arXiv:2305.01358v1 [cs.DS] 2 May 2023

基于样本的子序列自由性的距离近似

Omer Cohen Sidon* Dana Ron†

摘要

在这项工作中，我们研究基于样本的无分布模型中对子序列自由性距离进行近似的问题。对于给定的子序列（词） $w = w_1 \dots w_k$，若存在索引 $1 \le i_1 < \cdots < i_k \le n$，使得对于每个 $1 \le j \le k$都有 $t_{i_j} = w_j$，则称序列（文本） $T = t_1 \dots t_n$ 包含 w。否则，称 T 是 w-自由的。Ron 和 Rosin（ACM TOCT 2022）证明，在基于样本的无分布模型中，对子序列自由性进行单边错误测试所需且足够的样本数为 $\Theta(k/\epsilon)$。

记 $\Delta(T, w, p)$ 为在分布 p 下，T 到 w-自由性的距离：$[n] \to [0, 1]$。我们$\widehat{\quad}$希望获得一个估计值 Δ，使得以至少$2/3$的概率有 $|\Delta - \Delta(T, w, p)| \le \delta$，其中给定的距离参数为 δ。我们的主要结果是一种样本复杂度为 $\tilde{O}(k^2/\delta^2)$ 的算法。我们首先介绍一种适用于底层分布 p 为均匀分布的算法，然后说明如何对其进行修改，使其适用于任意（未知）分布 p。我们还证明，对 $1/\delta$ 的二次依赖是必要的。

*特拉维夫大学，omercs123@gmail.com

†特拉维夫大学，danaron@tau.ac.il。受以色列科学基金会（资助编号 1146/18）和 Kadar 家族奖资助。

1 引言

距离近似算法的定义见 [28], ; 它们是次线性算法，能够近似（以常数成功概率）对象与满足预先指定性质之间的距离$\mathcal{P}$。距离近似（以及与之密切相关的容错测试概念）是性质测试的扩展 [30, 19], ，其目标是区分满足某项性质的对象$\mathcal{P}$ 与远未满足该性质的对象。[1] 在这项工作中，我们研究子序列自由性这一性质。给定一个子序列（词） $w_1 \dots w_k$，它取自某个字母表 Σ；若一个序列（文本） $T = t_1 \dots t_n$取自 Σ，则称其为 w-自由的，如果不存在索引 $1 \le j_1 < \cdots < j_k \le n$使得 $t_{j_i} = w_i$ 对每个 $i \in [k]$都成立。[2]

在以往大多数关于性质测试和距离近似的研究中，算法可以查询访问对象，而满足所讨论性质 $\mathcal{P}$ 的距离定义为：对象与满足 $\mathcal{P}$ 的对象之间的最小汉明距离，再除以对象的大小。在这项工作中，我们考虑更具挑战性、且有时更合适的基于样本的模型；在该模型中，算法仅获得对象的一个随机样本。具体而言，当对象是一个序列 $T = t_1 \dots t_n$ 时，样本中的每个元素都是一个二元组 (j, t_j)。

我们同时研究两种情形：一种是每个索引 j 被（独立地）选取时所依据的底层分布为 $[n]$ 上的均匀分布；另一种是更一般的情形，即底层分布是某个任意的未知 $p : [n] \to [0,1]$。我们将前者称为基于均匀样本的模型，将后者称为分布无关的基于样本模型。（满足性质的）距离由底层分布决定。具体而言，它是根据 p 对索引 j 赋予的权重总和的最小值，其中这些索引必须经过修改，才能使序列 w 不含 t_j。因此，在基于均匀样本的模型中，距离度量就是汉明距离除以 n。

Ron 和 Rosin [29]研究了分布无关的基于样本模型中子序列自由性这一性质的测试问题。他们证明，在该模型中，对子序列自由性进行单边错误测试的样本复杂度为 $\Theta(k/\epsilon)$（其中 ϵ 是给定的距离参数）。一个自然的问题是，我们能否设计一种具有较小样本复杂度的次线性算法，来实际近似文本 T 到 w-自由性的距离。值得注意的是，一般而言，对某个性质进行容错测试（因而也包括距离近似）可能比测试该性质 [17, 3]困难得多。

1.1 我们的结果

下文中，当我们说从 T中均匀选取一个样本时，意思是对于每个样本点 (j, t_j)， j 均匀且独立地从 $[n]$ 中选取。这一说法可推广到底层分布为任意分布 p 的情形。

我们首先在基于均匀样本的模型中设计一种距离近似算法。令 $\Delta(T, w)$ 表示在均匀分布下，T 距离满足 w-自由的距离（这等于为了获得一个满足 w-自由的文本而必须修改的 T 中符号的比例），并令$\delta \in (0,1)$ 表示提供给该算法的误差参数。

[1]容错测试算法需要区分接近满足某个性质的对象与远离满足该性质的对象。[2]对于整数 x，我们用 $[x]$ 表示整数集合 $\{1, \dots, x\}$

定理 1.1 在均匀分布下，存在一种用于子序列自由性检测的基于样本的距离近似算法。该算法获取大小为 $\Theta\left(\frac{k^2}{\delta^2}\cdot\log\left(\frac{k}{\delta}\right)\right)$ 的样本，并输出一个估计值 $\widehat{\Delta}$，使得 $|\widehat{\Delta}-\Delta(T,w)|\le\delta$ 的概率至少为 $2/3$。[3]

接下来，我们将这一结果扩展到分布无关的基于样本模型。对于分布 $p\colon[n]\to[0,1]$，我们用 $\Delta(T,w,p)$ 表示在分布 p 下，T 与 w-自由性的距离（即根据 p，必须修改 T 中符号的最小权重，才能得到一个 w-自由文本）。

定理 1.2 存在一种针对子序列自由性的基于样本的分布无关距离近似算法。该算法从 T 中抽取大小为 $\Theta\left(\frac{k^2}{\delta^2}\cdot\log\left(\frac{k}{\delta}\right)\right)$ 的样本，这些样本服从某个未知分布 p，并输出一个估计值 Δ，使得 $|\Delta-\Delta(T,w,p)|\le\delta$，其成立概率至少为 $2/3$。

最后，我们探讨上界的紧确性。我们将证明（通过一个相当简单的论证），即使对于均匀分布，关于 $1/\delta$ 的二次依赖也是必要的。具体而言，记 k_d 为 w中不同符号的数量，我们给出一个下界 $\Omega(1/(k_d\delta^2))$，该下界适用于均匀分布（对每个含有 个不同符号、 w 取值为 k_d 、且 n 足够大、δ 足够小的 w 均成立——精确表述见定理4.1）。

1.2 我们的算法概述

我们的出发点是[29, 第 3 节中证明的在均匀分布下到 w-自由性的距离的结构刻画。1].[4] 为表述他们的刻画，我们引入 T 中 w 的副本概念，尤其是角色不相交副本。

$T=t_1\dots t_n$中的 $w=w_1\dots w_k$的一个副本是一个索引序列（$j_1,\dots,j_k$），使得 $1\le j_1<\cdots<j_k\le n$和 $t_{j_1}\dots t_{j_k}=w$。将一个副本表示为一个大小为 k的数组 C，其中 $C[i]=j_i$，这样会更加方便。若对于每个 $i\in[k]$，$\{C_\ell[i]\}$ 中的索引均互不相同（尽管对于 $i\ne i'$，可能存在 $C_\ell[i]=C_{\ell'}[i']$ （以及 $\ell\ne\ell'$）），则称副本集合 $\{C_\ell\}$ 为角色不相交的。在 w 的符号彼此均不相同这一特殊情况下，副本集合是角色不相交的，当且仅当它由互不相交的副本组成。Ron 和 Rosin 证明了 [29, 定理 3.4 + 断言 3.1]，即 $\Delta(T,w)$ 等于 w 在 T 中角色不相交副本的最大数量除以 n。

请注意，对子序列自由性进行单边错误的基于样本的测试时，其样本复杂度分析可转化为界定样本的大小：为了确保当 $\Delta(T,w)>\epsilon$ 时，样本中包含能够证明 T 不是 w-free 的证据，所需的样本大小既充分又必要。这里的证据表现为样本中存在 w 的一个副本，因此测试算法只需检查是否存在这样的副本。另一方面，距离近似问题具有更强的算法性质，因为对于给定的样本，算法必须执行什么并不是由问题本身决定的。

首先聚焦于均匀情形，Ron 和 Rosin 利用他们的刻画（更准确地说，利用如下方向：若 $\Delta(T,w)>\epsilon$，则 T 包含超过 ϵn 个 w的角色不相交副本），证明大小为 $\Theta(k/\epsilon)$ 的样本以至少 $2/3$ 的概率包含至少一个 w 副本。

[3]照例，对于任意 $\eta>0$，我们可以将成功概率提高到 $1-\eta$，代价是样本复杂度增加一个乘法因子 $O(\log(1/\eta))$。
[4]事实上，Ron 和 Rosin 指出：“这一刻画可能有助于证明关于子序列自由性性质测试以及（次线性的）距离近似的进一步结果。”

在这项工作中，我们更进一步，设计了一种算法，能够根据从 T 中均匀选取的样本，实际估算 T 中 w 的角色不相交副本数量（从而估算 $\Delta(T,w)$）。值得注意的是，对于恰好含有相同数量 w 角色不相交副本（因而到成为 w-free 的距离也相同）的文本，样本中出现一个副本的概率可能大不相同。[5]

在下一小节中，我们将讨论上述算法（适用于均匀情形），并在再下一小节中讨论无分布情形。

1.2.1 均匀情形

令 $R(T,w)$ 表示 T 中 w 的角色不相交副本数。简而言之，该算法通过计算 T 的少量前缀中 w 的符号出现次数的估计值，并利用这些估计值推导出 $R(T,w)$ 的估计值来实现。该算法及其分析的更精确描述基于若干组合性断言，我们将在下面介绍并简要讨论这些断言。

令 $R_i^j(T,w)$ 表示 $T, t_1 \dots t_j$ 的长度为 j 的前缀中， $w, w_1 \dots w_i$ 的长度为 i 的前缀的角色不相交副本数；令 $N_i^j(T,w)$ 表示 $t_1 \dots t_j$ 中符号 w_i 的出现次数。在我们的第一个组合性断言中，我们证明，对于任意 $i \in [k]$ 和 $j \in [n]$， $R_i^j(T,w)$ 的值都可以用 $j' \in [j]$ 对应的 $N_i^{j'}(T,w)$ 的值（特别是 $N_i^j(T,w)$）以及 $j' \in [j]$ 对应的 $R_{i-1}^{j'-1}(T,w)$ 的值来表示。换言之，我们建立了一个递归表达式；它意味着，如果我们知道每个 $j' \in [j]$ 对应的 $R_{i-1}^{j'-1}(T,w)$ 和 $N_i^{j'}(T,w)$，就能计算出 $R_i^j(T,w)$（最终计算出 $R(T,w) = R_k^n(T,w)$）。

在我们的第二个组合性断言中，我们证明，如果只需要对 $R(T,w)$ 进行近似，那么只需定义一个（同样以递归方式构造的）度量，该度量依赖于每个 $i \in [k]$ 对应的 $N_i^j(T,w)$ 的取值，但只需考虑 j 的相对少量取值，且这些取值均匀分布。更准确地说，每个这样的 j 都属于集合 $J = \{r \cdot \gamma n\}_{r=1}^{1/\gamma}$，其中 $\gamma = \Theta(\delta/k)$。我们证明，由于对于所选的 γ，每个区间 $[(r-1)\gamma n + 1, r\gamma n]$ 的大小均为 γn，因此可以确保上述度量（仅使用 $j \in J$）对 $R(T,w)$ 的近似误差不超过 $O(\delta n)$。

接着，我们证明，如果将这些 j 的取值（以及每个 $i \in [k]$）对应的每个 $N_i^j(T,w)$ 替换为足够精确的估计值，那么对 $R(T,w)$ 的近似将产生有界误差。最后，我们使用（均匀）抽样来获得这些估计值，样本大小为 $\tilde{O}(k^2/\delta^2)$。

1.2.2 无分布情形

在 [29, 第 4] 节中证明，给定一个词 w、一个文本 T 和一个分布 p，可以~定义一个词 $\widetilde{w}$ 和一个文本 T，使得以下结论成立。首先， $\Delta(T,w,p)$ 与 $\sim\Delta(T,\widetilde{w})$ 密切相关。其次，从 T 中按照 p 选取样本时观察到 w 的一个副本的概率，与从 T中均匀~选取样本时观察到 $\widetilde{w}$ 的一个副本的概率密切相关。

~我们使用上文所述的第一个关系（即 $\Delta(T,w,p)$ 与 $\Delta(T,\widetilde{w})$ 之间的关系）。然而，由于我们关注的是距离近似，而非单边错误测试，上文所述的第二个关系（即观察到 w in T 的一个副本的概率与观察到 $\sim\widetilde{w}$ in T 的一个副本的概率之间的关系）无法满足我们的需求，因此我们需要采取另一条（再次强调，更具算法性）路径，下面很快就会解释。

[5]例如，考虑 $w = 1 \dots k, T_1 = (1 \dots k)^{n/k}$ 和 $T_2 = 1^{n/k} \dots k^{n/k}$。

~理想情况下，我们希望从 T 中均匀抽样，然后使用该样本（以及 $\widetilde{w}$）运行前一小节中讨论的算法。然而，我们只能按照底层分布 p 对 T 进行抽样，无法直接从 T 中抽取均匀的~~样本。此外，由于 T 是基于（未知的） p 定义的，我们不清楚如何确定上述（均匀间隔的）索引子集 J。

为清晰起见，在继续当前论述时，我们作两个假设。第一，分布 p 满足：存在一个值 β，使得对于每个 $j \in [n]$， p_j/β 都是整数（无需知道 β 的值）。第二，在 w 中，不存在两个相同的连续符号。在这些假设下， $\widetilde{T} = t_1^{p_1/\beta} \dots t_n^{p_n/\beta}$、 $\widetilde{w} = w$和~Δ（T、$\widetilde{w}$） $= \Delta$（T、w、p），其中，对于整数 x， t_j^x是由 x 次重复抽样t_j组成的子序列。

我们针对无分布情形的算法（在上述假设下运行）首先抽取一个服从 p 分布的样本，并利用它从 $[n]$中选取一个（相对较小的）索引子集。将这些索引记为 b_0、 b_1、.....、 b_ℓ，其中 $b_0 = 0 < b_1 < \cdots < b_{\ell-1} < b_\ell = n$。我们原本希望确保每个区间 $[b_{u-1}+1, b_u]$ 相对于 p 的权重近似相同（就像均匀情形下考虑由子集 J 定义的区间那样）。更准确地说，我们希望每个区间的权重都相对较小，同时区间总数也不会太多。然而，由于某些单个索引 $j \in [n]$的概率p_j 可能很大，我们也允许存在权重较大的区间，但这些区间仅包含一个索引（且这样的区间数量很少）。

接下来，算法会再抽取一个样本，以便针对每个 $i \in [k]$ 和 $u \in [\ell]$，近似估计 T 的长度为 b_u 的前缀中符号 w_i 出现次数的权重，权重依据 p 计算。请注意， T 的前缀对应于 T 的前缀。此外，此类前缀中符号出现次数依据 p 计算的权重，转换为 T 中相应的~~前缀里符号的出现次数，并按 T 的长度归一化。随后，算法使用这些近似值来获得 $\Delta(T, \widetilde{w})$ 的估计值。

我们能期望很短~~连续~~子序列对在~~1.2.4~~节所述均匀情形算法~中的情况相反。然而，这总是由 T 中的单索引区间导致的（对于 j 使得 p_j 很大的情形）。每个这样的区间都对应于~ 中一个包含相同符号重复抽样的连续子序列，并且我们证明，这些区间不会带来额外误差。 T </style}

1.3 相关结果

正如我们先前提到的，与我们的工作最密切相关的是 Ron 和 Rosin 关于子序列自由性的分布无关基于样本测试的研究 [29]。关于性质测试的其他相关结果（例如，测试序列的其他性质、对其他类型性质进行基于样本的测试，以及分布无关测试（可能包含查询）），请参见 [29], 的引言，特别是第 1.4 节。关于另一类工作，即最长递增子序列的次线性近似，请参见 [26] 及其中的参考文献。这里我们简要讨论距离近似／容错测试的相关结果。

如前所述，距离近似和容错测试最早在 [28],中得到正式定义，并在 [17, 3]中被证明对某些性质而言难度显著更高。此前几乎所有研究结果都基于查询，且距离度量是在均匀分布下定义的。这些结果包括[20, 18, 1, 25, 15, 11, 22, 7, 24, 16, 27]。Kopparty 和 Saraf [23] 给出了针对若干分布族的线性容错测试查询算法结果。Berman、Raskhodnikova

和 Yaroslavtsev [5] 给出了单调性的容错（基于查询） L_p测试算法。Berman、Murzbulatov 和 Raskhodnikova [4] 给出了一种适用于均匀分布的图像性质基于样本的距离近似算法。

Canonne 等人 [12] 研究了布尔函数在各种偏序集上的 k-单调性。有限偏序集定义域 D 上的布尔函数是 k-单调的，当且仅当它在 D中的任意上升链上，取值 0 和 1 交替出现的次数至多为 k 。对于 $D = [n]$这一特殊情形， k- 单调性等价于不含长度为 $k+2$ 的 w，其中 $w_1 \in \{0,1\}$ 且对每个 $i \in [2, k+2]$都有 $w_i = 1 - w_{i-1}$。他们的一项结果给出了函数在均匀分布下满足 k-单调性的距离近似样本复杂度上界 $\widetilde{O}\left(\frac{k}{\delta^3}\right)$： f $[n] \to \{0,1\}$ （因此，当 w 是特定形式的二进制子序列时，也适用于 w-自由性）。该结果可推广到高维中的 k-单调性（代价是关于维度 d呈指数增长）。

Blum 和 Hu [9] 研究了分布无关主动设定下直线上 k-区间（布尔）函数的距离近似。在这种设定中，算法从函数定义域中获取一个无标签样本，并针对样本点的一个子集提出查询。聚焦于样本复杂度，他们证明，对于直线上的任意底层分布 p，大小为 $\widetilde{O}\left(\frac{k}{\delta^2}\right)$ 的样本足以将函数是 k-区间函数的距离近似到加性误差 δ 以内。这意味着，对于前一段所述的不含同一对 w 的特殊情形，存在一个样本复杂度相同的基于样本的分布无关距离近似算法，只需将 $k+2$ 替换为 $k+1$。

Blais、Ferreira Pinto Jr. 和 Harms [8] 引入了 VC 维数的一种变体，并用它证明了多种性质的分布无关测试样本复杂度的下界和上界。特别是，他们的一个结果表明， [9] 的结果中对 k 的线性依赖基本上是最优的。

最后，我们提到，在无分布情形下，我们通过抽样构造“权重几乎相等”的区间所采用的方法，让人联想到在处理非均匀分布时，其他测试场景中使用的技术[6, 21, 10]。

1.4 进一步研究

这项工作留下的主要开放问题，是缩小我们给出的上界与下界之间的差距，尤其是弄清其对 k 的精确依赖关系，或对由 w 确定的其他参数（例如 k_d）的依赖关系。第一作者的硕士论文 [13] 中给出了朝这一方向迈进的一步。

1.5 组织结构

在第 2 节中，我们介绍均匀分布下的距离近似算法。无分布情形下的算法见第 3节。在第 4 节中，我们证明下界。附录中给出了 Chernoff 界，以及若干技术性断言的证明。

2 均匀分布下的距离近似

本节讨论底层分布为均匀分布时的距离近似问题。正如引言中所述，Ron 和 Rosin 证明了 [29, Thm. 3.4] ， $\Delta(T, w)$（在均匀分布下， T 到 w-自由性的距离）等于

的角色不相交副本数量在 T中，除以 $n = |T|$ （其中角色不相交副本的定义见引言——参见第1.2节）。我们可以用 $T[j]$ 表示 j^{th} 的 T 符号（因此 $T[j] = t_j$）。

我们先介绍以下记号。

定义2.1 对于每个 $i \in [k]$ 和 $j \in [n]$，令 $N_i^j(T, w)$表示 T中长度为 j的前缀里符号 w_i的出现次数。令$R_i^j(T, w)$表示 $T[1, j]$中子序列 $w_1 \dots w_i$的角色不相交副本数。当 $i = k$和 $j = n$时，我们用简写 $R(T, w)$表示 $R_k^n(T, w)$（即 T中 w的角色不相交副本总数）。

注意，对于每个 $j \in [n]$，$R_1^j(T, w)$都等于 $N_1^j(T, w)$。

由于如上所述，$\Delta(T, w) = R(T, w)/n$，我们希望估计 $R(T, w)$。更具体地说，给定 $\delta > 0$，我们希望得到一个估计值 $\widehat{R}$，使得：$\left|\widehat{R} - R(T, w)\right| \leq \delta n$。为此，我们首先证明两个组合性断言。第一个断言表明，每个 $R_i^j(T, w)$的值都可以用 $j' \in [j]$ 对应的 $N_i^{j'}(T, w)$的值（尤其是 $N_i^j(T, w)$）以及 $j' \in [j]$对应的 $R_{i-1}^{j'-1}(T, w)$的值来表示。换言之，如果我们知道对每个 $j' \in [j]$而言 $R_{i-1}^{j'-1}(T, w)$和 $N_i^{j'}(T, w)$的值，就可以计算出 $R_i^j(T, w)$。

断言 2.1 对于每个 $i \in \{2, \dots, k\}$ 和 $j \in [n]$，

$$R_i^j(T, w) = N_i^j(T, w) - \max_{j' \in [j]} \left\{ N_i^{j'}(T, w) - R_{i-1}^{j'-1}(T, w) \right\} .$$

显然，$R_i^j(T, w) \leq N_i^j(T, w)$（对于每个 $i \in \{2, \dots, k\}$ 和 $j \in [n]$）成立，因为 $T[1, j]$ 中每个 $w_1 \dots w_i$ 的角色不相交副本都必须以 $T[1, j]$ 中一个不同的 w_i 出现次数结尾。断言 2.1 准确说明了 $R_i^j(T, w)$ 比 $N_i^j(T, w)$ 小多少。粗略地说，表达式 $\max_{j' \in [j]} \left\{ N_i^{j'}(T, w) - R_{i-1}^{j'-1}(T, w) \right\}$ 计入了 $T[1, j]$ 中无法用于构成 $T[1, j]$ 中 $w_1 \dots w_i$ 的角色不相交副本的 w_i 出现次数。

证明：为简化记号，我们针对 $i = k$ and$j = n$ 的情形证明该结论。对于一般的 $i \in \{2, \dots, k\}$ and $j \in [n]$，只需重命名索引，证明过程基本相同。由于在整个证明过程中 $T$and w 均为固定值，我们将使用简写 N_i^j 表示 $N_i^j(T, w)$，并使用简写 R_i^j 表示 $R_i^j(T, w)$。

为便于分析，我们首先描述一个简单的贪心过程，该过程在 T 中构造 $R = R_k^n$ 个彼此角色不相交的 w 副本。该过程的正确性由 [29, 命题 3.5] 以及一个简单的归纳论证得出（详细信息见附录 B ）。对于每个副本 C_m，其中 $m \in [R]$是一个大小为 k 的数组，其值单调递增；对于每个 $i \in [k]$ ，都有 $C_m[i] \in [n]$，且 $T[C_m[i]] = w_i$。此外，对于每个 $i \in [k]$，索引 $C_1[i]$、…、$C_R[i]$互不相同。对于每个 $m = 1$、…、R 以及 $i = 1$、…、k，该过程扫描 T，从 $T[C_m[i-1]+1]$ 开始（其中我们定义 $C_m[0]$ 为 0），到 $T[n]$ 结束，直到找到第一个满足 $T[j] = w_i$ 且 $j \notin \{C_1[i], \dots, C_{m-1}[i]\}$ 的索引 j。随后将 $C_m[i] = j$ 设为该值。对于 $i > 1$，在这种情况下，我们称该过程将 j 匹配到部分副本 $C_m[1]$、…、$C_m[i-1]$。

对于 $i \in [k]$，定义：$G_i = \{j \in [n] : T[j] = w_i\}$。另定义：$G_i^+ = \{j \in G_i : \exists m, C_m[i] = j\}$ 以及 $G_i^- = \{j \in G_i : \nexists m, C_m[i] = j\}$ （回顾一下，$C_m[i]$ 是第 m 个贪心副本中的第 i 个索引）。

[6]确实，如果对于 $i \neq i'$ 有 $w_i = w_{i'}$，那么对于每个 j，$N_i^j(T, w) = N_{i'}^j(T, w)$。

容易验证 $|G_i| = N_i^n$、 $\left|G_i^+\right| = R_i^n$ 和 $|G_i| = \left|G_i^+\right| + \left|G_i^-\right|$。为完成证明，我们将证明 $\left|G_i^-\right| = \max_{j\in[n]} \left\{N_i^j - R_{i-1}^{j-1}\right\}$。

令 j^* 为使 $\left\{N_i^j - R_{i-1}^{j-1}\right\}$ 最大的索引 j。在区间 $[j^*]$ 中， w_i 出现了 $N_i^{j^*}$ 次；而在区间 $[j^*-1]$ 中， $w_1 \dots w_{i-1}$ 只有 $R_{i-1}^{j^*-1}$ 个角色不相交副本。这意味着在区间 $[j^*]$ 中，至少有 w_i 的 $N_i^{j^*} - R_{i-1}^{j^*-1}$ 次出现不可能是任何贪心副本的第 i 个索引，因此我们有

$$\left|G_i^-\right| \geq N_i^{j^*} - R_{i-1}^{j^*-1} = \max_{j\in[n]} \left\{N_i^j - R_{i-1}^{j-1}\right\} . \tag{2.1}$$

另一方面，记 G_i^- 中最大的索引为 j^{**}。由于每个满足 $T[j] = w_i$ 的索引 $j \in [j^{**}]$，要么是某个副本的第 i 个元素，要么不是任何副本的第 i 个元素，$N_i^{j^{**}} = R_i^{j^{**}-1} + \left|G_i^-\right|$。我们声称 $R_i^{j^{**}-1} = R_{i-1}^{j^{**}-1}$。否则， $R_i^{j^{**}-1} < R_{i-1}^{j^{**}-1}$，此时索引 j^{**} 就必须是某个贪心副本的第 i 个元素。因此，

$$\left|G_i^-\right| = N_i^{j^{**}} - R_{i-1}^{j^{**}-1} \leq \max_{j\in[n]} \left\{N_i^j - R_{i-1}^{j-1}\right\} . \tag{2.2}$$

总之，

$$\left|G_i^-\right| = \max_{j\in[n]} \left\{N_i^j - R_{i-1}^{j-1}\right\} , \tag{2.3}$$

该断言得证。 ■

为了陈述下一个组合性断言，我们先引入另一个定义；它将在求得 $R(T,w)$ 的估计值时发挥核心作用。

定义 2.2 对于 $\ell \leq n$，令 $\mathcal{N}$ 为一个由非负数组成的 $k \times \ell$ 矩阵，其中我们用$\mathcal{N}_i^r$表示 $\mathcal{N}[i][r]$。对于每个 $r \in [\ell]$，令 $M_1^r(\mathcal{N}) = \mathcal{N}_1^r$；对于每个 $i \in \{2, \dots, k\}$，令

$$M_i^r(\mathcal{N}) \stackrel{\text{def}}{=} \mathcal{N}_i^r - \max_{r'\leq r} \left\{\mathcal{N}_i^{r'} - M_{i-1}^{r'}(\mathcal{N})\right\} .$$

当 $i = k$且 $r = \ell$时，我们用简写 $M(\mathcal{N})$ 表示 $M_k^\ell(\mathcal{N})$。

在第二个组合性断言中，我们证明，对于适当选取的矩阵 $\mathcal{N}$（其元素是 $\left\{N_i^j(T,w)\right\}_{i\in[k]}^{j\in[n]}$ 中所有取值的一个子集），可以界定 $M(\mathcal{N})$ 与 $R(T,w)$ 之间的差异。随后我们使用抽样来得到 $\mathcal{N}$ 的一个估计版本。

命题 2.2 设 $J = \{j_0, j_1, \dots, j_\ell\}$ 是满足 $j_0 = 0 < j_1 < j_2 < \cdots < j_\ell = n$ 的指标集。设 $\mathcal{N} = \mathcal{N}(J,T,w)$ 为一个矩阵，其元素为 $\mathcal{N}_i^r = N_i^{j_r}(T,w)$，对每个 $i \in [k]$ 和 $r \in [\ell]$ 均成立。于是我们有

$$|M(\mathcal{N}) - R(T,w)| \leq (k-1) \cdot \max_{\tau\in[\ell]} \{j_\tau - j_{\tau-1}\} .$$

证明：回顾一下， $M(\mathcal{N}) = M_k^\ell(\mathcal{N})$ 且 $R(T,w) = R_k^{j_\ell}(T,w)$。我们将证明，对于每个$i \in [k]$ 以及每个 $r \in [\ell]$， $\left|M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)\right| \leq (i-1) \cdot \max_{\tau\in[r]} \{j_\tau - j_{\tau-1}\}$。我们用归纳法对 i 进行证明。

对于 $i=1$ 以及每个 $r\in[\ell]$,

$$\left|M_1^r(\mathcal{N})-R_1^{j_r}(T,w)\right|=\left|N_1^{j_r}(T,w)-N_1^{j_r}(T,w)\right|=0\le(1-1)\cdot\max_{\tau\in[1]}\{j_\tau-j_{\tau-1}\}\ , \tag{2.4}$$

其中，第一个等号由 $\mathcal{N}$ 的设定以及 $M_1^r(\mathcal{N})$ 和 $R_1^{j_r}(T,w)$ 的定义得出。对于归纳步骤，我们假设该结论对 $i-1\ge1$ （以及每个 $r\in[\ell]$）成立，并证明它对 i 成立。我们有：

$$\begin{aligned}
&M_i^r(\mathcal{N})-R_i^{j_r}(T,w)\\
&\quad=\quad N_i^{j_r}(T,w)-\max_{b\in[r]}\left\{N_i^{j_b}(T,w)-M_{i-1}^b(\mathcal{N})\right\}-R_i^{j_r}(T,w) &(2.5)\\
&\quad=\quad \max_{j\in[j_r]}\left\{N_i^{j}(T,w)-R_{i-1}^{j-1}(T,w)\right\}-\max_{b\in[r]}\left\{N_i^{j_b}(T,w)-M_{i-1}^b(\mathcal{N})\right\}\ , &(2.6)
\end{aligned}$$

其中，等式（2.5）由 $\mathcal{N}$ 的设定和 $M_i^r(\mathcal{N})$ 的定义得出，而等式（2.6）由命题 2.1 推出。令 j^* 为使第一个最大项取得最大值的索引 $j\in[j_r]$，并令b^* 为满足 $j_{b^*}\le j^*$ 的最大索引。我们有：

$$\begin{aligned}
&\max_{j\in[j_r]}\left\{N_i^{j}(T,w)-R_{i-1}^{j-1}(T,w)\right\}-\max_{b\in[r]}\left\{N_i^{j_b}(T,w)-M_{i-1}^b(\mathcal{N})\right\}\\
&\quad\le\quad N_i^{j^*}(T,w)-R_{i-1}^{j^*-1}(T,w)-N_i^{j_{b^*}}(T,w)+M_{i-1}^{b^*}(\mathcal{N})\\
&\quad=\quad N_i^{j^*}(T,w)+R_{i-1}^{j_{b^*}}(T,w)-R_{i-1}^{j_{b^*}}(T,w)-R_{i-1}^{j^*-1}(T,w)-N_i^{j_{b^*}}(T.w)+M_{i-1}^{b^*}(\mathcal{N})\\
&\quad\le\quad \left(M_{i-1}^{b^*}(\mathcal{N})-R_{i-1}^{j_{b^*}}(T,w)\right)+\left(N_i^{j^*}(T,w)-N_i^{j_{b^*}}(T,w)\right)+\left(R_{i-1}^{j_{b^*}}(T,w)-R_{i-1}^{j^*-1}(T,w)\right)\\
&\quad\le\quad (i-2)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}+\left(j^*-j^{b^*}\right)+\left(j^{b^*}-(j^*-1)\right) &(2.7)\\
&\quad=\quad (i-2)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}+1\\
&\quad\le\quad (i-2)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}+\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}\\
&\quad=\quad (i-1)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}\ , &(2.8)
\end{aligned}$$

其中，在等式（2.7）中我们使用了归纳假设。结合等式（2.6）和（2.8），可得

$$M_i^r(\mathcal{N})-R_i^{j_r}(T,w)\le(i-1)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}\ . \tag{2.9}$$

与等式（2.6）类似，

$$R_i^{j_r}(T,w)-M_i^r(\mathcal{N})=\max_{b\in[r]}\left\{N_i^{j_b}(T,w)-M_{i-1}^b(\mathcal{N})\right\}-\max_{j\in[j_r]}\left\{N_i^{j}(T,w)-R_{i-1}^{j-1}(T,w)\right\}\ . \tag{2.10}$$

令 b^{**} 为使第一个最大项达到最大值的索引 $b\in[r]$。我们有

$$\begin{aligned}
&\max_{b\in[r]}\left\{N_i^{j_b}(T,w)-M_{i-1}^b(\mathcal{N})\right\}-\max_{j\in[j_r]}\left\{N_i^{j}(T,w)-R_{i-1}^{j-1}(T,w)\right\}\\
&\quad\le\quad N_i^{j_{b^{**}}}(T,w)-M_{i-1}^{b^{**}}(\mathcal{N})-N_i^{j_{b^{**}}}(T,w)+R_{i-1}^{j_{b^{**}}-1}(T,w)\\
&\quad\le\quad R_{i-1}^{j_{b^{**}}}(T,w)-M_{i-1}^{b^{**}}(\mathcal{N})\ \le\ \left|R_{i-1}^{j_{b^{**}}}(T,w)-M_{i-1}^{b^{**}}(\mathcal{N})\right|\\
&\quad\le\quad (i-2)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}\ \le\ (i-1)\max_{\tau\in[r]}\{j_\tau-j_{\tau-1}\}\ . &(2.11)
\end{aligned}$$

因此（结合公式（2.10）和（2.11））， [7]

$$R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \le (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ . \tag{2.12}$$

综合方程（2.9）和（2.12），可得

$$\left|M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)\right| \le (i-1)\max_{\tau\in[r]}\{j_\tau - j_{\tau-1}\} \ , \tag{2.13}$$

证明至此完成。 ■

在下一个断言中，我们给出 M（$\mathcal{N}$）$-M$（$\mathcal{N}$）之差的界，其中涉及两个矩阵（维度为 $k\times\ell$），并已知它们之间的 L_∞ 距离的上界。稍后我们将应用此断言，其中 $\mathcal{N}=\mathcal{N}$ 对应 $\mathcal{N}$，其定义见断言2.2；而 $\mathcal{N}$ 是一个矩阵，其中包含 N_i^r 的 $N_i^{j_r}$（T，w）的估计值（分别）。我们将在断言 $\widehat{\mathcal{N}}$ 中讨论如何得到2.4。

断言 2.3 令 $\gamma\in(0,1)$，并令 $\mathcal{N}$ 和 $\mathcal{N}$ 为两个 $k\times\ell$ 矩阵。若对每个 $i\in[t]$ 和 $r\in[\ell]$，

$$\left|\widehat{\mathcal{N}}_i^r - \widetilde{\mathcal{N}}_i^r\right| \le \gamma n \ ,$$

then

$$\left|M(\widehat{\mathcal{N}}) - M(\widetilde{\mathcal{N}})\right| \le (2k-1)\gamma n \ .$$

证明：我们将证明，对每个 $t\in[k]$ 以及每个 $r\in[\ell]$，$\left|M_t^r(\widehat{\mathcal{N}}) - M_t^r(\widetilde{\mathcal{N}})\right| \le (2t-1)\gamma n$。我们通过对 t 使用归纳法来证明这一点。
对于 $t=1$ 以及每个 $r\in[\ell]$，我们有

$$\left|M_1^r(\widehat{\mathcal{N}}) - M_1^r(\widetilde{\mathcal{N}})\right| \ = \ \left|\widehat{\mathcal{N}}_1^r - \widetilde{\mathcal{N}}_1^r\right| \ \le \ \gamma n \ . \tag{2.14}$$

现在假设该结论对 $t-1\ge 1$ 以及每个 $r\in[\ell]$ 都成立，并证明它对 t 也成立。对于任意$r\in[\ell]$，根据 $M_t^r(\cdot)$ 的定义，

$$\begin{aligned}
&\left|M_t^r(\widehat{\mathcal{N}}) - M_t^r(\widetilde{\mathcal{N}})\right| \\
&= \ \left|\widehat{\mathcal{N}}_t^r - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\} - \widetilde{\mathcal{N}}_t^r + \max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\}\right| \\
&\le \ \gamma n + \left|\max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\} - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\}\right| \ ,
\end{aligned} \tag{2.15}$$

其中，在最后一个不等式中我们用到了该结论的前提。假设方程（2.15）中的第一个最大项至少与第二个一样大（第二项大于第一项的情形可用类似方法处理），并令 r^* 为使第一个最大项达到最大值的索引。那么，

$$\begin{aligned}
&\left|\max_{r'\in[r]}\left\{\widetilde{\mathcal{N}}_t^{r'} - M_{t-1}^{r'}(\widetilde{\mathcal{N}})\right\} - \max_{r''\in[r]}\left\{\widehat{\mathcal{N}}_t^{r''} - M_{t-1}^{r''}(\widehat{\mathcal{N}})\right\}\right| \\
&\le \ \left|\left(\widetilde{\mathcal{N}}_t^{r^*} - \widehat{\mathcal{N}}_t^{r^*}\right) + \left(M_{t-1}^{r^*}(\widehat{\mathcal{N}}) - M_{t-1}^{r^*}(\widetilde{\mathcal{N}})\right)\right| \\
&\le \ \left|\widetilde{\mathcal{N}}_t^{r^*} - \widehat{\mathcal{N}}_t^{r^*}\right| + \left|M_{t-1}^{r^*}(\widehat{\mathcal{N}}) - M_{t-1}^{r^*}(\widetilde{\mathcal{N}})\right| \\
&\le \ \gamma n + (2t-3)\gamma n = (2t-2)\gamma n \ ,
\end{aligned} \tag{2.16}$$

[7] 实际上，$M_i^r(\mathcal{N}) \ge R_i^{j_r}(T,w)$ 成立，因此 $R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \le 0$；但为使归纳论证更简洁，我们证明 $R_i^{j_r}(T,w) - M_i^r(\mathcal{N})$ 的同一上界也适用于 $M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)$。

这里我们再次使用了该命题的前提以及归纳假设。将公式（2.15）与公式（2.16）结合，即可得出该命题。

下一个命题指出，对于 中的所有取值，我们都能获得良好的估计值 $\left\{N_i^{j_r}(T,\ w)\right\}_{i\in[k]}^{r\in[\ell]}$（只需样本量足够大）。其（标准）证明见附录B。

命题 2.4 对于任意 $\gamma \in (0,1)$ 和 $J=\{j_1,\ldots,j_\ell\}$（满足 $1\le j_1<\cdots<j_\ell=n$），从 T 中抽取大小为 $\Theta\left(\frac{\log(k\cdot\ell)}{\gamma^2}\cdot\right)$ 的样本，我们以至少 $2/3$ 的概率得到估计值 $\left\{\widehat{\mathcal{N}}_i^r\right\}_{i\in[k]}^{r\in[\ell]}$，使得

$$\left|\widehat{\mathcal{N}}_i^r - N_i^{j_r}(T,w)\right| \le \gamma n\ , \tag{2.17}$$

对于每个 $i\in[k]$ 和 $r\in[\ell]$，都有。

现在，我们可以针对均匀分布下的距离近似，重新陈述并证明主要定理。

定理 **1.1** 存在一种针对均匀分布下子序列自由性的基于样本的距离近似算法，抽取大小为 $\Theta\left(\frac{k^2}{\delta^2}\cdot\log\left(\frac{k}{\delta}\right)\right)$ 的样本，并输出估计值 $\widehat{\Delta}$，使得 $|\widehat{\Delta}-\Delta(T,\ w)|\le\delta$ 的概率至少为 $2/3$。[8]

虽然我们关注的是该算法的样本复杂度，但需要指出，其运行时间与样本大小呈线性关系。

证明：算法设定 $\gamma=\delta/(3k)$ 和 $J=\{\gamma n,2\gamma n,\ldots,n\}$。首先，在上述 γ 设置下应用命题 2.4，得到每个 $i\in[k]$ 和 $r\in[\ell]$ 对应的估计值 $\left\{\widehat{\mathcal{N}}_i^r\right\}$；以至少 $2/3$ 的 ~ 概率，这些估计值满足方程 (2.17) 中的陈述。若将命题 2.2 中定义的 $\mathcal{N}=\mathcal{N}$ 用作 $\mathcal{N}$，则命题 2.3 的前提成立。因此，我们可以应用命题 2.3，并将 ^ 与命题 2.2 以及 J 的定义结合起来，得到：对于矩阵 $\mathcal{N}$，以至少 $2/3$ 的概率，

$$\left|M(\widehat{\mathcal{N}})-R(T,w)\right| \le (2k-1)\gamma n+(k-1)\gamma n=(3k-2)\gamma n\le\delta n\ . \tag{2.18}$$

因此，该算法基于定义 $M(\widehat{\mathcal{N}})$，以迭代方式计算 $=M_k^\ell(\widehat{\mathcal{N}})$，2.2，^ ^并输出 $\Delta=M(\mathcal{N})/n$。由于 $R(T,\ w)/n=\Delta(T,\ w)$，定理得证。 ■

3 分布无关距离近似

如引言中所述，我们用于近似一般分布 p 下子序列自由性距离的算法，是通过将问题归约为近似均匀分布下的子序列自由性距离来实现的。不过，我们无法直接使用第 2 节中介绍的算法。主要有两个障碍，下面将简要说明。在归约中，给定一个词 w 以及从文本 T中获取样本的权限，其中样本服从 p，~ ~所定义的分布，我们定义一个词 $\widetilde{w}$ 和一个文本 T，使得如果能够得到 $\Delta(T,\widetilde{w})$的良好近似，就能得到 $\Delta(T,w,p)$的良好近似。（回想一下，$\Delta(T,w,p)$表示在分布 p下，T距离满足 w-自由性的距离。）然而，首先，我们实际上无法直接获取

[8]与往常一样，对于任意 $\eta>0$，我们可以将成功概率提高到 $1-\eta$，样本复杂度仅需乘以 $O(\log(1/\eta))$。

~服从均匀分布的样本 T，其次，我们无法像在第 J 节中那样，使用一组 2索引来诱导大小相同且有界的区间。

~我们分几个阶段来应对这些挑战（并精确定义 T 和 $\widetilde{w}$）。在第 3.1 节和 3.2 节中，我们首先根据 p进行抽样，以便在 T 中构造具有某些性质的区间（且这些性质以足够高的概率成立）。这些区间的作用将在接下来的其他小节中逐渐明了。

3.1 区间的构造与分类

我们先在本小节中定义 $[n]$ 中由 p （算法未知）决定的区间。随后，我们从 p 中抽样来构造区间；从某种意义上说，后者是前者的近似（我们之后会对此作形式化说明）。根据 p 所给出的（近似）权重，每个构造出的区间都会被分类为“重”或“轻”。理想情况下，我们希望所有区间都是轻区间，但又不能太轻，这样区间的数量就不会太多（正如我们在均匀分布下所做的那样，当时我们只需定义大小相等的区间）。然而，对于一般分布 p ，可能存在某些单个索引 $j \in [n]$ ，使得 p_j 很大，因此我们还需要允许重区间（每个重区间仅包含一个索引）。我们将使用以下两个定义。

定义 3.1 对于任意两个整数 $j_1 \le j_2$，令 $[j_1, j_2]$ 表示区间 $\{j_1, \ldots, j_2\}$。对于每个 j_1、 $j_2 \in [n]$，定义

$$\mathrm{wt}_p([j_1, j_2]) \stackrel{\mathrm{def}}{=} \sum_{j=j_1}^{j_2} p_j$$

为根据 p 得到的区间 $[j_1, j_2]$ 的权重。我们将使用简写 $\mathrm{wt}_p(j)$ 表示 $\mathrm{wt}_p([j, j])$。

定义 3.2 令 S 为一个大小为 s、元素取自 $[n]$ 的多重集。对于每个 $j \in [n]$，令 $N_S(j)$ 表示 S 中等于 j 的元素数量。对于每个 j_1、 $j_2 \in [n]$，定义

$$\mathrm{wt}_S([j_1, j_2]) \stackrel{\mathrm{def}}{=} \frac{1}{s} \sum_{j=j_1}^{j_2} N_S(j)$$

为根据 S 得到的区间 $[j_1, j_2]$ 的估计权重。我们将使用简写 $\mathrm{wt}_S(j)$ 表示 $\mathrm{wt}_S([j, j])$。

在下一个定义以及本节余下部分中，我们将使用

$$z \;=\; c_z \frac{k}{\delta}\,, \tag{3.1}$$

其中令 $c_z = 100$。

接下来，我们基于 p 定义上述区间集合。粗略来说，我们尽可能使各区间的权重均衡，同时考虑到某些索引的权重可能很大，因此将它们各自分配到一个单独的区间。

定义 3.3 按照以下迭代方式定义一个索引序列。令 $h_0 = 0$，并且对于 $\ell = 1\ 2\ \ldots$ $h_{\ell-1} < n$ h_ℓ $_p\ h_{\ell-1}+1\ > \frac{1}{8z}$ $h_\ell =$、……，只要，就按如下方式定义 $h_{\ell-1}+1$ h_ℓ $h'_\ell \in [h_{\ell-1}+1, n]$ $_p\ [h_{\ell-1}+1, h'_\ell]\ \le \frac{1}{4z}$ 。如果 wt ( ) ，则 。否则，令 为满足 wt ( ) 且对于每个 $h''_\ell \in [h_{\ell-1}+1, h'_\ell]$，$\mathrm{wt}_p(h'') \le \frac{1}{8z}$ 的最大索引。令 L 使得 $h_L = n$。

根据上述定义的索引 $\{h_\ell\}_{\ell=0}^{L}$，对于每个 $\ell \in [L]$，令 $H_\ell = [h_{\ell-1}+1, h_\ell]$，并令$\mathcal{H} = \{H_\ell\}_{\ell=1}^{L}$。我们将 $\mathcal{H}$ 划分为以下三个子集。令 $\mathcal{H}_{sin}$ 为所有满足 $|H| = 1$ 且 $\mathrm{wt}_p(H) > \frac{1}{8z}$ 的 $H \in \mathcal{H}$构成的子集。令 $\mathcal{H}_{med}$为所有满足 $|H| \neq 1$ 且 $\frac{1}{8z} \le \mathrm{wt}_p(H) \le \frac{1}{4z}$ 的 $H \in \mathcal{H}$ 构成的集合。令 $\mathcal{H}_{sml}$为所有满足 $\mathrm{wt}_p(H) < \frac{1}{8z}$ 的 $H \in \mathcal{H}$ 构成的集合。

注意，由于 $\mathrm{wt}_p(T) = 1$，因此 $|\mathcal{H}_{sin} \cup \mathcal{H}_{med}| \le 8z$。此外，由于每个 H'、$H'' \in \mathcal{H}_{sml}$ 之间至少要有一个 $H \in \mathcal{H}_{sin}$，因此我们还有 $|\mathcal{H}_{sml}| \le 8z + 1$。

根据其定义，$\mathcal{H}$ 由 p决定。接下来，我们根据 p 进行抽样，构造一个区间集合 $\mathcal{B}$（方式与定义3.3类似，但并不完全相同）。考虑一个从 p 中按 s_1 抽取的大小为 S_1 的样本（允许重复抽样），其中 s_1 将稍后确定。

定义 3.4 给定一个大小为 s_1 的样本 S_1（$[n]$ 中元素构成的多重集），按以下迭代方式确定一个索引序列。令 $b_0 = 0$，并且对于 $u = 1, 2, \ldots\ldots$，只要 $b_{u-1} < n$ 成立，就将 b_u 定义如下。如果 $\mathrm{wt}_{S_1}(b_{u-1}+1) > 1/z$，则 $b_u = b_{u-1}+1$。否则，令 b_u 为满足 $\mathrm{wt}_{S_1}([b_{u-1}+1, b'_u]) \le \frac{1}{z}$ 的最大索引 $b'_u \in [b_{u-1}+1, n]$。令 U 满足 $b_U = n$。

基于上述定义的索引 $\{b_u\}_{u=0}^{U}$，对于每个 $u \in [U]$，令 $B_u = [b_{u-1}+1, b_u]$，并令 $\mathcal{B} = \{B_u\}_{u=1}^{U}$。对于每个 $u \in [U]$，若 $\mathrm{wt}_{S_1}(B_u) > \frac{1}{z}$，则称 B_u 为重，否则为轻。

注意，每个重区间都只包含一个元素。

为了建立 $\mathcal{H}$ 与 $\mathcal{B}$ 之间的联系，我们基于样本 S_1 引入以下事件。

定义 3.5 记 E_1 为以下事件：

$$\forall H \in \mathcal{H}_{sin} \cup \mathcal{H}_{med},\ \ \frac{1}{2}\mathrm{wt}_p(H) \le \mathrm{wt}_{S_1}(H) \le \frac{3}{2}\mathrm{wt}_p(H)\ , \tag{3.2}$$

$$\forall H \in \mathcal{H}_{sml},\ \ \mathrm{wt}_{S_1}(H)\ \le \frac{1}{2z}\ . \tag{3.3}$$

断言 3.1 若样本 S_1 的大小为 $s_1 = 120z\log(240z)$，则

$$\Pr\left[E_1\right] \ge \frac{8}{10}\ ,$$

其中，概率是针对 S_1 的选择而言的。

证明：回顾一下，$\mathrm{wt}_p(H) \ge \frac{1}{8z}$ 对每个 $H \in \mathcal{H}_{sin} \cup \mathcal{H}_{med}$ 都成立。利用乘法型 Chernoff 界（见定理 A.1），我们得到对每个 $H \in \mathcal{H}_{sin} \cup \mathcal{H}_{med}$

$$\Pr\left[\left|\mathrm{wt}_{S_1}(H) - \mathrm{wt}_p(H)\right| > \frac{1}{2}\mathrm{wt}_p(H)\right] < 2\exp\left(-\frac{1}{12}\mathrm{wt}_p(H)s_1\right) < \frac{1}{120z}\ . \tag{3.4}$$

接下来回顾一下，对于每个 $H \in \mathcal{H}_{sml}$，$\mathrm{wt}_p(H) \le \frac{1}{4z}$。定义随机变量 $\{\chi_r\}_{r=1}^{s_1}$，使得对于每个 $r \in [s_1]$，$\chi_r = 1$ 以概率 $\frac{1}{4z}$ 取值，否则取值为 0。再次使用

乘法型 Chernoff 界，我们得到对于每个 $H \in \mathcal{H}_{sml}$

$$\begin{aligned}\Pr\left[\mathrm{wt}_{S_1}(H) > 2\frac{1}{4z}\right] &\leq \Pr\left[\sum_{r=1}^{s_1}\chi_r > 2\frac{1}{4z}\right] < \exp\left(-\frac{1}{3}\frac{1}{4z}s_1\right)\\ &= \exp\left(-\frac{1}{12z}s_1\right) < \frac{1}{240z}\,. \end{aligned} \tag{3.5}$$

对所有 $H \in \mathcal{H}_{sin} \cup \mathcal{H}_{med} \cup \mathcal{H}_{sml}$ 应用并合界（回想定义 3.3 后的讨论，$|\mathcal{H}_{sin} \cup \mathcal{H}_{med}| \leq 8z$ 以及 $|\mathcal{H}_{sml}| \leq 8z+1$），我们得到

$$\Pr[E_1] \geq 1 - 8z \cdot \frac{1}{120z} - 9z \cdot \frac{1}{240z} \geq \frac{8}{10}\,, \tag{3.6}$$

于是断言得证。 ■

断言 3.2 条　　件是在事件 E_1 的条件下，对于每个满足 B_u 为轻的 $u \in [U]$，$\mathrm{wt}_p(B_u) < \frac{6}{z}$.

证明：考虑一个轻区间 B_u。令 $\mathcal{H}_4$ 为满足 $\mathcal{H}_4 = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$ 的最小集合，其中 $\mathcal{H}_1 \subseteq \mathcal{H}_{sin}$、$\mathcal{H}_2 \subseteq \mathcal{H}_{med}$、$\mathcal{H}_3 \subseteq \mathcal{H}_{sml}$ 且 $B_u \subseteq \mathcal{H}_4$。令 $\mathcal{H}_5$ 为 $\mathcal{H}_4$ 的最小子集，使得 $\mathcal{H}_4 \setminus \mathcal{H}_5 \subseteq B_u$。容易验证 $\mathcal{H}_5 \subseteq \mathcal{H}_2 \cup \mathcal{H}_3$，并且 $|\mathcal{H}_5| \leq 2$。

在满足 E_1（定义 3.5、公式 (3.2)）的条件下，对于每个$H \in \mathcal{H}_1 \cup \mathcal{H}_2$，都有 $\mathrm{wt}_p(H) < 2\mathrm{wt}_{S_1}(H)$，因此对于每个 $H \in \mathcal{H}_1$，同样有 $\mathrm{wt}_{S_1}(H) > \frac{1}{16z}$。由于 B_u是轻的，$\mathrm{wt}_{S_1}(B_u) < \frac{1}{z}$，这意味着 $|\mathcal{H}_1| \leq 17$。如前所述，在任意一对区间 H'、$H'' \in \mathcal{H}_{sml}$之间至少存在一个区间 $H \in \mathcal{H}_{sin}$，这意味着 $|\mathcal{H}_3| \leq |\mathcal{H}_1| + 2 \leq 19$。因此，

$$\mathrm{wt}_p(B_u) \leq \sum_{H\in\mathcal{H}_4}\mathrm{wt}_p(H) = \sum_{H\in\mathcal{H}_1}\mathrm{wt}_p(H) + \sum_{H\in\mathcal{H}_2}\mathrm{wt}_p(H) + \sum_{\mathcal{H}\in H_3}\mathrm{wt}_p(H) \tag{3.7}$$

$$\leq 2\sum_{H\in\mathcal{H}_1}\mathrm{wt}_{S_1}(H) + 2\sum_{H\in\mathcal{H}_2}\mathrm{wt}_{S_1}(H) + |\mathcal{H}_3|\frac{1}{8z} \leq 2\mathrm{wt}_{S_1}(B_u) + 2\sum_{H\in\mathcal{H}_5}\mathrm{wt}_{S_1}(H) + \frac{19}{8z} \tag{3.8}$$

$$\leq 2\mathrm{wt}_{S_1}(B_u) + 4\max_{H\in\mathcal{H}_2\cup\mathcal{H}_3}\{\mathrm{wt}_{S_1}(H)\} + \frac{19}{8z} \tag{3.9}$$

$$\leq \frac{2}{z} + \frac{1}{z} + \frac{19}{8z} < \frac{6}{z}\,, \tag{3.10}$$

结论得证。 ■

3.2 区间的符号密度和权重估计

在本小节中，我们根据 p估计每个满足 $u \in U$的区间 $[b_u]$ 的权重及其符号密度，重点关注出现在 w中的符号。注意，$[b_u]$ 是区间B_1、.....、B_u的并集。我们首先介绍一些记号。

对于任意词 w^*、文本 T^*、$i \in [|w^*|]$ 和 $j \in [|T^*|]$，若 $T^*[j] = w_i^*$，则令 $I_i^j(T^*, w^*)$ 为 $= 1$，否则为 0。接下来定义

$$\xi_i^u = \sum_{j\in[b_u]} I_i^j(T, w)p_j\,. \tag{3.11}$$

考虑按照 p 选取的大小为 s_2 的样本 S_2 （允许重复抽样），其中 s_2 将在稍后设定。对于每个 $u \in [U]$ 和 $i \in [k]$，令

$$\breve{\xi}_i^u = \frac{1}{s_2} \sum_{j \in [b_u]} I_i^j(T, w) N_{S_2}(j) \,. \tag{3.12}$$

定义 3.6 事件 E_2 （基于 S_2）定义如下。对于每个 $i \in [k]$ 和 $u \in [U]$，

$$\left| \breve{\xi}_i^u - \xi_i^u \right| \leq \frac{1}{z} \,, \tag{3.13}$$

并且对于每个 $u \in [U]$

$$|\mathrm{wt}_{S_2}([b_u]) - \mathrm{wt}_p([b_u])| \leq \frac{1}{z} \,. \tag{3.14}$$

断言 3.3 如果样本 S_2 的大小为 $s_2 = z^2 \log(40kU)$，则

$$\Pr[E_2] \geq \frac{9}{10} \,,$$

其中，概率是针对 S_2 的选择而言的。

证明： 利用加性切尔诺夫界（参见定理 A.1），结合以下事实：$\mathbb{E}\left[N_{S_2}(\underset{s_2}{j}\, I_i^j)(T, w) \right] = p_j I_i^j(T, w)$，可得~~如下~~结果。

$$\Pr\left[\left| \breve{\xi}_i^u - \xi_i^u \right| > \frac{1}{z} \right] = \Pr\left[\left| \frac{1}{s_2} \sum_{j \in [b_u]} I_i^j(T, w) N_{S_2}(j) - \sum_{j \in [b_u]} I_i^j(T, w) p_j \right| > \frac{1}{z} \right] \tag{3.15}$$

$$< 2\exp(-2\frac{1}{z^2} s_2) \leq \frac{1}{20kU} \,. \tag{3.16}$$

对所有 $i \in [k]$ 和 $u \in [U]$应用并合界，可得其成立的概率至少为19 20，$\left| \breve{\xi}_i^u - \xi_i^u \right| \leq \frac{1}{z}$。再利用加性~~切~~尔诺夫界以及 $\mathbb{E}\left[N_{S_2}(\underset{s_2}{j}\,) \right] = p_j$)这一事实，可得

$$\Pr\left[|\mathrm{wt}_{S_2}([b_u]) - \mathrm{wt}_p([b_u])| > \frac{1}{z} \right] = \Pr\left[\left| \frac{1}{s_2} \sum_{j \in [b_u]} N_{S_2}(j) - \sum_{j \in [b_u]} p_j \right| > \frac{1}{z} \right] \tag{3.17}$$

$$\leq 2\exp(-2\frac{1}{z^2} s_2) \leq \frac{1}{20U} \,. \tag{3.18}$$

再次对所有 $u \in [U]$应用并合界，可得以至少19 20的概率有 $|\mathrm{wt}_{S_2}([b_u]) - \mathrm{wt}_p([b_u])| \leq \frac{1}{z}$。最后再应用一次并合界，可得 $\Pr[E_2] \geq \frac{9}{10}$ ■

3.3 从分布无关约化到均匀分布

在本小节中，我们将详细介绍上述从无分布情形到均匀情形的归约，使用前面各小节中定义的区间和估计量。我们先给出三个定义，这些定义取自 [29]，，并将在归约中用到。前两个定义针对拆分这一概念（此前的研究中也曾使用过这一概念的变体，例如 [14]）。

~ ~定义 3.7 对于文本 $T = t_1 \dots t_n$，若文本 T满足 $T = t_1^{\alpha_1} \dots t_n^{\alpha_n}$for ~some $\alpha_1 \dots \alpha_n \in \mathbb{N}^+$，则称其为 T的拆分。我们将拆分映射记作 ~ϕ T，它将 的每个符号（的索引）映射到其在 T中的来源。形式化地， ϕ: $[|T|] \to [n]$ 定义如下。对于每个 $\ell \in [|T|] = [\sum_{i=1}^{n} \alpha_i]$，令 $\phi(\ell)$ 为满足 $\sum_{r=1}^{i-1} \alpha_r < \ell < \sum_{r=1}^{i} \alpha_r$的唯一 $i \in [n]$。

~注意，根据此定义， ϕ 是一个非递减满射，满足 $T[\ell] = T[\phi(\ell)]$ 对~ ~每个 $\ell \in [|T|]$都成立。对于集合 $S \subseteq [|T|]$，我们令 $\phi(S) = \{\phi(\ell) : \ell \in S\}$。稍微滥用一下记号，对于任意 $i \in [n]$，我们用 $\phi^{-1}(i)$ 表示集合 $\left\{\ell \in [|\widetilde{T}|] : \phi(\ell) = i\right\}$；对于集合 $S \subseteq [n]$，我们令 $\phi^{-1}(S) = \left\{\ell \in [|\widetilde{T}|] : \phi(\ell) \in S\right\}$

定义 3.8 给定文本 $T = t_1 \dots t_n$及其对应的概率分布 $p = (p_1, \dots, p_n)$，(T, p) 的一种拆分是文本 T及其对应的概率分布 ~^$p = (\hat{p}_1, \dots, \hat{p}_{|\widetilde{T}|})$，其中 T是 T 的一种拆分，并且对于每个 $i \in [n]$，都有 $\sum_{\ell \in \phi^{-1}(i)} \hat{p}_\ell = p_i$。

第三个 定义的是一个词集，其中任意两个连续符号都不相同 .

定义 3.9 令 $\mathcal{W}_c = \{w : w_{j+1} \neq w_j, \forall j \in [k-1]\}$。

3.3.1 从分布无关约化到均匀分布的基础

~ ~令 $\widetilde{w}$ 为一个长度为 k 的词， T 为一个长度为 $\widetilde{n}$的文本。本小节将证明一个命题，该命题^给出了估计矩阵 $\mathcal{N}$（的归一化版本）满足的充分条件，在这些条件下，它可以~用于以较小的加性误差估计 $\Delta(T, \widetilde{w})$。

我们首先陈述一个与命题 2.2类似的命题，只有一个虽小但重要的区别：~会考虑 T 中的区间（由指标集 J确定），这些区间由同一个符号的重复抽样组成。由于其证明与命题 2.2的证明非常相似，因此将其推迟到附录B。~回顾一下， $M(\cdot)$ 在定义 2.2中已有定义，而 $R(T, \widetilde{w})$ 表示角色不相交的~副本 $\widetilde{w}$ 在 T中的数量。

命题 3.4 设 $J = \{j_0, j_1, \dots, j_\ell\}$是满足 $j_0 = 0 < j_1 < j_2 < \dots < j_\ell = \widetilde{n}$的指标集。设 $\mathcal{N}$是一个矩阵，其元素对于每个 $i \in [\widetilde{k}]$和 $r \in [\ell]$均为 $\mathcal{N}_i^r = N_i^{j_r}(\widetilde{T}, \widetilde{w})$。设 $J' = \{r \in$ ~$[\ell] : T[j_{r-1}+1] = \dots = T[j_r]\}$。则

$$\left| M(\mathcal{N}) - R(\widetilde{T}, \widetilde{w}) \right| \leq (\widetilde{k} - 1) \cdot \max_{r \in [\ell] \setminus J'} \{(j_r - j_{r-1})\} \ .$$

下面的观察可以通过归纳法轻松证明。

^ ~观察 3.5 设 $\mathcal{N}$是一个大小为 $k \times \ell$的矩阵。则

$$\frac{1}{\widetilde{n}} M(\widehat{\mathcal{N}}) = M\left(\frac{\widehat{\mathcal{N}}}{\widetilde{n}}\right) \ . \tag{3.19}$$

接下来的命题将作为我们从一般的无分布情形归约到均匀情形的基础。

^ ~断言 3.6 设 $\mathcal{N}$为一个 $k \times \ell$ 矩阵， $J = \{j_0, j_1, j_2, \dots, j_\ell\}$ 为满足 $j_0 = 0 < j_1 < j_2 < \dots < j_\ell = \widetilde{n}$ 的指标集，并设 c_1 和 c_2 为常数。假设满足以下条件。

1. 对于每个 $r \in [\ell]$，如果 $j_r - j_{r-1} > c_1 \cdot \frac{\delta\widetilde{n}}{\widetilde{k}}$，那么 $\widetilde{T}[j_{r-1}+1] = \cdots = \widetilde{T}[j_r]$。

2. 对于每个 $i \in [\widetilde{k}]$ 和 $r \in [\ell]$，$\left|\widehat{\mathcal{N}}_i^r - N_i^{j_r}(\widetilde{T}, \widetilde{w})\right| \le c_2 \cdot \frac{\delta\widetilde{n}}{\widetilde{k}}$。

然后，

$$\left| M\left(\frac{\widehat{\mathcal{N}}}{\widetilde{n}}\right) - \Delta(\widetilde{T}, \widetilde{w}) \right| \le (c_1 + 2c_2)\delta \, .$$

证明：令 $\mathcal{N}$ 为矩阵，其条目对于每个 $i \in [\widetilde{k}]$ 和 $r \in [\ell]$ 都是 $\mathcal{N}_i^r = N_i^{j_r}(\widetilde{T}, \widetilde{w})$。我们使用命题 3.4 和当前命题前提中的条目 1，得到 $\left|M(\mathcal{N}) - R(\widetilde{T}, \widetilde{w})\right| \le c_1\delta\widetilde{n}$。我们还使用断言 2.3 和当前命题前提中的条目 2，得到 $\left|M(\widehat{\mathcal{N}}) - M(\mathcal{N})\right| \le 2c_2\delta\widetilde{n}$。结合这些界限，我们得到 $\left|M(\widehat{\mathcal{N}}) - R(\widetilde{T}, \widetilde{w})\right| \le (c_1+2c_2)\delta\widetilde{n}$。

通过应用观察 3.5 以及以下事实 $\frac{R(\widetilde{T}, \widetilde{w})}{\widetilde{n}} = \Delta(\widetilde{T}, \widetilde{w})$。 ■

3.3.2 建立针对 $w \in \mathcal{W}_c$ 且 p 量化的归约

为便于阅读，我们首先讨论这样一种特殊情形：$w \in \mathcal{W}_c$ (回忆定义 3.9)，且存在 $\beta \in (0, 1)$ 使得 p_j/β 对每个 $j \in [n]$ 都是整数。稍后我们将说明如何处理一般情形；在这种情形下，我们会借助 [29] 中的技术，并引入一些实现算法所需的新技术。

对于本小节这种情形，令 $\widetilde{n} = \frac{n}{\beta}$ 对每个 $j \in [n]$ 都成立，因此 $|\widetilde{T}| = \frac{1}{\beta}$。定义 $\widetilde{p}$ 为 $\widetilde{p}_j = \beta$ 对每个 $j \in [|\widetilde{T}|]$，使得 $\widetilde{p}$ 为均匀分布。由于 $p_j = \beta \cdot \alpha_j$，对每个 $j \in [n]$，我们有 $(T, \widetilde{p})$ 是 (T, p) 的拆分（回忆定义 3.8），因此根据 [29, 结论 4。4] (利用假设 $w \in \mathcal{W}_c$)，

$$\Delta(\widetilde{T}, w, \widetilde{p}) = \Delta(T, w, p) \, . \tag{3.20}$$

记 $\widetilde{n} = |\widetilde{T}|$。我们首先定义一个由以下对象构成的区间集合 $[\widetilde{n}]$，其中 $\{b_0, \ldots, b_U\}$ 和 $\mathcal{B} = \{B_1, \ldots, B_U\}$ 的定义见第 3.1节，而 ϕ 的定义见定义 3.8。

定义 3.10 令 $b_0 = 0$，对于每个 $u \in [U]$，令 $b_u = \max\{h \in [\widetilde{n}] : \phi(h) = b_u\}$。对于每个$u \in [U]$，令 $\widetilde{B}_u = [\widetilde{b}_{u-1} + 1, \widetilde{b}_u]$，并定义 $\widetilde{\mathcal{B}} = \left\{\widetilde{B}_u\right\}_{u=1}^{U}$。

接下来，我们引入一种记号，根据 $\widetilde{p}$ 表示这些区间并集的权重。对于每个 $i \in [k]$ 和 $u \in [U]$，

$$\widetilde{\xi}_i^u = \sum_{j \in [\widetilde{b}_u]} I_i^j(\widetilde{T}, w)\widetilde{p}_j \, . \tag{3.21}$$

注意，

$$\widetilde{\xi}_i^u = \frac{1}{\widetilde{n}} N_i^{b_u}(\widetilde{T}, w) \, . \tag{3.22}$$

断言 3.7 对于每个 $i \in [k]$ 和 $u \in [U]$

$$\widetilde{\xi}_i^u = \xi_i^u \, ,$$

其中 ξ_i^u 的定义见公式 (3.11)。

证明：

$$\xi_i^u = \sum_{j\in[b_u]} I_i^j(T,w)p_j = \sum_{j\in[b_u]} I_i^j(T,w) \sum_{\widetilde{j}\in\phi^{-1}(j)} \widetilde{p}_{\widetilde{j}} \tag{3.23}$$

$$= \sum_{j\in[b_u]} \sum_{\widetilde{j}\in\phi^{-1}(j)} I_i^j(T,w)\widetilde{p}_{\widetilde{j}} = \sum_{j\in[b_u]} \sum_{\widetilde{j}\in\phi^{-1}(j)} I_i^{\widetilde{j}}(\widetilde{T},w)\widetilde{p}_{\widetilde{j}} = \sum_{\widetilde{j}\in\widetilde{b}_u} I_i^{\widetilde{j}}(\widetilde{T},w)\widetilde{p}_{\widetilde{j}} = \widetilde{\xi}_i^u \,, \tag{3.24}$$

至此，命题得证。 ■

现在我们可以陈述并证明以下引理。

引理 3.8 设 w 是 $\mathcal{W}_c$ 中长度为 k 的一个词， T 是长度为 n 的文本， p 是 $[n]$ 上的一个分布，并且存在 $\beta \in \in(0,1)$，使得对每个 $j \in [n]$， p_j/β 都是整数。存在一种算法，对于给定的参数 $\delta \in \in(0,1)$，从 T 中按照 p 进行分布抽取一个大小为 $\Theta\left(\frac{k^2}{\delta^2}\cdot\log\left(\frac{k}{\delta}\right)\right)$ 的样本，并输出一个估计值 Δ，使得 $|\Delta - \Delta(T,w,p)| \le \delta$ 的概率至少为 $2/3$。

与单一 形式情形一样，该算法的运行时间与样本大小呈线性关系 .

证明：算法首先抽取一个大小为 $s_1 = 120z\log(240z)$ 的样本 S_1，并按照定义 3.4 构造区间集合 $\mathcal{B}$。接着，算法按照 $40kU$ 再抽取一个大小为 $s_2 = z^2\log(40kU)$ 的样本 S_2，并据此定义一个大小为 $k\times U$ 的估计矩阵 $\widehat{\xi}$，如下所示。对于每个 $i\in[k]$ 和 $u\in[U]$，令 $\widehat{\xi}[i][u] = \check{\xi}_i^u$，其中 $\check{\xi}_i^u$ 如方程 (3.12) 中所定义。最后，该算法输出 $\Delta = M(\xi)$，其中 M 如定义 2.2 中所定义。

T，$\widehat{\ }$我们主要证明概率至少为 $3/5$ 以证明根据 s_1 的设定，应用断言 3.1 可知，$\widetilde{\ }$事件以至少 $8/10$ 的概率发生 E_1，如定义 3.5所述。根据 s_2的设定，应用断言 3.3 可知，事件以至少 $9/10$的概率发生 E_2，如定义 3.6所述。下文均假设这两个事件同时发生（其概率至少为 $7/10$）。

为了应用断言 3.6，我们令 $\widetilde{w} = w$， $J = \left\{\widetilde{b}_0,\ \widetilde{b}_1,\ \ldots\ldots,\ \widetilde{b}_U\right\}$（回顾定义 3.10），并令 $\widehat{\mathcal{N}} = \widetilde{n}\widehat{\xi}$，其中 $\widehat{\xi}$ 如上所定义。此外，我们令 $c_1 = \frac{1}{2}$ 并令 $c_2 = \frac{1}{4}$。接下来，我们证明该断言前提中的两个条件均满足。

为证明条目 1 得到满足，我们首先注意到，由于 $\widetilde{p}$ 是均匀的，因此对于每个 $u\in U$，$\widetilde{\ }\mathrm{wt}_{\widetilde{p}}(b_u) = \frac{b_u - b_{u-1}}{\widetilde{n}}$。我们利用断言 3.2 的推论（回顾一下，我们以 E_1 为条件），据此，对于每个满足 $\frac{b_u - b_{u-1}}{\widetilde{n}} \ge \frac{6}{z}$ 的 u， B_u 都是重的（因为对于每个 $u\in U$，$\mathrm{wt}_{\widetilde{p}}(\widetilde{B}_u) = \mathrm{wt}_p(B_u)$）。根据定义 3.4，这意味着 B_u 只包含一个索引，因此 $T[b_{u-1}+1] = \cdots = T[b_u]$。根据 z 的定义（公式（3.1））以及 c_1 的设定，该条目得到满足。

为证明条目 2 得到满足，我们结合 E_2（定义 3.6，公式（3.13））的定义与断言 3.7，由此得到 $|\widehat{\xi}_i^u - \widetilde{\xi}_i^u| \le \frac{1}{z}$ 对于每个 $i\in[k]$ 以及 $u\in[U]$。根据公式（3.22）、 z 的定义以及 c_2 的设定，我们得到该条目得到满足。

应用断言 3.6 后，我们根据 $\widetilde{\ }$ $\widehat{\ }$ c_1 和 $\widetilde{\ }$ c_2 的设定，该值至多为 δ。由于 $\widetilde{p}$ 是均匀分布， $\Delta(T,w) = \Delta(T,w,\widetilde{p})$，并且由于 $\widetilde{\ }\Delta(T,w,\widetilde{p}) = \Delta(T,w,p)$（根据方程 (3.20)），引理得证。 ■

在接下来的小节中，我们将讨论一般情形：此时不一定有$w\in\mathcal{W}_c$，也不一定存在某个值 β，使得对每个 $j\in[n]$， p_j/β 都是整数。在这种一般情形下，我们需要采取一些额外步骤，才能进行拆分。首先，

将其归约到一个略有不同的分布，然后再归约到 $w \in \mathcal{W}_c$。虽然这遵循 [29], ，但为了实现我们的算法，我们需要在此过程中说明如何定义 $\widehat{\ }$ $\widehat{\ }$估计矩阵 $\xi\ (= \mathcal{N}/\widetilde{n})$ 以及相应的指标集 J ，以便应用命题 3.6，这与引理 3.8的证明中所示的方法类似。

3.4 量化分布

设 $\eta = c_\eta \frac{1}{nz}$，其中 $c_\eta = \frac{1}{16}$。我们通过对每个 $j \in [n]$，将 p_j“舍入”到最接近的更大整数倍 η，来定义 $\ddot{p}$。也就是说，对每个 $j \in [n]$，有 $\ddot{p}_j = \left\lceil \frac{p_j}{\eta} \right\rceil \eta$。根据此定义， $L_1(\ddot{p}, p) = \sum_{j=1}^n |\ddot{p}_j - p_j| \le \eta n = \frac{c_\eta}{z}$。我们将 $\dot{p}$ 定义为 $\ddot{p}$ 的归一化版本。也就是说，令 $\zeta = \frac{1}{\sum_{j=1}^n \ddot{p}_j}$，则对每个 $j \in [n]$，设 $\dot{p}_j = \zeta \ddot{p}_j$，并注意到 $\zeta \le 1$。观察可知， $L_1(\dot{p}, \ddot{p}) = \sum_{j=1}^n |\zeta \ddot{p}_j - \ddot{p}_j| = \frac{|\zeta - 1|}{\zeta} = \frac{1}{\zeta} - 1$。另请注意， $\frac{1}{\zeta} = \sum_{j=1}^n (p_j + (\ddot{p}_j - p_j)) = 1 + \sum_{j=1}^n (\ddot{p}_j - p_j)$。我们有$L_1(\dot{p}, \ddot{p}) = \frac{1}{\zeta} - 1 = \sum_{j=1}^n (\ddot{p}_j - p_j) \le \sum_{j=1}^n |\ddot{p}_j - p_j| = L_1(p, \ddot{p})$。利用三角不等式可得

$$L_1(p, \dot{p}) \le L_1(p, \ddot{p}) + L_1(\ddot{p}, \dot{p}) \le 2L_1(p, \ddot{p}) \le 2\frac{c_\eta}{z} . \tag{3.25}$$

根据 [29,命题 4.1] ，我们有

$$|\Delta(T, w, p) - \Delta(T, w, \dot{p})| \le L_1(p, \dot{p}) . \tag{3.26}$$

Fin另外，注意对于每个 $j \in [n]$， $\dot{p}_j$ 都是 $\zeta\eta$ 的整数倍，因为 $\dot{p}_j = \zeta\eta \left\lceil \frac{p_j}{\eta} \right\rceil$.

对于每个 $i \in [k]$ 和 $u \in [U]$，令

$$\dot{\xi}_i^u = \sum_{j \in [b_u]} I_i^j(T, w) \dot{p}_j . \tag{3.27}$$

命题 3.9 对于每个 $i \in [k]$ 和 $u \in [U]$，

$$\left| \dot{\xi}_i^u - \xi_i^u \right| \le \sum_{j \in [b_u]} |\dot{p}_j - p_j| , \tag{3.28}$$

并且对于每个 $u \in [U]$

$$|\mathrm{wt}_{\dot{p}}([b_u]) - \mathrm{wt}_p([b_u])| \le \sum_{j \in [b_u]} |\dot{p}_j - p_j| . \tag{3.29}$$

证明：等式（3.28）由三角不等式以及以下事实可得：对于每个 $j \in [n]$， $i \in [k]$， $I_i^j\ (T,\ w) \le 1$

$$\left| \dot{\xi}_i^u - \xi_i^u \right| = \left| \sum_{j \in [b_u]} I_i^j(T, w)(\dot{p}_j - p_j) \right| \le \sum_{j \in [b_u]} |\dot{p}_j - p_j| . \tag{3.30}$$

方程（3.29）由三角不等式得出

$$|\mathrm{wt}_{\dot{p}}([b_u]) - \mathrm{wt}_p([b_u])| = \left| \sum_{j \in [b_u]} \dot{p}_j - p_j \right| \le \sum_{j \in [b_u]} |\dot{p}_j - p_j| , \tag{3.31}$$

由此，结论得证。 ■

3.5 处理 $w \notin \mathcal{W}_c$

我们原本希望考虑对 $T = t_1 t_2 \dots t_n$ 进行 $\dot{p}$-拆分，然后利用拆分前后距 w-自由性的距离之间的关系。然而，我们只知道在 $w \in \mathcal{W}_c$ 的情况下这些距离之间存在这种联系。因此，我们将采用从一般的 w 到 $w \in \mathcal{W}_c$ 的归约，正如 [29]，在证明引理 4.8 时所做的那样。不失一般性，假设 0 是一个不会出现在 w 或 T 中的符号（如果 Σ 中没有这样的符号，那么我们将 Σ 扩展为 $\Sigma \cup \{0\}$）。令 $w' = w_1 0 w_2 0 \dots w_{k-1} 0 w_k 0$、$T' = t_1 0 t_2 0 \dots t_n 0$ 和 $p' =$（$\dot{p}_1/2$、$\dot{p}_1/2$、……、$\dot{p}_n/2$、$\dot{p}_n/2$）。注意，w' 属于 $\mathcal{W}_c$。由 [29，引理 4.6]，

$$\Delta(T', w', p') = \frac{1}{2}\Delta(T, w, \dot{p}) \,. \tag{3.32}$$

这里同样定义一个由 $[2n]$ 构成的区间集合。

定义 3.11 令 $b'_0 = 0$。定义 U'，$\{b'_u\}_{u=1}^{U'}$ 以及函数 f：$[U'] \to [U]$ 使用算法 1。对于每个 $u \in [U']$ 令 $B'_u = [b'_{u-1} + 1, b'_u]$，并定义 $\mathcal{B}' = \{B'_u\}_{u=1}^{U'}$。

算法 1

输入：$U, \{b_v\}_{v=1}^{U}$，对每个 $v \in U$ 指明 $B_v = \{b_v\}_{v=}^{U}$ 是否 [1] 为重或轻。
输出：U'，$\{b'_u\}_{u=1}^{U'}$。

```
u = 1, v = 1
当 v ≤ U 时执行
    如果 B_v 是重的，那么
        b'_u = 2b_v − 1, b'_{u+1} = 2b_v
        f(u) = v, f(u + 1) = v
        v = v + 1,  u = u + 2
    else
        b'_u = 2b_v
        f(u) = v
        v = v + 1, u = u + 1
    结束 if
结束 while
U' = 最大 {f^{-1}(U)}
```

直观来说，算法 1 确保在 T中，重区间之后的 0 会在 T'中自身成为单索引区间。另一方面，其余的 0 会与 T中包含其左邻居的轻区间合并，形成一个新区间，其权重与 T中的轻区间相同。它还设置了函数 f，该函数将 T' 中的区间映射到 T中对应的区间。

对于每个 $i \in [2k]$ 和 $u \in [U']$，令

$$\xi_i'^u = \sum_{j \in [b'_u]} I_i^j(T', w') p'_j \,. \tag{3.33}$$

观察 3.10 对于每个 $u \in [U']$，以及每个满足 $2 \nmid i$（即 $w'_i \neq 0$）的 $i \in [2k]$，

$$\xi'^u_i = \frac{1}{2}\xi^{f(u)}_{\frac{i+1}{2}} , \tag{3.34}$$

而如果 $2 \mid i$（意为 $w'_i = 0$），

$$\xi'^u_i = \frac{1}{2}\begin{cases} \mathrm{wt}_{\dot{p}}\left([b_{f(u)}]\right) & \text{if } B_{f(u)} \text{is light} \\ \mathrm{wt}_{\dot{p}}\left([b_{f(u)}]\right) & \text{if } B_{f(u)} \text{is heavy and } T'[b'_u] = 0 \\ \mathrm{wt}_{\dot{p}}\left([b_{f(u)-1}]\right) & \text{if } B_{f(u)} \text{is heavy and } T'[b'_u] \neq 0 . \end{cases} \tag{3.35}$$

3.6 通过拆分实现均匀分布

回顾一下，η 和 ζ 在第 3.4 节开头处定义。令 $T = t'^{\alpha_1}_1 \dots t'^{\alpha_{2n}}_{2n}$，其中 $\alpha_j = \left\lceil \frac{p_j}{\eta} \right\rceil$ 对每个 $j \in [2n]$ 都成立。将分布 $\widetilde{p}$ 定义为 $\widetilde{p}_j = \frac{1}{2}\zeta\eta$，对每个 $j \in [|\widetilde{T}|]$ 都成立，因此 $\widetilde{p}$ 是均匀分布（回顾一下，ζ 和 η 在第 3.4 节中定义）。由于 $p'_j = \frac{1}{2}\zeta\eta \cdot \alpha_j = \sum_{\tilde{j} \in \phi^{-1}(j)} \widetilde{p}_{\tilde{j}}$，对每个 $j \in [2n]$，我们得到（T，$\widetilde{p}$）是（T'，p'）的一个拆分（回顾定义 3.8）。

我们再次运用 [29, 定理 4.4]，，该定理表明拆分保持与 w-自由性的距离，以证明

$$\Delta(\widetilde{T}, w', \widetilde{p}) = \Delta(T', w', p') . \tag{3.36}$$

记 $\widetilde{n} = |\widetilde{T}| = \frac{2}{\zeta\eta}$。接下来我们定义 $[\widetilde{n}]$ 的一个区间集合。

定义 3.12 令 $\widetilde{b}_0 = 0$，并且对于每个 $u \in [U']$，令 $\widetilde{b}_u = \max\{h \in [\widetilde{n}] : \phi(h) = b'_u\}$。对于每个 $u \in [U']$，令 $\widetilde{B}_u = [\widetilde{b}_{u-1} + 1, \widetilde{b}_u]$，并定义 $\widetilde{\mathcal{B}} = \left\{\widetilde{B}_u\right\}_{u=1}^{\widetilde{b}_{U'}}$。

对于每个 $i \in [2k]$ 和 $u \in [U']$，令

$$\widetilde{\xi}^u_i = \sum_{j \in [\widetilde{b}_u]} I^j_i(\widetilde{T}, w')\widetilde{p}_j , \tag{3.37}$$

并注意到

$$\widetilde{\xi}^u_i = \frac{1}{\widetilde{n}} N^{\widetilde{b}_u}_i(\widetilde{T}, w') . \tag{3.38}$$

命题的证明　下一个断言的证明与断言 3.7 的证明几乎完全相同，因此省略 .

命题 3.11 对于每个 $i \in [2k]$ 和 $u \in [U']$，

$$\widetilde{\xi}^u_i = \xi'^u_i .$$

对于本小节的最后一个断言，回顾一下，事件 E_1 是在定义 3.5 中定义的。

命题 3.12 在给定事件 E_1 的条件下，对于每个 $u \in [U']$，如果 $\mathcal{B}_{f(u)}$ 是轻的，则 $\mathrm{wt}_{\widetilde{p}}(\mathcal{B}_u) < \frac{6}{z} + \frac{c_\eta}{z}$。

证明：考虑任意满足 $\mathcal{B}_{f(u)}$ 是轻的 $u \in [U']$。在以 E_1 为条件时，E_1 在定义 3.5 中定义，断言 3.2 的结论成立，因此 $\mathrm{wt}_p(\mathcal{B}_{f(u)}) < \frac{6}{z}$。此外，容易验证，若 $\mathcal{B}_{f(u)}$ 是轻的，则 $\left|\mathrm{wt}_p(\mathcal{B}_{f(u)}) - \mathrm{wt}_{p'}(\mathcal{B}'_u)\right| \leq L_1(p, \dot{p})$。由于 $\mathrm{wt}_{p'}(\mathcal{B}'_u) = \mathrm{wt}_{\widetilde{p}}(\widetilde{\mathcal{B}}_u)$ 且 $L_1(p, \dot{p}) \leq \frac{c_\eta}{z}$，结论得证。 ■

3.7 无分布情形的估计量

对于每个 $i \in [2k]$ 和 $u \in [U']$，令 $x(u,i)$ 取以下值

$x(u,i) = 1$ 若 $2 \nmid i$,$x(u,i) = 2$ 若 $2 \mid i$ 且 $B_{f(u)}$ 是轻的，$x(u,i) = 2$（也）若 $2 \mid i$ 且 $B_{f(u)}$ 是重的，并且 $T'[b'_u] = 0$,$x(u,i) = 3$ 若 $2 \mid i$ 且 $B_{f(u)}$ 是重的，并且 $T'[b'_u] \neq 0$。

定义如下估计量。对于每个 $i \in [2k]$ 和 $u \in [U']$

$$\widehat{\xi}_i^u = \frac{1}{2} \begin{cases} \breve{\xi}_{\frac{i+1}{2}}^{f(u)} & \text{if } x(u,i) = 1 \\ \mathrm{wt}_{S_2}\left([b_{f(u)}]\right) & \text{if } x(u,i) = 2 \\ \mathrm{wt}_{S_2}\left([b_{f(u-1)}]\right) & \text{if } x(u,i) = 3\,. \end{cases} \tag{3.39}$$

对于下一个断言，请回忆事件 E_2 是在定义 3.6中定义的。

引理 3.13 在事件 E_2的条件下，对于每个 $i \in [2k]$ 和 $u \in [U']$

$$\left|\widehat{\xi}_i^u - \widetilde{\xi}_i^u\right| \leq \frac{c_\eta}{z} + \frac{1}{2z}\,. \tag{3.40}$$

证明： 利用三角不等式以及断言 3.9、观察结果 3.10 和断言 3.11，我们得到，对于每个 $i \in [2k]$ 和 $u \in [U']$

$$\begin{aligned} \left|\widehat{\xi}_i^u - \widetilde{\xi}_i^u\right| \quad &\leq \quad \frac{1}{2} \sum_{r \in B_{f(u)}} |\dot{p}_j - p_j| \\ &+ \quad \frac{1}{2} \begin{cases} \left|\breve{\xi}_{\frac{i+1}{2}}^{f(u)} - \xi_{\frac{i+1}{2}}^{f(u)}\right| & \text{if } x(u,i) = 1 \\ \left|\mathrm{wt}_{S_2}\left([b_{f(u)}]\right) - \mathrm{wt}_p\left([b_{f(u)}]\right)\right| & \text{if } x(u,i) = 2 \\ \left|\mathrm{wt}_{S_2}\left([b_{f(u-1)}]\right) - \mathrm{wt}_p\left([b_{f(u-1)}]\right)\right| & \text{if } x(u,i) = 3\,. \end{cases} \end{aligned} \tag{3.41}$$

使用等式 （3.25）且由于我们已对 E_2 进行条件限定，因此得到所需的不等式。 ■

~我们证明另一个断言，以建立 $\Delta(T, w', \widetilde{p})$ 与 $\Delta(T, w, p)$ 之间的联系。

引理 3.14

$$|2\Delta(\widetilde{T}, w', \widetilde{p}) - \Delta(T, w, p)| \leq L_1(p, \dot{p})\,. \tag{3.42}$$

证明： 该断言由合并以下等式得到：（3.26）、（3.32）和（3.36 ）. ■

3.8 一般情形的收尾

我们现在可以重新表述 并证明本节的主要定理（如引言中所述） .

定理 **1.2** 存在一种用于子序列自由性的基于样本的分布无关距离近似算法。该算法从服从未知分布 p 的 T 中获取大小为 $\Theta\left(\frac{k^2}{\delta^2} \cdot \log\left(\frac{k}{\delta}\right)\right)$ 的样本，并输出一个估计值 Δ，使得 $|\Delta - \Delta(T, w, p)| \leq \delta$ 的概率至少为 2 3。

与特殊情形 $w \in \mathcal{W}_c$ 以及量化分布 p一样，该算法的运行时间与样本大小呈线性关系。定理 1.2 的证明与引理 3.8的证明类似，但两者有几个重要区别。为求完整起见，下面给出详细证明。

证明：该算法首先抽取一个大小为 $s_1 = 120z\log(240z)$ 的样本 S_1，并按照定义 3.4 构造区间集合 $\mathcal{B}$。接着，算法根据另一个大小为 $s_2 = z^2\log(40kU)$ 的样本 S_2，按照定义 3.2 为每个 $u \in U$计算向量 $\mathrm{wt}_{S_2}([b_u])$，并定义一个大小为 $k \times U$ 的矩阵 $\check{\xi}$，如下所示。对于每个 $i \in [k]$ 和 $u \in [U]$，令 $\check{\xi}[i][u] = \check{\xi}_i^u$，其中 $\check{\xi}_i^u$ 如方程 (3.12) 所定义。然后，算法使用算法 1 定义 $w' = w_1 0 w_2 0 \dots w_k 0$ 和 $\mathcal{B}'$，后者是定义 3.11 中所定义的 U' 个区间的集合；据此，它还得到函数 $f\colon [U'] \to [U]$。随后，该算法定义一个大小为 $2k \times U'$ 的矩阵 ξ，如下所示。对于每个 $i \in [2k]$ 和 $u \in [U']$，令

$$\hat{\xi}[i][u] = \hat{\xi}_i^u,$$

其中 $\hat{\xi}_i^u$ 如方程 (3.39) 所定义。最后，算法输出 $\Delta = 2M(\xi)$，其中 M 如定义 2.2 所定义。

我们希望应用命题 3.6 来证明 $|\Delta - \Delta(T, w, p)| \le \delta$ 以至少 23 的概率成立。根据 s_1 的设定，应用命题 3.1 可知，事件 E_1（定义见定义 3.5）以至少 $8\,10$ 的概率成立。根据 s_2 的设定，应用命题 3.3 可知，事件 E_2（定义见定义 3.6）以至少 $9\ 10$ 的概率成立。此后我们假设这两个事件都成立（它们同时成立的概率至少为 $7/10$）。

为应用断言 3.6，我们设 $\widetilde{w} = w'$，$J = \left\{\widetilde{b}_0,\ \widetilde{b}_1,\ \dots\dots,\ \widetilde{b}_{U'}\right\}$（回忆定义 3.12），并令 $\widehat{\mathcal{N}} = \widetilde{n}\widehat{\xi}$，其中 $\widehat{\xi}$ 如上所定义。此外，我们设 $c_1 = \frac{1}{8}$ 和 $c_2 = \frac{1}{8}$。接下来我们证明，该断言前提中的所有条件均已满足。

为证明条目 1 得到满足，我们首先注意到，对每个 $u \in [U']$，以下结论成立。由于 $\widetilde{p}$ 是 $[\widetilde{n}]$ 上的均匀分布，$\frac{b_u - b_{u-1}}{\widetilde{n}} = \mathrm{wt}_{\widetilde{p}}(\widetilde{\mathcal{B}}_u)$。因此，若 $\frac{b_u - b_{u-1}}{\widetilde{n}} \ge \frac{25}{4}\frac{1}{z}$，则 $\mathrm{wt}_{\widetilde{p}}(\widetilde{\mathcal{B}}_u) \ge \frac{25}{4}\frac{1}{z}$；根据命题 3.12（回顾一下，我们是在 E_1 的条件下进行讨论），这意味着 $\mathcal{B}_{f(u)}$ 是重的。这又意味着 $\mathcal{B}'_u$ 只包含一个索引，因此 $T[b_{u-1}+1] = \dots = T[b_u]$。根据 z 的定义（公式（3.1））以及 c_1 的设定，可知该条目得到满足。

为证明条目 2 得到满足，我们使用命题 3.13，它表明对每个 $\left|\widehat{\xi}_i^u - \widetilde{\xi}_i^u\right| \le \frac{c_\eta}{z} + \frac{1}{2z}$ 和 $i \in [2k]$ 以及 $u \in [U']$，根据 c_2 的设定以及公式（3.38）和z 与 c_η 的定义（后者在第 3.4 节开头定义），可知该条目得到满足。

应用断言 3.6 后，我们得到 $|\Delta - 2\Delta(T, w')| \le 2(c_1\delta + 2c_2\delta)$；根据 c_1 和 c_2 的设定，该值至多为 $\frac{3\delta}{4}$。由于 $\widetilde{p}$是均匀分布，因此 $\Delta(\widetilde{T}, w') = \Delta(\widetilde{T}, w', \widetilde{p})$。利用断言 3.14 和公式（3.25），我们得到 $|2\Delta(\widetilde{T}, w', \widetilde{p}) - \Delta(T, w, p)| \le 2\frac{c_\eta}{z}$；根据 z和c_η 的定义，该值至多为 $\frac{\delta}{4}$，故断言得证。 ■

4 距离近似的下界

本节给出从文本 T的 w-无子序列性进行距离近似所需样本数量的下界。当底层分布为均匀分布时，该下界成立。

定理 4.1 设 k_d 为 w 中不同符号的数量。在均匀分布下，任何针对 w-自由性的距离近似算法都必须抽取大小为 $\Omega(\frac{1}{k_d\delta^2})$ 的样本，条件是

$on\ \delta \le \frac{1}{300k_d}\ and\ n > \max\left\{\frac{8k}{\delta}, \frac{200}{k_d\delta^2}\right\}$.

注意，如果 $\delta \ge 1/k_d$，则该算法可以直接输出 0。这是因为， T 中 w 的互不相交副本数量至多等于 T 中 w 里出现次数最少的符号的出现次数。该数量的上界为 $\frac{n}{k_d}$，因此与 w-自由性的距离至多为 $\frac{1}{k_d}$。在这种情况下无需抽样，因此只有平凡的下界成立。证明将处理 $\delta \in (0, \frac{1}{300k_d}]$ 的情形。

证明：该证明基于区分无偏硬币与带有微小偏差的硬币的难度。具体细节如下。

令 $V = \{v_1, \ldots, v_{k_d}\}$ 为 w 中不同符号组成的集合，并令 0 为不属于 V 的一个符号。我们定义文本上的两个分布 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 如下。对于每个 $\tau \in [\frac{n}{k_d}]$ 和 $\rho \in [0,1]$，令 λ_ρ^τ 为一个随机变量，其以 ρ 的概率取值为 0，以 $1-\rho$ 的概率取值为 v_1。令 $\delta' = 3k_d\delta$，并考虑文本上的以下两个分布

$$\mathcal{T}_1 = \left[\lambda^1_{\frac{1}{2}}, v_2, v_3, \ldots, v_{k_d}, \lambda^2_{\frac{1}{2}}, v_2, v_3, \ldots, v_{k_d}, \ldots\ldots, \lambda^{n/k_d}_{\frac{1}{2}}, v_2, v_3, \ldots, v_{k_d}\right], \tag{4.1}$$

$$\mathcal{T}_2 = \left[\lambda^1_{\frac{1}{2}+\delta'}, v_2, v_3, \ldots, v_{k_d}, \lambda^2_{\frac{1}{2}+\delta'}, v_2, v_3, \ldots, v_{k_d}, \ldots\ldots, \lambda^{n/k_d}_{\frac{1}{2}+\delta'}, v_2, v_3, \ldots, v_{k_d}\right]. \tag{4.2}$$

也就是说，这两个分布的支持集都包含由 n/k_d 个大小均为 k_d 的块组成的文本。对于 $i \in \{2, \ldots, k_d\}$，每个块中的第 i 个符号为 v_i。这两个分布的区别仅在于每个块中第一个符号的选取方式。在 $\mathcal{T}_1$ 中，它以 $1/2$ 的概率为 0，以 $1/2$ 的概率为 v_1；而在 $\mathcal{T}_2$ 中，它以 $1/2+\delta' = 1/2+3\delta k_d$ 的概率为 0，以 $1/2-\delta'$ 的概率为 v_1。

对于 $b \in \{1,2\}$，考虑根据 T_b 选择一个文本 $\mathcal{T}_b$ （记作 $T_b \sim \mathcal{T}_b$），并令 O_b为文本中 v_1 的出现次数（因此 O_b 是一个随机变量）。注意到$\mathbb{E}[O_1] = \frac{n}{2k_d}$ 且 $\mathbb{E}[O_2] = \frac{n}{2k_d} - 3\delta n$。应用加性切尔诺夫界（定理A.1），并利用定理关于 n的前提，

$$\Pr_{T_1\sim\mathcal{T}_1}\left[O_1 < \mathbb{E}[O_1] - \delta n/8\right] \le \exp(-2(k_d\delta/8)^2 \cdot n/k_d) \le \frac{1}{100}, \tag{4.3}$$

and

$$\Pr_{T_2\sim\mathcal{T}_2}\left[O_2 < \mathbb{E}[O_2] + \delta n/8\right] \le \exp(-2(k_d\delta/8)^2 \cdot n/k_d) \le \frac{1}{100}. \tag{4.4}$$

对于 $b \in \{1,2\}$，令 $R_b = R(T_b, w)$（回忆一下， $R(T_b, w)$表示T_b中 w 的不相交副本数，并注意 R_b 是一个随机变量）。注意到 $R_1 \ge O_1 - k + 1$，且 $R_2 \le O_2$。

因此，根据公式（4.3），若根据 T_1 选择 $\mathcal{T}_1$ 并利用前提$n > \frac{8k}{\delta}$，则R（T_1， w）$\ge \frac{n}{2k_d} - \frac{1}{8}\delta n - k + 1 \ge \frac{n}{2k_d} - \frac{2}{8}\delta n$的概率至少为 99/100；根据公式（4.4），若根据 T_2 选择 $\mathcal{T}_2$，则 R（T_2， w）$\le \frac{n}{2k_d} - 3\delta n + \frac{1}{8}\delta n = \frac{n}{2k_d} - \frac{23}{8}\delta n$ 的概率至少为 99/100。

反设结论不成立，即存在一种用于子序列自由性的基于样本的距离近似算法，其所需样本量为 $Q(k_d, \delta) = 1/(ck_d\delta^2)$，其中 c 是某个充分大的常数；该算法输出到 w-自由性的距离估计值，加性误差至多为δ，且这一结果以至少 $2/3$ 的概率成立。考虑在 $T_1 \sim \mathcal{T}_1$ 或 $T_2 \sim \mathcal{T}_2$上运行该算法。令L 表示样本落在形如 $j = \ell \cdot k_d + 1$ （其中 ℓ 为整数）的索引上的次数。根据马尔可夫不等式，$L > 10 \cdot Q(k_d, \delta)/k_d = 10/(ck_d^2\delta^2)$ 的概率至多为 $1/10$。

根据上述结论，若我们在

$$T_1 \sim \mathcal{T}_1 \qquad 2/3 - 1/100 -$$

上运行该 $1/10$

$$\widehat{\Delta} \geq \frac{n}{2k_d} - \frac{10}{8} \qquad L \leq 10/\ ck_d^2\delta^2$$

算法，则算法以至少该概率输出一个估计值，且（）。类似地，若我们在 $T_2 \sim \mathcal{T}_2$上运行该算法，则算法以至少 $2/3 - 1/100 - 1/10$ 的概率输出一个估计值 $\widehat{\Delta} \leq \frac{n}{2k_d} - \frac{15}{8}$，且 $L \leq 10/(ck_d^2\delta^2)$。（两种情况下，概率均取自对$T_b \sim \mathcal{T}_b$的选择、算法获得的样本，以及算法可能使用的额外内部随机性。）根据 $\mathcal{T}_1$ 和 $\mathcal{T}_2$的定义，这意味着可以区分无偏硬币和偏置为 $3k_d\delta$ 的硬币，且成功概率至少为 $2/3-1/100-1/10 > \frac{8}{15}$，所用样本量为 $\frac{1}{c'k_d^2\delta^2}$，这与 Bar-Yosef[2, 定理 8]（取$m = 2$、 $\epsilon = 3k_d\delta$时的结论相矛盾。由于我们有 $\delta < \frac{1}{300k_d}$，因此 $\epsilon < \frac{1}{96}$，这符合所引定理的要求）。 ■

参考文献

[1] Nir Ailon、Bernard Chazelle、Seshadhri Comandur 和 Ding Liu。估计到单调函数的距离。《随机结构与算法》，31(3):371–383，2007年。[2] Ziv Bar-Yossef。通过信息论得到的采样下界。载于第35届年度ACM计算理论研讨会论文集，第335–344页，2003年。[3] Omri Ben-Eliezer、Eldar Fischer、Amit Levi 和 Ron D. Rothblum。具有（极）短 PCPP 的困难性质及其应用。载于第11届理论计算机科学创新会议（ITCS）论文集，第9:1–9:27页，2020年。[4] Piotr Berman、Meiram Murzabulatov 和 Sofya Raskhodnikova。图像性质的容错测试器。《ACM算法汇刊》，18(4):1–39，2022年。文章编号37。[5] Piotr Berman、Sofya Raskhodnikova 和 Grigory Yaroslavtsev。Lp测试。载于第46届年度ACM计算理论研讨会论文集，第164–173页，2014年。[6] Hadley Black、Deeparnab Chakrabarty 和 C. Seshadhri。单调性测试的定义域约简：针对 $o(d)$ 维布尔函数的测试器。载于第31届年度ACM-SIAM离散算法研讨会论文集，第1975–1994页，2020年。[7] Eric Blais、Cl´ement L Canonne、Talya Eden、Amit Levi 和 Dana Ron。容错junta测试及其与次模优化和函数同构的联系。《ACM计算理论汇刊》，11(4):1–33，2019年。[8] Eric Blais、Renato Ferreira Pinto Jr. 和 Nathaniel Harms。VC维数与分布无关的基于样本的测试。载于第53届年度ACM计算理论研讨会论文集，第504–517页，2021年。[9] Avrim Blum 和 Lunjia Hu。主动容错测试。载于第31届计算学习理论会议（COLT）论文集，第474–497页，2018年。[10] Mark Braverman、Subhash Khot、Guy Kindler 和 Dor Minzer。通过超立方体嵌入改进单调性测试器。载于第13届理论计算机科学创新会议（ITCS）论文集，第25:1–25:24页，2024年。

[11] Andrea Campagna、Alan Guo 和 Ronitt Rubinfeld。用于连通性和直径的局部重构器与容错测试器。载于第17届随机化与计算国际研讨会论文集，第411–424页，2013年。[12] Cl´ement L Canonne、Elena Grigorescu、Siyao Guo、Akash Kumar 和 Karl Wimmer。测试k-单调性：布尔函数的兴衰。《计算理论》，15(1):1–55，2019年。本文曾发表于2017年ITCS会议论文集。[13] Omer Cohen Sidon。子序列自由性的基于样本的距离近似，2023年。硕士论文，特拉维夫大学。[14] Ilias Diakonikolas 和 Daniel Kane。一种测试离散分布性质的新方法。载于第56届年度IEEE计算机科学基础研讨会论文集，第685–694页，2016年。[15] Shahar Fattal 和 Dana Ron。高维中到单调性的距离近似。《ACM算法汇刊》，6(3):1–37，2010年。[16] Nimrod Fiat 和 Dana Ron。图性质的高效距离近似。《ACM-SIAM离散算法研讨会》第32届年会论文集，第1618–1637页，2021年。[17] Eldar Fischer 和 Lance Fortnow。布尔性质的容错测试与非容错测试。《计算理论》，2:173–183，2006年。[18] Eldar Fischer 和 Ilan Newman。图性质的测试与估计。《SIAM计算杂志》，37(2):482–501，2007年。[19] Oded Goldreich、Shafi Goldwasser 和 Dana Ron。性质测试及其与学习和近似的联系。《ACM杂志》，45:653–750，1998年。[20] Venkat Guruswami 和 Atri Rudra。容错局部可测试码。载于第9届随机化与计算国际研讨会论文集，第306–317页，2005年。[21] Nathaniel Harms 和 Yuichi Yoshida。乘积分布下测试与学习的下采样，2022年。[22] Carlos Hoppen、Yoshiharu Kohayakawa、Richard Lang、Hanno Lefmann 和 Henrique Stagni。估计到遗传图性质的距离。《离散数学电子札记》，61:607–613，2017年。[23] Swastik Kopparty 和 Shubhangi Saraf。容错线性测试与局部可测试码。载于第13届随机化与计算国际研讨会论文集，第601–614页，2009年。[24] Amit Levi 和 Erik Waingarten。通过对图进行拒绝抽样，得到容错junta测试和单调性测试的下界。载于第10届理论计算机科学创新会议（ITCS）论文集，第52:1–52:20页，2019年。[25] Sharon Marko 和 Dana Ron。有界度图和一般稀疏图中的距离近似。《算法汇刊》，5(2)，2009年。文章编号22。

[26] Ilan Newman 和 Nithin Varma。《LIS 估计的新次线性算法与下界》。载于《自动机、语言与编程：第 48 届国际研讨会》，第 100:1–100:20 页，2021 年。

[27] Ramesh Krishnan S Pallavoor、Sofya Raskhodnikova 和 Erik Waingarten。布尔函数单调性的距离近似。随机结构与算法，60(2):233– 260，2022.

[28] Michal Parnas、Dana Ron 和 Ronitt Rubinfeld。容错性质测试与距离近似。《计算机与系统科学杂志》，72(6):1012–1042，2006年。

[29] Dana Ron 和 Asaf Rosin。单边错误下子序列无关性的最优分布无关基于样本测试。ACM计算理论汇刊，14(4):1–31，2022。该工作的扩展摘要发表于 SODA 2021 论文集。

[30] Ronitt Rubinfeld 和 Madhu Sudan。具有应用于程序测试的多项式的鲁棒刻画。SIAM计算杂志，25(2)：252–271，1996。

A Chernoff 界

定理 A.1 设 χ_1，...，χ_m 为 m 独立随机变量，其中 $\chi_i \in [0,1]$ 对每个$1 \le i \le m$均成立。令 $p \overset{\text{定义}}{=} \frac{1}{m}\sum_i \mathbb{E}[\chi_i]$。那么，对于每个 $\gamma \in (0,1],$，以下界成立：

- （加法形式）

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i > p+\gamma\right] < \exp\left(-2\gamma^2 m\right) \tag{A.1}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i < p-\gamma\right] < \exp\left(-2\gamma^2 m\right) \tag{A.2}$$

- （乘法形式）

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i > (1+\gamma)p\right] < \exp\left(-\gamma^2 pm/3\right) \tag{A.3}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\chi_i < (1-\gamma)p\right] < \exp\left(-\gamma^2 pm/2\right) \tag{A.4}$$

B 缺失的证明

断言 B.1 作为断言证明的一部分所描述的贪心算法 *2.1* 会找到一个由以下对象的角色不相交副本构成的最大规模集合 w 位于 T中。

证明：我们首先引入有序角色不相交副本的概念。根据 [29,定义 3.8], 若两个角色不相交副本 $C = (i_1, \ldots, i_k)$ 和 $C' = (i'_1, \ldots, i'_k)$ 属于 w 在 T 中，并且 C' 对每个 $j \in [k]$ 都成立，则它们是有序的，且 C' 后接于 C。由角色不相交副本组成的序列 $(C_1, \ldots, C_m)$

w 在 T 中是一个有序角色不相交副本序列，当且仅当对每个 $r \in [m-1]$，都有 C_{r+1}后接于 C_r。

根据[29, 断言 3。5], 对于 w 中的 T的每一组角色不相交副本，都存在 w 中 T 的一组等大小的有序角色不相交副本。由于断言 2.1 的证明中描述的贪心算法在 w 中 T找到了 w 中 T的一组有序角色不相交副本序列，接下来只需证明不存在其他更长（更大）的 w 中的 T有序角色不相交副本序列。

将贪心算法在 T中找到的 w有序角色不相交副本序列记为 $\mathcal{C} = (C_1, \ldots, C_{|\mathcal{C}|})$。反设与断言相反，存在一个更长的序列，$\sim$ $\sim$ $\sim\mathcal{C} = (C_1, \ldots, C_{|\widetilde{\mathcal{C}}|})$，它由 T中的 w有序角色不相交副本组成。接下来，我们通过对 m 和 i 进行归纳，证明对于每一对 $(m, i) \in [|\mathcal{C}|] \times [k]$，都有 $C_m[i] \leq C_m[i]$，这将与计数器假设矛盾。

对于每个 $m \in [|\mathcal{C}|]$ 和 $i = 1$，根据贪心算法的定义，$C_m[1]$ 是 T 中 w_1 第 m 次出现的索引 $\sim$ $\sim$ $\sim$。由于 $\mathcal{C}$ 有序，因此 $C_1[1] < C_2[1] < \cdots < C_m[1]$，我们有 $\sim$ $\sim$，即 $C_m[1]$ 是 T 中 w_1 第 $m' \geq m$ 次出现的索引。因此，对于每个 $m \in [|\mathcal{C}|]$，都有 $C_m[1] \leq C_m[1]$。

为了证明 $i > 1$ 时 (m, i) 的结论，我们用归纳法假设该结论对 $\sim(m, i-1)$ 以及 $(m-1, i)$ 成立。为便于论证（从而使 C_{m-1} 和 C_{m-1} 对 $m = 1$ 也有定义），我们定义 $C_0[i] = C_0[i] = -k+i$ $C_m[i-1] \leq C_m[i-1]$。根据归纳假设，且 $C_{m-1}[i] < C_{m-1}[i]$。由于副本的索引始终严格递增，$C_m[i-1] \leq C_m[i]$、$\sim$ $\sim$ $\sim$ $\sim$；又由于 $\mathcal{C}$ 有序，$C_{m-1}[i] < C_m[i]$。因此，$C_m[i-1] < C_m[i]$ 且 $C_{m-1}[i] < C_m[i]$。根据算法的定义，$C_m[i]$ 是大于 $C_{m-1}[i]$ 的、随后出现的 $\sim$ w_i $C_m[i-1]$ 的首次出现的索引。由于 $T[C_m[i]] = w_i$，我们得到 $C_m[i] \leq C_m[i]$，证毕。

最后，根据计数器假设，这意味着对于每个 $m \in [|C|]$，都有 $\sim C_m[i] < C_{|C|+1}[i]$，并且 $i \in [k]$。但这与算法在经过 $C_{|\mathcal{C}|}$ 后仍未找到任何角色不相交副本这一事实相矛盾。■

命题 **2.4** 的证明：令 $s = \log(6k \cdot \ell) \frac{2\gamma^2}{}$。我们从 $[n]$ 中均匀选取 s 个样本，彼此独立地随机抽取（允许重复抽样）。将第 q 个抽样索引记为 ρ_q。对于每个 $i \in [k]$、$r \in [\ell]$ 和 $q \in [s]$，定义随机变量 $\chi^{i,r}_q$：当且仅当 $\rho_q \in [j_r]$ 和 $T[\rho_q] = w_i$ 时取值为 1；否则 $\chi^{i,r}_q = 0$。

对于每个 $i \in [k]$ 和 $r \in [\ell]$，令

$$\widehat{\mathcal{N}}_i^r = \frac{n}{s} \sum_{q=1}^{s} \chi_q^{i,r} , \tag{B.1}$$

并注意到

$$\mathbb{E}\left[\chi_q^{i,j}\right] = \frac{N_i^{j_r}(T, w)}{n} . \tag{B.2}$$

根据加性切尔诺夫界（参见定理 A.1）以及 s的设定，我们得到

$$\begin{aligned} \Pr\left[\left|\widehat{\mathcal{N}}_i^r - N_i^{j_r}(T, w)\right| > \gamma n\right] &= \Pr\left[\left|\frac{n}{s}\sum_{q=1}^{s}\chi_q^{i,r} - N_i^{j_r}(T, w)\right| > \gamma n\right] \\ &= \Pr\left[\left|\frac{1}{s}\sum_{q=1}^{s}\chi_q^{i,r} - \frac{N_i^{j_r}(T, w)}{n}\right| > \gamma\right] \\ &< 2\exp\left(-2\gamma^2 s\right) = \frac{1}{3k \cdot \ell} . \end{aligned} \tag{B.3}$$

对所有配对 $(i,r) \in [k] \times [\ell]$ 应用并合界，我们得到：对每个 $i \in [k]$ 和 $r \in [\ell]$，其概率至少为2 3

$$\left|\widehat{\mathcal{N}}_i^r - N_i^{j_r}(T,w)\right| \le \gamma n \ , \tag{B.4}$$

如所要求。 ■

~命题 3.4 的证明：为简便起见，我们分别用 T 和 w 代替 T 和 $\widetilde{w}$。回顾一下，$M(\mathcal{N}) = M_k^\ell(\mathcal{N})$，且 $R(T,w) = R_k^{j_\ell}(T,w)$。我们将证明，对于每个 $i \in [k]$以及每个 $r \in [\ell]$，都有 $\left|M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)\right| \le (i-1) \cdot \max_{\tau \in [r]\setminus J'} \{j_\tau - j_{\tau-1}\}$。我们通过对 i进行归纳法来证明这一点。

对于 $i = 1$ 以及每个 $r \in [\ell]$:

$$\begin{aligned} \left|M_1^r(\mathcal{N}) - R_1^{j_r}(T,w)\right| &= \left|N_1^{j_r}(T,w) - N_1^{j_r}(T,w)\right| \\ &= 0 \le (1-1) \cdot \max_{\tau \in [1]\setminus J'} \{j_\tau - j_{\tau-1}\} \ , \end{aligned} \tag{B.5}$$

其中，第一个等式由 $\mathcal{N}$的设定以及 $M_1^r(\mathcal{N})$和$R_1^{j_r}(T,w)$的定义得出。归纳步骤中，我们假设该命题对 $i-1 \ge 1$ （以及每个 $r \in [\ell]$）成立，并证明它对i也成立。我们有，

$$\begin{aligned} & M_i^r(\mathcal{N}) - R_i^{j_r}(T,w) \\ &= N_i^{j_r}(T,w) - \max_{b \in [r]} \left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} - R_i^{j_r}(T,w) \qquad \text{(B.6)} \\ &= \max_{j \in [j_r]} \left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\} - \max_{b \in [r]} \left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} \ , \qquad \text{(B.7)} \end{aligned}$$

其中，等式（B.6）由 $\mathcal{N}$ 的设定以及 M_i^r（$\mathcal{N}$）的定义得出，而等式（B.7）由命题 2.1 推出。记 j^* 为使第一项 max 取最大值的一个索引 $j \in [j_r]$，并令b^* 为满足 $j_{b^*} \ge j^*$ 的最小索引。我们有:

$$\begin{aligned} & \max_{j \in [j_r]} \left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\} - \max_{b \in [r]} \left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} \\ &\le N_i^{j^*}(T,w) - R_{i-1}^{j^*-1}(T,w) - N_i^{j_{b^*}}(T,w) + M_{i-1}^{b^*}(\mathcal{N}) \\ &= N_i^{j^*}(T,w) + R_{i-1}^{j_{b^*}}(T,w) - R_{i-1}^{j_{b^*}}(T,w) - R_{i-1}^{j^*-1}(T,w) \\ &\quad -N_i^{j_{b^*}}(T.w) + M_{i-1}^{b^*}(\mathcal{N}) \\ &\le \left(M_{i-1}^{b^*}(\mathcal{N}) - R_{i-1}^{j_{b^*}}(T,w)\right) + \left(N_i^{j^*}(T,w) - N_i^{j_{b^*}}(T,w)\right) \\ &\quad + \left(R_{i-1}^{j_{b^*}}(T,w) - R_{i-1}^{j^*-1}(T,w)\right) \\ &\le (i-2) \max_{\tau \in [r]\setminus J'} \{j_\tau - j_{\tau-1}\} \\ &\quad + \begin{cases} 0 & \text{, if } T[j_a^*] = \cdots = T[j_b^*]] \\ \max_{\tau \in [r]\setminus J'} \{j_\tau - j_{\tau-1}\} & \text{otherwise} \end{cases} \\ &\le (i-1) \max_{\tau \in [r]\setminus J'} \{j_\tau - j_{\tau-1}\} \ , \end{aligned} \tag{B.8}$$

在第三个不等式中，我们使用了归纳假设，以及以下事实：如果没有$T[j_{b^*}] = \cdots = T[j^*]$，那么 $\left(N_i^{j^*}(T,w) - N_i^{j_{b^*}}(T,w)\right) + \left(R_{i-1}^{j_{b^*}}(T,w) - R_{i-1}^{j^*-1}(T,w)\right) \leq (j_{b^*} - j^* + 1) \leq \max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\}$。

结合公式 (B.7) 和 (B.8)，我们得到

$$M_i^r(\mathcal{N}) - R_i^{j_r}(T,w) \leq (i-1)\max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\}\ . \tag{B.9}$$

与公式 (B.7) 类似：

$$R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) = \max_{b\in[r]}\left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} - \max_{j\in[j_r]}\left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\}\ . \tag{B.10}$$

令 b^{**} 为使第一个 max 项取最大值的索引 $b\in[r]$。我们有：

$$\begin{aligned}
&\max_{b\in[r]}\left\{N_i^{j_b}(T,w) - M_{i-1}^b(\mathcal{N})\right\} - \max_{j\in[j_r]}\left\{N_i^j(T,w) - R_{i-1}^{j-1}(T,w)\right\}\\
\leq\ & N_i^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N}) - N_i^{j_{b^{**}}}(T,w) + R_{i-1}^{j_{b^{**}}-1}(T,w)\\
\leq\ & R_{i-1}^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N})\\
\leq\ & \left|R_{i-1}^{j_{b^{**}}}(T,w) - M_{i-1}^{b^{**}}(\mathcal{N})\right|\\
\leq\ & (i-2)\max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\} \leq (i-1)\max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\}\ .
\end{aligned} \tag{B.11}$$

因此（结合公式 (B.10) 和 (B.11)），[9]

$$R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \leq (i-1)\max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\} \tag{B.12}$$

结合公式 (B.9) 和 (B.12)，我们得到

$$\left|M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)\right| \leq (i-1)\max_{\tau\in[r]\setminus J'}\{j_\tau - j_{\tau-1}\}\ , \tag{B.13}$$

证明至此完成。 ■

[9]实际上，$M_i^r(\mathcal{N}) \geq R_i^{j_r}(T,w)$ 成立，因此 $R_i^{j_r}(T,w) - M_i^r(\mathcal{N}) \leq 0$ 也成立；但为了使归纳论证更简单，我们对 $R_i^{j_r}(T,w) - M_i^r(\mathcal{N})$ 证明与 $M_i^r(\mathcal{N}) - R_i^{j_r}(T,w)$ 相同的上界。