm \frac{L}{n} \right\} \right| \ge \frac{9\Lambda_{\max} L}{(1-p)^2} \,\middle|\, N_X \right] \\ &\le C \exp\left(-\frac{9\Lambda_{\max} L}{(1-p)^2} + \frac{8\Lambda_{\max} L N_X}{n} \right). \end{aligned}$$

由于 $N_X \le N \le Cn$ 根据泊松尾界，该值以高概率至多为 Ce^{-cL}，因此由有界差分不等式可得该引理。 □

接下来，我们利用此前引理 C.1 和 C.2中得到的平均势结果，证明大多数位置满足 $\mathcal{C}_x(\pi)$。

引理 C.5。设

$$\mathscr{S}(\pi)=\left\{x\in[0,1]\colon \sum_{k=0}^{KL-1}\mathbf{1}\{\mathcal{C}^c_{\frac{\lfloor nx\rfloor-k}{n}}(\pi)\}\geq\frac{KL}{6}\ \ or\ \sum_{k=1}^{KL}\mathbf{1}\{\mathcal{C}^c_{\frac{\lfloor nx\rfloor+k}{n}}(\pi)\}\geq\frac{KL}{6}\right\}$$

存在一个不依赖于 K,L 的常数 $C>0$，使得以高概率

$$\frac{1}{N_X}\sum_{i=1}^{N_X}P\left(X_i\in\mathscr{S}(\pi)\mid X,Y\right)\leq CL^{-\delta/2},\quad \frac{1}{N_X}\sum_{i=1}^{N_X}P\left(X_i\in\mathscr{S}(\pi)\mid X,Y,\Gamma_{X_i^{\pm KL}}(\pi)=\varnothing\right)\leq CL^{-\delta/2}.$$

证明。 对于全局吉布斯测度，我们有

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}\frac{1}{KL}\sum_{i=-KL}^{KL}\mathbf{1}\left\{\mathcal{C}^c_{\frac{l+i}{n}}(\pi)\right\}\leq\frac{C}{n}\sum_{l=1}^{n}\mathbf{1}\left\{\mathcal{C}^c_{\frac{l}{n}}(\pi)\right\}\\
&\leq\frac{C}{n}\sum_{l=1}^{n}\sum_{k\in[N_X]:\pi(k)\neq\varnothing}\mathbf{1}\left\{X_k\leq\tfrac{l}{n}<Y_{\pi(k)}\text{ or }X_k>\tfrac{l}{n}\geq Y_{\pi(k)}\right\}\mathbf{1}\left\{\left|Y_{\pi(k)}-X_k\right|\geq L/n\right\}\\
&\leq\frac{C}{n}\sum_{k\in[N_X]:\pi(k)\neq\varnothing}n\left|Y_{\pi(k)}-X_k\right|\mathbf{1}\left\{n\left|Y_{\pi(k)}-X_k\right|\geq L\right\}
\end{aligned}$$

应用 (41) 和引理C.1,

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}\frac{1}{KL}\sum_{i=-KL}^{KL}P[\mathcal{C}^c_{\frac{l+i}{n}}(\pi)\mid X,Y]\leq\frac{C}{L^\delta}E\left(\frac{1}{n}\sum_{k\in[N_X]:\pi(k)\neq\varnothing}(n\left|Y_{\pi(k)}-X_k\right|)^{1+\delta}\,\middle|\,X,Y\right)\\
&\leq C'L^{-\delta}\text{ w.h.p.}
\end{aligned}\tag{75}$$

由此及马尔可夫不等式可得

$$\frac{1}{n}\sum_{l=1}^{n}P\left(\tfrac{l}{n}\in\mathscr{S}(\pi)\mid X,Y\right)\leq CL^{-\delta}\text{ w.h.p.}\tag{76}$$

然后令 $N_l=|\{i\colon X_i\in[\frac{l}{n},\frac{l+1}{n})\}|$，并注意到 $\sum_l N_l=N_X$，我们有

$$\frac{1}{N_X}\sum_{i=1}^{N_X}P\left(X_i\in\mathscr{S}(\pi)\mid X,Y\right)\leq\frac{1}{N_X}\sum_{l=1}^{n}\underbrace{P\left(\tfrac{l}{n}\in\mathscr{S}(\pi)\mid X,Y\right)}_{:=p_l}N_l\leq L^{-\delta/2}+\frac{1}{N_X}\sum_{l\in[n]:\,p_l\geq L^{-\delta/2}}N_l.$$

指标 $l\in[n]$ where $p_l\geq L^{-\delta/2}$ 的数量有如下界:

$$\sum_{l=1}^{n}\mathbf{1}\left\{p_l\geq L^{-\delta/2}\right\}\leq L^{\delta/2}\sum_{l=1}^{n}p_l\leq CL^{-\delta/2}n\text{ w.h.p.}$$

此类指标 $l\in[n]$ 的子集至多有 $\binom{n}{CL^{-\delta/2}n}\leq 2^n$ 个。因此，根据二项尾界并对所有此类子集应用并集界，存在一个常数 $C>0$，使得

$$\sup_{S\subseteq[n]:|S|\leq CL^{-\delta/2}n}\sum_{l\in S}N_l\leq C'L^{-\delta/2}N_X\text{ w.h.p.}$$

因此我们得到

$$\frac{1}{N_X}\sum_{l=1}^{N_X} P\left(X_i \in \mathscr{S}(\pi) \mid X, Y\right) \le CL^{-\delta/2} \text{ w.h.p.}$$

对于局部吉布斯测度，对每个固定的 $l \in [n]$，在事件 $\Gamma_{\underline{l \pm KL}}(\pi) = \varnothing$ 上我们类似地有

$$\begin{aligned}
&\sum_{i=-KL}^{KL} \mathbf{1}\left\{\mathcal{C}^c_{\frac{l+i}{n}}(\pi)\right\} \\
&\le \sum_{i=-KL}^{KL} \sum_{k\in[N_X]:\pi(k)\neq\varnothing,\, X_k\in(\frac{l-KL}{n},\frac{l+KL}{n}]} \mathbf{1}\left\{\left|Y_{\pi(k)} - X_k\right| \ge \tfrac{L}{n}\right\} \mathbf{1}\left\{X_k \le \tfrac{l+i}{n} < Y_{\pi(k)} \text{ or } X_k > \tfrac{l+i}{n} \ge Y_{\pi(k)}\right\} \\
&\le \sum_{k\in[N_X]:\pi(k)\neq\varnothing,\, X_k\in(\frac{l-KL}{n},\frac{l+KL}{n}]} n\left|Y_{\pi(k)} - X_k\right| \mathbf{1}\left\{n\left|Y_{\pi(k)} - X_k\right| \ge L\right\}.
\end{aligned}$$

然后应用（41）和引理 C.2，再对 $l \in [n]$取平均，

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{n}\frac{1}{KL}\sum_{i=-KL}^{KL} P[\mathcal{C}^c_{(l+i)/n}(\pi) \mid X, Y, \Gamma_{\underline{l\pm KL}}(\pi) = \varnothing] \\
&\le \frac{C}{nL^\delta}\sum_{l=1}^{n} E\left[\frac{1}{KL}\sum_{k\in[N_X]:\pi(k)\neq\varnothing,\, X_k\in(\frac{l-KL}{n},\frac{l+KL}{n}]} (n\left|Y_{\pi(k)} - X_k\right|)^{1+\delta} \,\middle|\, X, Y, \Gamma_{\underline{l\pm KL}}(\pi) = \varnothing\right] \\
&\le C'L^{-\delta} \text{ w.h.p.}
\end{aligned}$$

然后再次由马尔可夫不等式，我们类似于（76）得到

$$\frac{1}{n}\sum_{l=1}^{n} P(\tfrac{l}{n} \in \mathscr{S}(\pi) \mid X, Y, \Gamma_{\underline{l\pm KL}}(\pi) = \varnothing) \le CL^{-\delta}.$$

证明的其余部分与全局吉布斯测度下的证明完全相同。 □

引理 *C.3* 的证明。我们有

$$\begin{aligned}
\frac{1}{N_X}\sum_{i=1}^{N_X}\mathbf{1}\{\mathcal{G}_i^c(\pi)\} &\le \frac{1}{N_X}\sum_{i=1}^{N_X}\mathbf{1}\left\{\sum_{k=0}^{KL-1}\mathbf{1}\left\{\mathcal{L}^c_{X_i^{-k}}(\pi)\right\} \ge \frac{KL}{3} \text{ or } \sum_{k=1}^{KL}\mathbf{1}\left\{\mathcal{L}^c_{X_i^k}(\pi)\right\} \ge \frac{KL}{3}\right\} \\
&\le \frac{1}{N_X}\sum_{i=1}^{N_X}\left(\mathbf{1}\left\{\sum_{k=0}^{KL-1}\mathbf{1}\left\{\mathcal{A}^c_{X_i^{-k}}\right\} \ge \frac{KL}{6}\right\} + \mathbf{1}\left\{\sum_{k=1}^{KL}\mathbf{1}\left\{\mathcal{A}^c_{X_i^k}\right\} \ge \frac{KL}{6}\right\}\right. \\
&\qquad \left. + \mathbf{1}\left\{\sum_{k=0}^{KL-1}\mathbf{1}\left\{\mathcal{C}^c_{X_i^{-k}}(\pi)\right\} \ge \frac{KL}{6} \text{ or } \sum_{k=1}^{KL}\mathbf{1}\left\{\mathcal{C}^c_{X_i^k}(\pi)\right\} \ge \frac{KL}{6}\right\}\right) \\
&\le \frac{1}{N_X}\sum_{i=1}^{N_X}\frac{6}{KL}\sum_{k=-KL}^{KL}\mathbf{1}\left\{\mathcal{A}^c_{X_i^k}\right\} + \frac{1}{N_X}\sum_{i=1}^{N_X}\mathbf{1}\{X_i \in \mathscr{S}(\pi)\}.
\end{aligned}$$

结论便由引理C.4 和C.5得证。 □

C.1.3 相关性衰减

We 现在通过证明后验分布的相关性衰减来证明定理 2.4 .

定理的证明 2.4。 简记为 $P(\cdot) = P(\cdot \mid X, Y)$。我们采用与定理 2.9 的证明中类似的耦合论证。设 π, π' 为从后验分布$P(\pi \mid X, Y)$ 中独立抽取的两个样本。在 (X, Y) 的条件下，固定 $i \in [N_X]$ 并定义位置列表

$$\begin{aligned}
\mathcal{Q}_i^-(\pi) &= \left\{k \in \{-KL-1, \ldots, 0\} : \mathcal{L}_{X_i^k}(\pi) \text{ holds}\right\}, \\
\mathcal{Q}_i^+(\pi) &= \left\{k \in \{1, \ldots, KL\} : \mathcal{L}_{X_i^k}(\pi) \text{ holds}\right\}, \\
\mathcal{Q}_i(\pi) &= \mathcal{Q}_i^-(\pi) \cup \mathcal{Q}_i^+(\pi),
\end{aligned}$$

并定义类似 地 $\mathcal{Q}_i^-(\pi')\mathcal{Q}_i^+(\pi')$ $\mathcal{Q}_i(\pi')$（对于 π'）。在事件 $\mathcal{G}_i(\pi) \cap \mathcal{G}_i(\pi')$ 上，我们必有

$$|\mathcal{Q}_i^-(\pi) \cap \mathcal{Q}_i^-(\pi')| \geq \frac{KL}{3}, \qquad |\mathcal{Q}_i^+(\pi) \cap \mathcal{Q}_i^+(\pi')| \geq \frac{KL}{3}.$$

对任意 $m \in \{-KL-1, \ldots, KL\}$，令 $\mathcal{F}_m(\pi)$ 表示由 $\mathcal{Q}_i(\pi)$ 以及边界变量列表给出的信息集

$$\{\Gamma_{X_i^l}(\pi) : l \in \mathcal{Q}_i(\pi) \setminus [m-L, m+L]\}.$$

我们来给出下界

$$P(\Gamma_{X_i^m}(\pi) = \varnothing \mid \mathcal{L}_{X_i^m}(\pi), \mathcal{F}_m(\pi)).$$

考虑任意满足 $\mathcal{L}_{X_i^m}(\pi)$ 的 π，以及 $\Gamma_{X_i^m}(\pi) = \Gamma \neq \varnothing$。我们由 π 构造出部分双射 $\tilde{\pi}$，除 Γ 中的匹配外，其余匹配均保持不变，并对每个满足 $\in \Gamma$ 的 (k, m)，令 $\tilde{\pi}(k) = \varnothing$ 且 $\tilde{\pi}^{-1}(m) = \varnothing$。根据构造，$\Gamma_{X_i^m}(\tilde{\pi}) = \varnothing$，因此 $\mathcal{L}_{X_i^m}(\tilde{\pi})$ 成立。此外， π 和$\tilde{\pi}$ 具有相同的长程匹配，即 $|X_j - Y_{\pi(j)}| > L/n$ 当且仅当 $|X_j - Y_{\tilde{\pi}(j)}| > L/n$，因此$\mathcal{F}_m(\pi) = \mathcal{F}_m(\tilde{\pi})$。回顾下界 $V(\varepsilon) \geq -C$，以及 $U_n(x) \geq -C$，

$$\frac{\exp(-H_n(\pi))}{\exp(-H_n(\tilde{\pi}))} = \exp\left(-\sum_{(k,m)\in\Gamma} U_n(X_k) + U_n(Y_m) + V(n\,|X_k - Y_m|)\right) \leq \exp\left(C|\Gamma|\right). \tag{77}$$

对于每个固定的边界条件 $\Gamma \neq \varnothing$，从 π 到 $\tilde{\pi}$ 的这个映射是单射。在事件$\mathcal{A}_{X_i^m} \subset \mathcal{L}_{X_i^m}$($\pi$发生时，$L' = CL$ 在 $\{X_1, \ldots, X_{N_X}, Y_1, \ldots, Y_{N_Y}\}$ 中至多有 $[X_i^{m-L}, X_i^{m+L}]$个点，因此满足 $\Gamma \neq \varnothing$的边界条件 $\mathcal{L}_{X_i^m}$的可能数量至多为 $e^{CL\log L}$。接着，按照定理2.9中的相同论证，我们得到

$$P(\Gamma_{X_i^m}(\pi) = \varnothing \mid \mathcal{L}_{X_i^m}(\pi), \mathcal{F}_m(\pi)) \geq \exp(-CL\log L) =: \iota(L). \tag{78}$$

同样的论证 也适用于在条件中加入 $\Gamma_{X_i^{\pm KL}} = \varnothing$ 后 π 的后验分布，因此

$$P(\Gamma_{X_i^m}(\pi) = \varnothing \mid \mathcal{L}_{X_i^m}(\pi), \mathcal{F}_m(\pi), \Gamma_{X_i^{\pm KL}}(\pi) = \varnothing) \geq \iota(L). \tag{79}$$

设定

$$\begin{aligned}
M_- \equiv M_-(\pi, \pi') &= \min\{m : m \in \mathcal{Q}_i^-(\pi) \cap \mathcal{Q}_i^-(\pi'), \Gamma_{X_i^m}(\pi) = \Gamma_{X_i^m}(\pi') = \varnothing\}, \\
M_+ \equiv M_+(\pi, \pi') &= \max\{m : m \in \mathcal{Q}_i^+(\pi) \cap \mathcal{Q}_i^+(\pi'), \Gamma_{X_i^m}(\pi) = \Gamma_{X_i^m}(\pi') = \varnothing\}
\end{aligned}$$

在约定 $M_- = 0$ 或 $M_+ = 0$ 下，若不存在这样的指标，则我们如同定理2.9 的论证那样得到

$$\begin{aligned}\mathsf{TV}\left(\widehat{P}_i, P_i\right) &\le 2P(M_- = 0 \text{ or } M_+ = 0) + 2P(M_- = 0 \text{ or } M_+ = 0 \mid \Gamma_{X_i^{\pm KL}}(\pi) = \varnothing) \\ &\le C\Big((1-\iota(L)^2)^{K/3} + P(\mathcal{G}_i^c(\pi)) + P(\mathcal{G}_i^c(\pi) \mid \Gamma_{X_i^{\pm KL}}(\pi) = \varnothing)\Big).\end{aligned}$$

于是，对位置 C.3 应用引理并取平均后，定理 $i \in [N_X]$即得。 □

C.2 定理 2.7 的证明

C.2.1 局部弱收敛到极限点过程

令 $\mathcal{D}_N = \{N, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing, \{(\bar{X}_i,\ \bar{Y}_i)\}_{i=1}^N,\ \pi_X^*,\ \pi_Y^*\}$。给定 $\mathcal{D}_N$，令 $\bar{I} = \pi_X^*(I) \sim$ Unif $\{\mathcal{S}_{XY} \cup \mathcal{S}_X\}$为 $\{X_1, \dots, X_N\}$ 中随机中心化点的索引，并考虑 $\mathbb{R}^2$ 上的点过程

$$\eta_N = \eta_N^{XY} + \eta_N^X + \eta_N^Y + \eta_N^\varnothing, \qquad \eta_N^\dagger = \sum_{i \in \mathcal{S}_\dagger} \delta_{n(\bar{X}_i - \bar{X}_{\bar{I}}, \bar{Y}_i - \bar{X}_{\bar{I}})} \text{ for } \dagger \in \{XY, X, Y, \varnothing\}$$

，其随机性由 I 的选择所诱导。我们定义极限点过程如下

$$\eta = \eta^{XY} + \eta^X + \eta^Y + \eta^\varnothing$$

：(i) 抽样 $X \sim \Lambda$。(ii) 抽样 $\varepsilon \sim V(\cdot)$，并以 $\mathbb{P}[\dagger_0 = XY] = p$ 和 $\mathbb{P}[\dagger_0 = X] = 1-p$ 抽样原点标记 $\dagger_0 \in \{XY, X\}$。(iii) 令

$$\mu_X(\mathrm{d}x\,\mathrm{d}y) = \frac{\Lambda(X)}{(1-p)^2} q(y-x)\mathrm{d}x\,\mathrm{d}y, \quad (p_{XY}, p_X, p_Y, p_\varnothing) = \Big(p^2, p(1-p), p(1-p), (1-p)^2\Big),$$

并抽样独立的泊松点过程

$$\left(\eta^{XY}\right)' \sim \mathrm{PPP}(p_{XY}\mu_X), \quad \left(\eta^X\right)' \sim \mathrm{PPP}(p_X\mu_X), \quad \eta^Y \sim \mathrm{PPP}(p_Y\mu_X), \quad \eta^\varnothing \sim \mathrm{PPP}(p_\varnothing\mu_X).$$

(iv) 设

$$\begin{cases} \eta^{XY} = \left(\eta^{XY}\right)' + \delta_{(0,\varepsilon)}, & \eta^X = \left(\eta^X\right)' \quad \text{if } \dagger_0 = XY, \\ \eta^X = \left(\eta^X\right)' + \delta_{(0,\varepsilon)}, & \eta^{XY} = \left(\eta^{XY}\right)' \quad \text{if } \dagger_0 = X. \end{cases}$$

引理 C.6. 在 $\mathcal{D}_N$条件下，关于均匀随机索引 $I \sim$均匀$\{\mathcal{S}_{XY} \cup \mathcal{S}_X\}$，当 $n \to \infty$时几乎必然，

$$(\eta_N^{XY}, \eta_N^X, \eta_N^Y, \eta_N^\varnothing) \xrightarrow{w} (\eta^{XY}, \eta^X, \eta^Y, \eta^\varnothing).$$

证明。 根据命题B.12 和 B.13，只需证明对于任意有理矩形简单函数 s^{XY}，s^X，s^Y，$s^\varnothing \in C_c^+(\mathbb{R}^2)$，当 $n \to \infty$ 时，该结论几乎必然成立。

$$\begin{aligned}\Psi_{\eta_N}(s) &:= \mathbb{E}\left[\exp\left\{\sum_{\dagger \in \{XY,X,Y,\varnothing\}} s^\dagger(x,y)\eta_N^\dagger(\mathrm{d}x\,\mathrm{d}y)\right\} \,\middle|\, \mathcal{D}_N\right] \\ &\to \Psi_\eta(s) := \mathbb{E}\exp\left\{\sum_{\dagger \in \{XY,X,Y,\varnothing\}} s^\dagger(x,y)\eta^\dagger(\mathrm{d}x\,\mathrm{d}y)\right\}\end{aligned}$$

设 $s^\dagger = \sum_{\ell=1}^m a_\ell^\dagger \mathbf{1}_{R_\ell^\dagger}$，不妨假设对每个 $\dagger \in \{XY, X, Y, \varnothing\}$，$m$ 都相同。定义 $\mathcal{N}_\dagger^{(i)}(R_\ell^\dagger) = |\{j \in \mathcal{S}_\dagger \colon j \neq i,\ n(\bar{X}_j - \bar{X}_i,\ \bar{Y}_j - \bar{X}_i) \in R_\ell^\dagger\}|$，并令

$$\mathcal{N}_\dagger^{(i)} = \Big(\mathcal{N}_\dagger^{(i)}(R_1^\dagger), \dots, \mathcal{N}_\dagger^{(i)}(R_m^\dagger)\Big), \qquad a^\dagger = (a_1^\dagger, \dots, a_m^\dagger), \qquad \varepsilon_i = n(\bar{Y}_i - \bar{X}_i).$$

Then

$$\Psi_{\eta_N}(s) = \frac{1}{|\mathcal{S}_{XY}| + |\mathcal{S}_X|} \sum_{i \in \mathcal{S}_{XY} \cup \mathcal{S}_X} \exp\left\{ -s^{XY}(0, \varepsilon_i)\mathbf{1}_{i \in \mathcal{S}_{XY}} - s^X(0, \varepsilon_i)\mathbf{1}_{i \in \mathcal{S}_X} - \sum_{\dagger \in \{XY, X, Y, \varnothing\}} (a^\dagger)^\top \mathcal{N}_\dagger^{(i)} \right\}.$$

在给定 $N, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing$ 以及 X_i 的条件下，对任意固定的 $i \in \mathcal{S}_{XY} \cup \mathcal{S}_X$，注意到 $\{n(X_j - X_i, Y_j - X_i)\}_{j \in \mathcal{S}_\dagger \setminus \{i\}}$ 对每个 $\dagger \in \{XY, X, Y, \varnothing\}$ 都是独立同分布的。于是

$$\begin{aligned}
&\mathbb{E}[\Psi_{\eta_N}(s) \mid N, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing] \\
&= \frac{|\mathcal{S}_{XY}|}{|\mathcal{S}_{XY}| + |\mathcal{S}_X|} \mathbb{E}\left[e^{-s^{XY}(0,\varepsilon_i)} \mathbb{E}\left[e^{-(a^\dagger)^\top \xi^{XY}} \,\middle|\, \bar{X}_i \right]^{|\mathcal{S}_{XY}|-1} \prod_{\dagger \in \{X, Y, \varnothing\}} \mathbb{E}\left[e^{-(a^\dagger)^\top \xi^\dagger} \,\middle|\, \bar{X}_i \right]^{|\mathcal{S}_\dagger|} \right] \\
&\quad + \frac{|\mathcal{S}_X|}{|\mathcal{S}_{XY}| + |\mathcal{S}_X|} \mathbb{E}\left[e^{-s^{X}(0,\varepsilon_i)} \mathbb{E}\left[e^{-(a^\dagger)^\top \xi^{X}} \,\middle|\, \bar{X}_i \right]^{|\mathcal{S}_X|-1} \prod_{\dagger \in \{XY, Y, \varnothing\}} \mathbb{E}\left[e^{-(a^\dagger)^\top \xi^\dagger} \,\middle|\, \bar{X}_i \right]^{|\mathcal{S}_\dagger|} \right],
\end{aligned}$$

其中 $\xi^\dagger$ 是 $\{e_1, \dots, e_m\} \cup \{0\}$ 中的一个向量，其在给定 $\bar{X}_i$ 条件下的分布为

$$\mathbb{P}\left[\xi^\dagger = 0 \,\middle|\, \bar{X}_i\right] = 1 - \sum_{\ell=1}^m \rho_{\ell,n}^\dagger(\bar{X}_i), \qquad \mathbb{P}\left[\xi^\dagger = e_\ell \,\middle|\, \bar{X}_i\right] = \rho_{\ell,n}^\dagger(\bar{X}_i) \quad \text{for } 1 \le \ell \le m,$$

并且 $\rho_{\ell,n}^\dagger(\bar{X}_i)$ 对矩形 $R_\ell^\dagger$ 的定义与式（53）中相同。注意到

$$\frac{1}{n}\Big(|\mathcal{S}_{XY}|, |\mathcal{S}_X|, |\mathcal{S}_Y|, |\mathcal{S}_\varnothing|\Big) \to \frac{1}{(1-p)^2}(p_{XY}, p_X, p_Y, p_\varnothing)$$

并应用与式（54）类似的控制收敛论证，当 $n \to \infty$ 时，我们几乎必然得到

$$\begin{aligned}
&\mathbb{E}[\Psi_{\eta_N}(s) \mid N, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing] \\
&\to \Big(p\, \mathbb{E}_{\varepsilon \sim q}[e^{-s^{XY}(0,\varepsilon)}] + (1-p)\mathbb{E}_{\varepsilon \sim q}[e^{-s^{X}(0,\varepsilon)}] \Big) \\
&\qquad \times \mathbb{E}_{X \sim \Lambda}\left[\exp\left(\frac{\Lambda(X)}{(1-p)^2} \sum_{\dagger \in \{XY, X, Y, \varnothing\}} p_\dagger \sum_{\ell=1}^m (e^{-t_\ell^\dagger} - 1) \iint_{R_\ell^\dagger} q(y-x)\, \mathrm{d}x\, \mathrm{d}y \right) \right]
\end{aligned}$$

右侧恰为 $\Psi_\eta(s)$。与精确匹配的引理 B.14中的集中性论证类似，可知 $\Psi_{\eta_N}(s) - \mathbb{E}[\Psi_{\eta_N}(s) \mid N, \mathcal{S}_{XY}, \mathcal{S}_X, \mathcal{S}_Y, \mathcal{S}_\varnothing] \to 0$ 几乎必然成立，从而证明了该引理。 □

推论 C.7. 给定 $\mathcal{D}_N$，令 I 为 $[N_X]$中的一个均匀随机索引，并考虑以下对象的联合分布

$$\Big(\{n(X_i - X_I, Y_{\pi^*(i)} - X_I)\}_{i \in \mathrm{dom}(\pi^*)}, \{n(X_i - X_I)\}_{i \in [N_X] \setminus \mathrm{dom}(\pi^*)}, \{n(Y_j - X_I)\}_{j \in [N_Y] \setminus \mathrm{range}(\pi^*)} \Big) \quad (80)$$

作为定义在 $\mathbb{R}^2$ 上的点过程，以及定义在 $\mathbb{R}$上的两个点过程。类似地，将定义 $(\mathbf{X}_{\Lambda,p},\mathbf{Y}_{\Lambda,p},\pi^*_{\Lambda,p})$ *2.5* 中的对象识别为点过程的联合分布

$$\Big(\{(\mathsf{X},\pi^*_{\Lambda,p}(\mathsf{X}))\}_{\mathsf{X}\in\mathrm{dom}(\pi^*_{\Lambda,p})},\mathbf{X}_{\Lambda,p}\setminus\mathrm{dom}(\pi^*_{\Lambda,p}),\mathbf{Y}_{\Lambda,p}\setminus\mathrm{range}(\pi^*_{\Lambda,p})\Big). \tag{81}$$

那么，在给定 $\mathcal{D}_N$ 的条件下，对 $I\in[N_X]$ 的均匀随机选择，几乎必然地，当 $n\to\infty$ 时，

$$\Big(\{n(X_i-X_I,Y_{\pi^*(i)}-X_I)\}_{i\in\mathrm{dom}(\pi^*)},\{n(X_i-X_I)\}_{i\in[N_X]\setminus\mathrm{dom}(\pi^*)},\{n(Y_j-X_I)\}_{j\in[N_Y]\setminus\mathrm{range}(\pi^*)}\Big)$$
$$\xrightarrow{w}(\mathbf{X}_{\Lambda,p},\mathbf{Y}_{\Lambda,p},\pi^*_{\Lambda,p}).$$

证明。 (80) 是由 η_N^{XY}、 X 在 η_N^X 下的边缘分布，以及 Y 在 η_N^Y 下的边缘分布组成的元组。按照引理 B.14 中类似的论证计算拉普拉斯泛函，可以验证 (81) 与由 η^{XY}、 X 在 η^X 下的边缘分布，以及Y 在 η^Y 下的边缘分布组成的元组具有相同的联合分布律。由引理 C.6，结论随即得出。 □

C.2.2 命题2.8的证明

命题 *2.8* 的证明。 设 $M'\geq M>0$是两个局部性参数，并令 $\widehat{P}_i^M$、 $\widehat{P}_i^{M'}$分别表示算法 1 的相应输出。定理2.4 意味着，对于所有充分大的 n，几乎必然地，

$$\frac{1}{N_X}\sum_{i=1}^{N_X}\mathsf{TV}(\widehat{P}_i^M,\widehat{P}_i^{M'})\leq\frac{1}{N_X}\sum_{i=1}^{N_X}\mathsf{TV}(P_i,\widehat{P}_i^M)+\mathsf{TV}(P_i,\widehat{P}_i^{M'})\leq o_M(1).$$

对任意 $i\in[N_X]$，根据算法 1，全变差($\widehat{P}_i^M$， $\widehat{P}_i^{M'}$) 是点过程 (80) 的函数，并且在这些点的个数的条件下，连续依赖于这些点的位置。然后由推论 C.7 以及连续映射定理，几乎必然

$$\mathbb{E}_{\Lambda,p}[\mathsf{TV}(Q_M^0,Q_{M'}^0)]=\lim_{n\to\infty}\frac{1}{N_X}\sum_{i=1}^{N_X}\mathsf{TV}(\widehat{P}_i^M,\widehat{P}_i^{M'})=o_M(1).$$

证明的其余部分与命题2.14的证明相同。 □

定理 *2.7* 的证明。固定任意局部性参数 $M>0$，并令 $\widehat{P}_i^M$为算法 1应用于 M 的输出。由三角不等式，我们有

$$\left|\frac{1}{N_X}\sum_{i=1}^{N_X}f(P_i,\pi^*(i))-\mathbb{E}_{\Lambda,p}[f(\mathsf{Q},\pi^*_{\Lambda,p}(0))]\right|\leq(\mathrm{I})+(\mathrm{II})+(\mathrm{III})$$

其中

$$(\mathrm{I})=\left|\frac{1}{N_X}\sum_{i=1}^{N_X}f(P_i,\pi^*(i))-\frac{1}{N_X}\sum_{i=1}^{N_X}f(\widehat{P}_i^M,\pi^*(i))\right|,$$
$$(\mathrm{II})=\left|\frac{1}{N_X}\sum_{i=1}^{N_X}f(\widehat{P}_i^M,\pi^*(i))-\mathbb{E}_{\Lambda,p}[f(Q_M^0,\pi^*_{\Lambda,p}(0))]\right|,$$
$$(\mathrm{III})=\left|\mathbb{E}_{\Lambda,p}[f(Q_M^0,\pi^*_{\Lambda,p}(0))]-\mathbb{E}_{\Lambda,p}[f(\mathsf{Q},\pi^*_{\Lambda,p}(0))]\right|.$$

由于 $f(\cdot)$ 是 全变差连续的，(I) $\leq o_M(1)$ 对所有充分大的 n几乎必然成立，由定理 2.4，且 (III) $\leq o_M(1)$ 由命题 2.8。由推论 C.7的局部弱收敛及连续映射定理，(II) $\to 0$ 当 $n\to\infty$时几乎必然成立，从而证明了该定理。 □

致谢

本研究　　部分由 NSF DMS-2142476 和斯隆研究奖学金资助　　.

参考文献

[1] David Aldous 和 J Michael Steele。《客观方法：概率组合优化与局部弱收敛》。载于离散结构上的概率，第 1–72 页。施普林格，2004年。[2] Yeganeh Alimohammadi 和 Kiana Asgari。《带有学习得到的距离度量的 Mallows模型：抽样与最大似然估计》。*arXiv* 预印本 *arXiv:2507.08108*，2025年。[3] Miguel A Arcones 和 Evarist Gin´e。u过程的极限定理。《概率年刊》，第 1494–1542 页，1993年。[4] In´es Armend´ariz、Pablo A Ferrari、Pablo Groisman 和 Florencia Leonardi。$\mathbb{Z}_d$上的置换的有限循环吉布斯测度。《统计物理学杂志》，158(6):1213–1233，2015年。[5] Mohsen Bayati、Margot Gerritsen、David F Gleich、Amin Saberi 和 Ying Wang。大型稀疏网络对齐问题的算法。载于*2009*年*IEEE*第九届国际数据挖掘会议，第 705–710 页。IEEE，2009年。[6]Volker Betz。正则格上的随机排列。《统计物理学杂志》，155(6):1222–1248，2014年。[7] Volker Betz、Daniel Ueltschi 和 Yvan Velenik。带循环权重的随机排列。《应用概率年刊》，21(1):312–331，2011年。[8]Marek Biskup 和 Thomas Richthammer。一维离散点集上的置换吉布斯测度。《应用概率年刊》，25(2):898–929，2015年。[9] Jacopo Borga、Sayan Das、Sumit Mukherjee 和 Peter Winkler。随机排列的大偏差原理。《国际数学研究通知》，2024(3):2138–2191，2024年。[10] Shuxiao Chen、Sizun Jiang、Zongming Ma、Garry P Nolan 和 Bokai Zhu。具有低秩信号的数据集单向匹配。*arXiv* 预印本 *arXiv:2204.13858*，2022年。[11] Michael Chertkov、Lukas Kroc、F Krzakala、M Vergassola 和 L Zdeborov´a。通过在图像间传递消息推断粒子追踪实验。《美国国家科学院院刊》，107(17):7663–7668，2010年。[12] Richard Combes。McDiarmid不等式的推广。载于*2024*年*IEEE*国际信息论研讨会（*ISIT*），第 79–84 页，2024年。[13] Osman Emre Dai、Daniel Cullina 和 Negar Kiyavash。相关高斯数据库近乎精确对齐的可实现性。载于*2020*年*IEEE*国际信息论研讨会（*ISIT*），第 1230–1235 页。IEEE，2020年。[14] Osman Emre Dai、Daniel Cullina 和 Negar Kiyavash。高斯数据库对齐与高斯植入匹配。*arXiv* 预印本 *arXiv:2307.02459*，2023年。

[15] Daryl J Daley 和 David Vere-Jones。《点过程理论导论：第一卷：基础理论与方法》。施普林格，2003年。[16] Daryl J Daley 和 David Vere-Jones。《点过程理论导论：第二卷：一般理论与结构》。施普林格，2008年。[17] Pinar Demetci、Rebecca Santorella、Bj¨orn Sandstede、William Stafford Noble 和 Ritambhara Singh。Scot：采用最优传输的单细胞多组学对齐。《计算生物学杂志》，29(1):3–18，2022年。[18] Jian Ding、Yihong Wu、Jiaming Xu 和 Dana Yang。植入匹配问题：尖锐阈值与无限阶相变。《概率论及相关领域》，187(1- 2):1–71，2023年。[19] Richard Phillips Feynman。氦中 λ 转变的原子理论。《物理评论》，91(6):1291，1953年。[20] Yan V Fyodorov 和 Stephen Muirhead。空间随机置换模型的带状结构。《概率论及相关领域》，179(3):543–587，2021年。[21] Laleh Haghverdi、Aaron TL Lun、Michael D Morgan 和 John C Marioni。通过匹配互为最近邻校正单细胞 RNA 测序数据中的批次效应。《自然·生物技术》，36(5):421–427，2018年。[22] Ilya Korsunsky、Nghia Millard、Jean Fan、Kamil Slowikowski、Fan Zhang、Kevin Wei、Yuriy Baglaenko、Michael Brenner、Po-ru Loh 和 Soumya Raychaudhuri。借助 Harmony 快速、灵敏且准确地整合单细胞数据。《自然·方法》，16(12):1289–1296，2019年。[23] Dmitriy Kunisky 和 Jonathan Niles-Weed。几何植入匹配的强恢复。载于*2022*年*ACM-SIAM*离散算法年度研讨会论文集（*SODA*），第 834–876 页。SIAM，2022年。[24] G¨unter Last 和 Mathew Penrose。《泊松过程讲义》，数理统计学会教材第 7 卷。剑桥大学出版社，剑桥，2018年。[25] Jiayi Ma、Xingyu Jiang、Aoxiang Fan、Junjun Jiang 和 Junchi Yan。从手工特征到深度特征的图像匹配：综述。《国际计算机视觉杂志》，129(1):23–79，2021年。[26] Cheng Mao 和 Yihong Wu。置换混合的学习：成对比较群与组合矩方法。《统计年鉴》，50(4):2231–2255，2022年。[27] Erik Meijering、Oleh Dzyubachyk 和 Ihor Smal。细胞与粒子追踪方法。《酶学方法》，504:183–200，2012年。[28] Mehrdad Moharrami、Cristopher Moore 和 Jiaming Xu。植入匹配问题：相变与精确结果。《应用概率年刊》，31(6):2663–2720，2021年。[29] Sumit Mukherjee。置换上指数族的估计。《统计年鉴》，44(2):853–875，2016年。

[30] Ashwin Pananjady、Martin J Wainwright 和 Thomas A Courtade。打乱数据的线性回归：置换恢复的统计与计算极限。*IEEE* 信息论汇刊，64(5):3286–3300，2017年。[31] Sidney I Resnick。极值、正则变差与点过程，第4卷。施普林格科学与商业媒体，2008年。[32] Adrian Sayers、Yoav Ben-Shlomo、Ashley W Blom 和 Fiona Steele。概率记录链接。国际流行病学杂志，45(3):954–964，2016年。[33] Geoffrey Schiebinger、Jian Shu、Marcin Tabaka、Brian Cleary、Vidya Subramanian、Aryeh Solomon、Joshua Gould、Siyan Liu、Stacie Lin、Peter Berube 等。通过最优传输分析单细胞基因表达，识别重编程过程中的发育轨迹。细胞，176(4):928–943，2019年。[34] Maciej Skorski。Beta 分布的 Bernstein型界。现代随机过程：理论与应用，10(2):211–228，2023年。[35]Tim Stuart、Andrew Butler、Paul Hoffman、Christoph Hafemeister、Efthymia Papalexi、William M Mauck、Yuhan Hao、Marlon Stoeckius、Peter Smibert 和 Rahul Satija。单细胞数据的全面整合。细胞，177(7):1888–1902，2019年。[36] Andras Suto。玻色气体中的渗流相变。物理学*A*辑：数学与一般物理，26(18):4689，1993年。[37] Hoa Thi Nhu Tran、Kok Siong Ang、Marion Chevrier、Xiaomeng Zhang、Nicole Yee Shin Lee、Michelle Goh 和 Jinmiao Chen。单细胞RNA测序数据批次效应校正方法的基准评估。基因组生物学，21(1):12，2020年。[38] Haoyu Wang、Yihong Wu、Jiaming Xu 和 Israel Yolou。几何模型中的随机图匹配：完全图情形。载于学习理论会议，第3441–3488页。PMLR，2022年。[39] Chenyang Zhong。带有 l^1 和 l^2 距离的 Mallows置换模型 I：命中与运行算法及混合时间。*arXiv*预印本 *arXiv:2112.13456*，2021年。[40] Bokai Zhu、Shuxiao Chen、Yunhao Bai、Han Chen、Guanrui Liao、Nilanjan Mukherjee、Gustavo Vazquez、David R McIlwain、Alexandar Tzankov、Ivan T Lee 等。利用共享和独有特征实现稳健的单细胞匹配与多模态分析。《自然·方法》，20(2):304–315，2023年。