arXiv:2302.03380v2 [cs.LG] 4 May 2023

相关数据库检测中的相变

多尔·埃利梅莱赫* 瓦西姆·胡莱赫尔†

2023年5月5日

摘要

我们研究检测两个高斯数据库 $\mathbf{X} \in \mathbb{R}^{n\times d}$ 和 $\mathbf{Y}^{n\times d}$之间相关性的问题。每个数据库都由n 个用户和d 个特征组成。该问题与社交媒体、计算生物学等领域的分析相关。我们将其表述为一个假设检验问题：在零假设下，这两个数据库在统计上相互独立；而在备择假设下，存在一个未知的置换 σ，作用于n 个用户（即行置换），使得 $\mathbf{X}$ 与 $\mathbf{Y}^{\sigma}$的一个置换版本具有 ρ相关性。我们确定了最优检验发生相变时的尖锐阈值，具体取决于n 和d所处的渐近情形。具体而言，我们证明，若 $\rho^2 d \to 0$，当$d \to \infty$时，则无论n取何值，弱检测（表现略优于随机猜测）在统计上都是不可能的。这与一种简单检验的表现相符，该检验对 $\mathbf{X}^T\mathbf{Y}$ 的所有元素之和进行阈值判定。此外，当d 固定时，我们证明对于任何 $\rho < \rho^\star$，强检测（错误概率趋于零）都是不可能的，其中 $\rho^\star$ 是d的显式函数；而只要 $\rho^2 d \to 0$，弱检测同样不可能。这些结果弥补了近期相关研究中的重要空白。

1 引言

数据库对齐以及量化不同数据库之间的关系，是现代统计学应用中最基础的任务之一。在许多情况下，观测到的数据库具有高维、无标签、含噪且被打乱等特点，使得推断任务颇具挑战性。举例来说，两个数据库之间的这类推断任务可以表述为一个假设检验问题：在零假设下，数据库在统计上不相关；而在备择假设下，

*D. Elimelech 任职于以色列贝尔谢巴 84105 的本-古里安大学电气与计算机工程学院（电子邮箱：doreli@post.bgu.ac.il）。D. Elimelech 的研究工作得到以色列科学基金会资助（资助编号：1058/18）。

†W. Huleihel 任职于以色列特拉维夫 6997801 的特拉维夫大学电气工程与系统系（电子邮箱：wasimh@tauex.tau.ac.il）。W. Huleihel 的研究工作得到以色列科学基金会资助（资助编号：1734/21）。

存在一种置换，可以打乱其中一个数据库，并使两个数据库相关。那么，观测到这两个数据库后，在什么条件下我们能够判断它们是否相关?

事实证明，上述一般推断问题与许多领域相关，例如计算生物学 [SXB08、KHP12], 社会网络分析 [NS08、NS09],计算机视觉 [BBM05，CSS06], 以及数据匿名化/隐私保护系统。一个具体的经典例子是：考虑两个数据源（例如 Netflix 和 IMDb），它们分别提供一组实体（例如用户）的特征列表。这些特征可能是用户的各种属性，例如姓名、用户标识和评分。在许多情况下，由于隐私问题，特征的可靠标签要么不可用，要么已被删除（以隐藏敏感的唯一身份识别信息）。一般而言，这使得人们无法直接识别两个数据源中对应于同一用户的特征对。尽管如此，我们仍希望，如果两个数据库之间的相关性足够大，就有可能识别出这两个数据库之间的对应关系，并在特征列表之间建立对齐 [NS08、NS09]。

不久前，[CMK18、DCK19]提出并研究了数据对齐问题。这是一个看似简单的概率模型，刻画了上述情形。具体而言，有两个数据库 $\mathbf{X} \in \mathbb{R}^{n\times d}$ 和 $\mathbf{Y}^{n\times d}$，每个数据库都由n 名用户组成，每名用户具有d个特征。存在一个未知的置换／对应关系，将 $\mathbf{X}$ 中的用户与 $\mathbf{Y}$ 中的用户匹配起来。对于一对匹配的数据库条目，其特征依据已知分布而相互依赖；对于未匹配的条目，其特征则相互独立。目标是恢复未知置换，并推导统计保证，确定在何种相关程度、n和d条件下可以或不可能恢复。粗略而言，对于一大类概率分布，这个恢复问题已经得到了充分理解。例如，在高斯情形下，将相关系数记为 ρ， [DCK19] 证明了，当 $\rho^2 = 1 - o(n^{-4/d})$ 时，可以实现完美恢复；而当 $\rho^2 = 1 - \omega(n^{-4/d})$时则不可能，其中 n 、$d \to \infty$。

上述恢复问题的检测对应问题也在[KN22a、KN22b]中得到了研究。如上所述，核心问题是：给定两个数据库，我们能否判断它们是否相关? 研究表明，若 $\rho^2 d \to \infty$，则可以通过对 $\mathbf{X}^T\mathbf{Y}$ 的元素之和进行阈值判定，实现高效检测（且错误概率趋于零）。另一方面，研究还表明，若 $\rho^2 d\sqrt{n} \to 0$ 且$d = \Omega(\log n)$，那么从信息论角度而言，检测是不可能的（即存在下界）；同样，当n、$d \to \infty$时也是如此。遗憾的是，这两个界之间显然存在较大差距。尤其值得注意的是，前述上界完全不依赖n，这意味着n在检测问题中并不起显著作用；而下界则强烈依赖n 。这正是本文的主要目标：我们希望根据 n 和 d 紧致刻画检测边界。

乍看之下，人们可能会怀疑，这一差距的根源在于所提出的算法。事实上请注意，在备择分布下，潜在置换表示一种隐藏的对应关系，数据库正是在这种对应关系下相关联的。因此，“和阈值化”方法忽略了这种隐藏的组合结构，因而看起来并非最优。然而，事实证明，在$d \to \infty$且独立于n的情形下，这种简单方法实际上出奇地最优：我们证明，只要 $\rho^2 d \to 0$，弱检测（表现略优于随机猜测）在信息论上便

	弱检测		强检测	
渐近行为	**可能**	**不可能**	**可能**	**不可能**
$n,\ d \to \infty$	$\Omega(d^{-1})^*$	$o(d^{-1})$	$\omega(d^{-1})^*$	$(1-\varepsilon)d^{-1}$
$d \to \infty, n$ 常数	$\Omega(d^{-1})^*$	$o(d^{-1})$	$\omega(d^{-1})^*$	$O(d^{-1})$
$n \to \infty, d$ 常数	$\rho^2 \geq \frac{60\log 2}{d}^*$	$o(1)$	$1-o(n^{-\frac{2}{d-1}})$	$\rho^\star(d)$

表 1: 我们关于 ρ^2 在不同渐近情形下弱检测和强检测界限的总结。标记为 $*$ 的界限由 [KN22b] 的上界推出。

不可能；而如果 $\rho^2 \lessapprox 1/d$（也就是说，对于某个$\varepsilon > 0$，$\rho^2 \leq (1-\varepsilon)/d$），那么强检测（错误概率趋于零）在信息论上不可能。然而，当d固定时，这种情况会发生变化。在这种情况下，我们证明，对于任意 $\rho < \rho^\star$，强检测都不可能，其中 $\rho^\star$ 是d的一个显式函数；而当 $\rho = 1 - o(n^{-2/(d-1)})$时，则可能。后者可通过统计超过某个阈值的经验成对相关性的数量来实现。最后，我们证明，当 $\rho^2 d \to 0$时，弱检测（表现略优于随机猜测）不可能；而对于任意 $\rho > \rho^{\star\star}$则可能，其中 $\rho^{\star\star}$ 同样是一个显式函数。以往工作的界限以及上述新结果总结于表 1。

现在简要介绍其他相关工作。在 [DCK20] 中，研究了数据库对齐置换的部分恢复问题。在 [SGE19] 中，基于典型性的框架建立了匹配成功恢复的充要条件。此外，[BE21] 和 [BE22] 分别提出并研究了特征删除和特征重复情况下的置换恢复问题。最近，[Tam 22] 分析了高斯数据库的联合相关性检测与对齐问题。最后，我们指出，数据库对齐和检测问题与多种植入匹配问题密切相关，特别是图对齐问题；这些领域已有许多令人振奋且有趣的研究成果，也发展出有用的数学技巧，参见例如 MMX21、DWXY21、PG11、DMWX18、WXY20、MWXY21、WXY22、Gan20]，以及其中的许多参考文献。

记号。对于任意 $n \in \mathbb{N}$，整数集合 $\{1, 2, \ldots, n\}$ 记作 $[n]$。令 S_n 表示 $[n]$上的所有置换的集合。对于给定的置换 $\sigma \in S_n$，令 σ_i 表示 σ 将 $i \in [n]$映射到的值。我们用 log 表示自然对数函数（即以 e 为底）。随机向量用大写字母表示，例如 X，其转置记为 X^T。由 n 个随机向量组成的集合记作 $X = (X_1, \ldots, X_n)$。记号 $(X_1, \ldots, X_n) \sim P_X^{\otimes n}$表示随机向量 $(X_1, \ldots, X_n)$ 按照 P_X独立同分布（i.i.d.）。我们用 $\mathcal{N}(\eta, \Sigma)$表示均值向量为 η、协方差矩阵为 Σ的多元正态分布。令 Poisson(λ)

表示参数为 λ的泊松分布。$n \times n$ 单位矩阵记为 $I_{n\times n}$，而 0_d 表示全零的 d 维列向量。令 $\mathcal{L}(Y)$表示随机变量 Y 的分布律，即概率分布。对于概率测度 $\mathbb{P}$ 和 $\mathbb{Q}$，令 $d_{TV}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int |d\mathbb{P} - d\mathbb{Q}|$ 表示全变差距离。对于空间 Ω上的概率测度 μ，我们用 $\mu^{\otimes d}$ 表示 μ 在乘积空间 Ω^d上的乘积测度（共 d 个）。对于测度 $\nu \ll \mu$（即相对于 μ绝对连续的测度），我们将相对于 μ的拉东-尼科迪姆导数 ν（为简洁起见）记作 $\frac{\nu}{\mu}$。对于函数 f、g: $\mathbb{N} \to \mathbb{R}$，若存在 $c > 0$使得对所有 n 都有 $f(n) \le cg(n)$（且 $f(n) \ge cg(n)$），则称 $f = O(g)$（且 $f = \Omega(g)$）。若 $\lim_{n\to\infty} f(n)/g(n) = 0$，则称 $f = o(g)$；若 $g = o(f)$，则称 $f = \omega(g)$。

2 问题表述与主要结果

概率模型。考虑以下二元假设检验问题。在零假设 $\mathcal{H}_0$ 下，高斯数据库 X 和 Y 独立生成，其中$X_1, \ldots, X_n$，$Y_1, \ldots, Y_n \sim \mathcal{N}(0_d, I_{d\times d})$。令 $\mathbb{P}_0$ 表示由此得到的关于(X, Y)的分布。在备择假设 $\mathcal{H}_1$ 下，数据库 X 和 Y 通过置换σ 相关，其中 $\sigma \in \mathbb{S}_n$ 未知，而相关系数 $\rho \in [-1,1] \setminus \{0\}$已知。具体而言，

$$\begin{aligned} \mathcal{H}_0 &: (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) \\ \mathcal{H}_1 &: (X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\\rho & 1\end{bmatrix}\right), \end{aligned} \tag{1}$$

对于某个置换 $\sigma \in \mathbb{S}_n$。对于固定的 $\sigma \in \mathbb{S}_n$，我们将假设 $\mathcal{H}_1$ 下 (X, Y) 的联合分布测度记为 $\mathbb{P}_{\mathcal{H}_1|\sigma}$。图 1 直观展示了我们的概率模型。

学习问题。我们的问题的检验函数是一个函数 $\phi: \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \{0,1\}$，旨在判定发生的是哪个假设：$\mathcal{H}_0$ 还是 $\mathcal{H}_1$。检验ϕ 的风险定义为其第一类错误概率与（最坏情况下的）第二类错误概率之和，即：

$$R(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(X, Y) = 1] + \max_{\sigma\in\mathbb{S}_n} \mathbb{P}_{\mathcal{H}_1|\sigma}[\phi(X, Y) = 0]. \tag{2}$$

我们假设检测问题的极小极大风险为

$$R^\star \triangleq \inf_{\phi:\mathbb{R}^{n\times d}\times\mathbb{R}^{n\times d}\to\{0,1\}} R(\phi). \tag{3}$$

我们指出，R 是 ρ、d 和 n 的函数；但为了便于阅读，记号中省略了这些变量。

我们研究在多个渐近情形下该检测问题的可行性与不可能性。这些情形由参数序列$(\rho, d, n) = (\rho_k, d_k, n_k)_{k\in\mathbb{N}}$ 刻画。我们考虑 d_k 和 n_k 有界或趋于无穷大的情形。因此，每当我们使用渐近记号，例如，

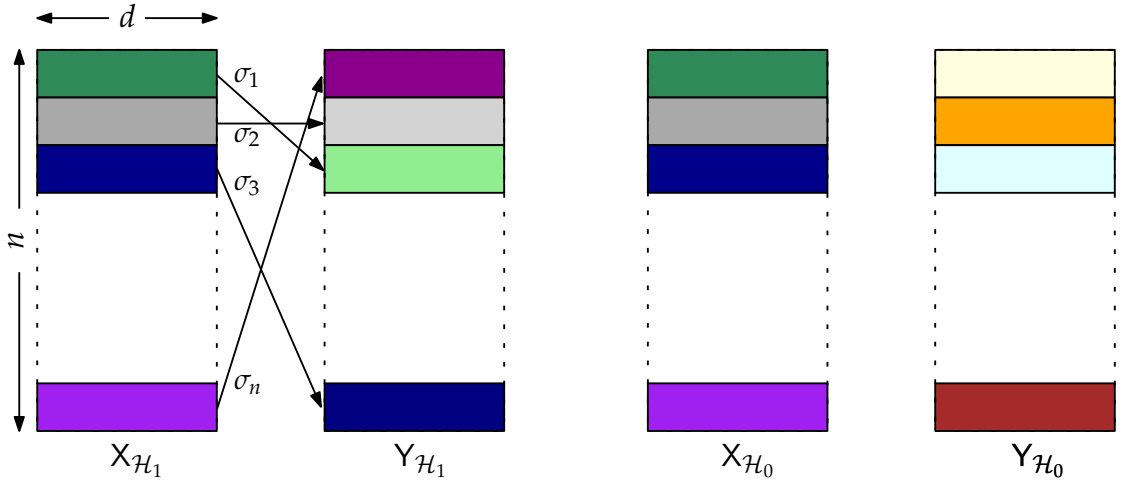


图1： 概率模型示意图。在左侧是处于假设 $\mathcal{H}_1$ 下的数据库 X 和 Y，其中相关向量以相似的颜色标记。右侧是在零假设下的不相关数据库。

$\rho^2 = o(\cdot)$、$\rho^2 = \Omega(\cdot)$等，都应结合上述序列来理解。例如，条件 $\rho^2 = o(d^{-1})$ 意味着序列(ρ,d,n) 满足 $\rho_k^2 d_k \to 0$，当$k \to \infty$时。

定义1。 若一个序列 $(\rho,d,n) = (\rho_k,d_k,n_k)_k$ 满足以下条件，则称其为：

1. 对于强检测是可容许的，如果 $\lim_{k\to\infty} \mathsf{R}^\star = 0$。
2. 对于弱检测是可容许的，如果 $\operatorname{limsup}_{k\to\infty} \mathsf{R}^\star < 1$

.

显然，强检测的可容许性意味着弱检测的可容许性。

显然，强检测的可容许性指的是存在能够以趋近于 1 的概率正确检测的算法；而弱检测则意味着存在渐近地优于随机猜测所发生假设的算法。排除弱检测/强检测可能性的一种有效方法，是考虑放宽后的平均情形问题，其中置换服从均匀分布，而不是任意选取。在这种情况下，风险函数由零假设分布 $\mathbb{P}_{\mathcal{H}_0}$ 与备择假设下的分布 $\mathbb{P}_{\mathcal{H}_1}$ 之间的全变差距离刻画。特别地，可以证明，

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) = o(1) \implies \lim_{k\to\infty} \mathsf{R}^\star = 1, \tag{4}$$

and

$$d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) \le 1 - \Omega(1) \implies \liminf_{k\to\infty} \mathsf{R}^\star > 0, \tag{5}$$

分别对应于弱检测和强检测的不可能性。我们还将在第 4 节讨论式（4）和式（5）中给出的关系。

主要结果。 本节将介绍我们关于不同渐近区域中弱检测和强检测的可容许性阈值及不可能性阈值的结果：(a) n 和 d 均趋于无穷，(b) n 为常数且 d 趋于无穷，以及 (c) d 为常数且 n 趋于无穷。我们先给出上界。如引言中所述，[KN22b]提出了以下简单检验：

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{I}\left\{\mathsf{sign}(\rho)\sum_{i,j=1}^{n} X_i^T Y_j > |\rho|\frac{dn}{2}\right\}. \tag{6}$$

我们有如下结果。

定理1. *[KN22b，定理 1]* 考虑 (1) 中的检测问题。则，

$$\mathsf{R}(\phi_{\mathsf{sum}}) \le 2\cdot\exp\left(-\frac{d\rho^2}{60}\right). \tag{7}$$

定理1的含义是，如果 $\rho^2 = \omega(\mathrm{d}^{-1})$，那么 ϕ_{sum} 能够实现强检测。此外，如果 $\rho^2 > {}^{60}\log 2\ \mathrm{d}$，那么 ϕ_{sum} 能够实现弱检测。注意，(7)中的上界完全独立于 n。我们观察到，当 d 为常数时，界 (7) 永远无法保证使用 ϕ_{sum}实现强检测，即使在 $\rho^2 = 1$ 的平凡情形下也是如此（此时可以实现零风险检测）。需要强调的是，上述现象是内在的；可以证明，与 $\mathrm{R}(\phi_{\mathsf{sum}})$ 相关的边界$\rho^2 = \omega(\mathrm{d}^{-1})$无法改进，这并非用于建立 (7) 的界定技术所导致的假象。

为了在 d 为常数（且 ρ^2 小于 1）的情形下实现强检测，我们提出以下备择检测算法。对于 $i \in [n]$，，定义 $\bar{X}_i$、$X_i/\|X_i\|_2$ 以及 $\bar{Y}_i$、$Y_i/\|Y_i\|_2$。然后，定义检验

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{I}\left\{\sum_{i,j=1}^{n}\mathbb{I}\left\{\mathsf{sign}(\rho)\bar{X}_i^T\bar{Y}_j \ge |\rho|\right\} \ge \frac{1}{2}n\mathcal{P}_{d,\rho}\right\}, \tag{8}$$

其中 $\mathcal{P}_{d,\rho} \triangleq \mathbb{P}_\rho\left(\bar{X}_1^T\bar{Y}_1 \ge \rho\right)$，并且

$$\mathbb{P}_\rho \triangleq \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right). \tag{9}$$

以下定理表明，只要 ρ 趋近于 1 的速度足够快，强检测便可能实现。

定理 2. 考虑 (1) 中的检测问题，并固定 $2 \le d \in \mathbb{N}$。那么，$\mathsf{R}(\phi_{\mathsf{count}}) \to 0$，当 $n \to \infty$ 时，如果 $\rho^2 = 1 - o(n^{-\frac{2}{d-1}})$。

在我们的工作之前，对于 d 固定且独立于n的情况，强检测（错误概率趋于零）是否能够实现尚不明确。定理首次证明，在满足非平凡条件时，可能实现强检测

（即，$\rho^2 < 1$）。我们还指出，我们的分析仅在$d \geq 2$时成立。另一方面，$d = 1$的情况仍然是个谜。事实证明，当$d = 1$ 时，许多其他备择假设检验也会失效。我们怀疑，只要$d = 1$，这种现象可能源于概率结构本身，而非算法造成的偶然结果。我们将这个引人深思的问题留待未来研究。

注 1（恢复与检测）。如引言中所述，置换 σ 的恢复问题曾在 *[DCK19]* 中得到研究。研究表明，当 d 为常数时，若 $\rho^2 = 1 - o(n^{-\frac{4}{d}})$，则可通过最大似然估计量实现恢复；而若 $\rho^2 = 1 - \omega(n^{-\frac{4}{d}})$，则不可能恢复。因此，上述定理 2 表明，即使 d 固定，检测问题在统计上也比恢复问题容易。事实上，将最大似然估计量记为 $\hat{\sigma}^{\mathsf{ML}}$，考虑以下检验

$$\phi_{\mathsf{max}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{I}\left\{\mathsf{sign}(\rho)\sum_{i=1}^{n} X_i^T Y_{\hat{\sigma}_i^{\mathsf{ML}}} > |\rho|\frac{dn}{2}\right\}. \tag{10}$$

文献 *[KN22b]* 声称，作为 *[DCK19]*, 的推论，该检验在相同的恢复保证下实现强检测，即当 $\rho^2 = 1 - o(n^{-\frac{4}{d}})$ 时。然而，我们怀疑这一说法并不准确，因为作者忽略了对第一类错误概率的分析。此外，值得一提的是，*[Tam22]* 提出并分析了一个与 (8) 中类似的检验。研究表明，当 $\frac{\rho^2}{4} = 1 - o(n^{-\frac{4}{d}})$ 时，检测是可能的。然而，这一条件显然毫无意义，因为 $|\rho| \leq 1$。最后，我们还指出，在其他情形下，恢复问题的阈值与检测问题相比有显著不同的表现；例如，若 $d = \omega(\log n)$，则恢复阈值为 $\log n\, d$，即当 $\rho^2 \gg \log n\, d$ ($\rho^2 \ll \log n\, d$) 时，恢复是可能的（不可能的）。另一方面，对于检测，阈值为 $\frac{1}{d}$，与 n 无关。

在定理 3中，我们给出下界，确定弱检测不可能的阈值。

定理 3（弱检测不可能性）。只要 $\rho^2 = o(d^{-1})$，弱检测就不可能。也就是说，对于满足 $\rho^2 = o(d^{-1})$ 的序列 $(\rho,d,n) = (\rho_k,d_k,n_k)_k$：

- 如果 d 是 k 的任意函数，且 $n \to \infty$，那么 $\lim_{k\to\infty} \mathsf{R}^\star = 1$。
- 如果 n 为常数且 $d \to \infty$，那么 $\lim_{k\to\infty} \mathsf{R}^\star = 1$。

也就是说，(ρ,d,n) 不满足弱检测的条件。

我们的第二个下界涉及强检测不可能性。在d 为常数的情况下。在定理 4 中，我们证明对于任意d，都存在 $\rho^\star = \rho^\star(d)$，使得对于 $\rho^2 < \rho^\star$，强检测不可能。在其余情况下，我们证明函数d^{-1} 是强检测的阈值。

定理 4（强检测不可能性）。 在以下任一情形下，序列（ρ，d，n）均不适用于强检测：

1. $d \in \mathbb{N}$ 和 $\rho \in (-1,1)$ 是常数，使得 $d < \log(\rho^2)\log(1-\rho^2)$，且 $n \to \infty$。

2. $n,d \to \infty$，且 $\rho^2 < (1-\varepsilon)d^{-1}$，其中某个固定的 $\varepsilon > 0$ 不依赖于 n 和 d。

3. $d \to \infty$，n 为常数，且 $\rho^2 = O(d^{-1})$。

我们指出，函数 $d^\star(\rho^2) = \log(\rho^2)/(1-\rho^2)$ 作为函数 $(0,1) \to (0,\infty)$ 是可逆的。记其反函数为 $\rho^\star$，则定理 4 中得到的条件等价于 $\rho^2 < \rho^\star(d)$。

3 上界

以下不失一般性地假设 $\rho > 0$。回顾一下，对于 $i \in [n]$，我们定义 $\bar{X}_i$、$\|X_i\|_2$ 和 $\bar{Y}_i$、$\|Y_i\|_2$。此外，令 $\mathcal{Q}_d$、ρ、$\mathbb{P}_0(\bar{X}_1^T \bar{Y}_1 \geq \rho)$，以及 $\mathcal{P}_d$、ρ、$\mathbb{P}_\rho(\bar{X}_1^T \bar{Y}_1 \geq \rho)$，其中 $\mathbb{P}_\rho$ 在式（9）中定义，且 $\mathbb{P}_0 \equiv \mathbb{P}_{\rho=0}$。考虑式（8）中的检验。我们首先对第一类错误概率进行上界估计。由马尔可夫不等式可得

$$\mathbb{P}_{\mathcal{H}_0}(\phi_{\text{count}} = 1) = \mathbb{P}_{\mathcal{H}_0}\left(\sum_{i,j=1}^n \mathbb{I}\left\{\bar{X}_i^T \bar{Y}_j \geq \rho\right\} \geq \frac{1}{2} n \mathcal{P}_{d,\rho}\right) \tag{11}$$

$$\leq \frac{2n^2 \mathcal{Q}_{d,\rho}}{n\mathcal{P}_{d,\rho}}. \tag{12}$$

另一方面，我们对第二类错误概率作如下界定。在 $\mathcal{H}_1$ 下，由于我们提出的检验对于 X 和 Y 的重新排序具有不变性，因此不失一般性地假设潜在置换为恒等置换，即 $\sigma = \mathrm{Id}$。于是，由切比雪夫不等式可得

$$\mathbb{P}_{\mathcal{H}_1}(\phi_{\text{count}} = 0) = \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i,j=1}^n \mathbb{I}\left\{\bar{X}_i^T \bar{Y}_j \geq \rho\right\} < \frac{1}{2} n \mathcal{P}_{d,\rho}\right) \tag{13}$$

$$\leq \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i=1}^n \mathbb{I}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\} < \frac{1}{2} n \mathcal{P}_{d,\rho}\right) \tag{14}$$

$$\leq \frac{4 \cdot \mathrm{Var}_\rho\left(\sum_{i=1}^n \mathbb{I}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right)}{n^2 \mathcal{P}_{d,\rho}^2}, \tag{15}$$

其中 Var_ρ 表示关于 $\mathbb{P}_\rho$ 的方差；对于随机变量 $\sum_{i=1}^n \mathbb{I}\{\bar{X}_i^T \bar{Y}_i\}$，它等于假设 $\mathcal{H}_1$ 下的方差。注意到

$$\mathrm{Var}_\rho\left(\sum_{i=1}^n \mathbb{I}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right) = \sum_{i=1}^n \mathrm{Var}_\rho\left(\mathbb{I}\left\{\bar{X}_i^T \bar{Y}_i \geq \rho\right\}\right) \tag{16}$$

$$= n\mathcal{P}_{d,\rho}(1 - \mathcal{P}_{d,\rho}), \tag{17}$$

最终得到，

$$\mathbb{P}_{\mathcal{H}_1}\left(\phi_{\mathsf{count}}=0\right) \leq \frac{4(1-\mathcal{P}_{d,\rho})}{n\mathcal{P}_{d,\rho}} \leq \frac{4}{n\mathcal{P}_{d,\rho}}. \tag{18}$$

接下来，我们推导 $\mathcal{Q}_{d,\rho}$ 和 $\mathcal{P}_{d,\rho}$的界。我们先给出 $\mathcal{P}_{d,\rho}$的下界。设$Z_1,\ldots,Z_d$ 为独立同分布的 $\mathcal{N}(\mathbf{0}, I_{d\times d})$ 随机向量，且与$X_1,\ldots,X_n$相互独立。令$Y_i' \triangleq \rho X_i + \sqrt{1-\rho^2}Z_i$。注意，在 $\mathbb{P}_\rho$下，$(\mathbf{X},\mathbf{Y})$ 与$(\mathbf{X},\mathbf{Y}')$同分布。因此，在我们的分析中，不失一般性地假设 $\mathbf{Y} = \mathbf{Y}'$。在这一假设下，我们有

$$|\bar{X}_1^T\bar{Y}_1|^2 = \frac{\rho^2\|X_1\|_2^4 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2^2 X_1^T Z_1 + (1-\rho^2)|X_1^T Z_1|^2}{\rho^2\|X_1\|_2^4 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2^2 X_1^T Z_1 + (1-\rho^2)\|X_1\|_2^2\|Z_1\|_2^2} \tag{19}$$

$$\geq \frac{\rho^2\|X_1\|_2^4 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2^2 X_1^T Z_1}{\rho^2\|X_1\|_2^4 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2^2 X_1^T Z_1 + (1-\rho^2)\|X_1\|_2^2\|Z_1\|_2^2} \tag{20}$$

$$= \frac{\rho^2\|X_1\|_2^2 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2\|Z_1\|_2\cos(\varphi_{X_1Z_1})}{\rho^2\|X_1\|_2^2 + 2\rho\sqrt{1-\rho^2}\|X_1\|_2\|Z_1\|_2\cos(\varphi_{X_1Z_1}) + (1-\rho^2)\|Z_1\|_2^2}, \tag{21}$$

这里我们用到了这样一个事实：对于$X_1,Z_1 \in \mathbb{R}^d$，内积$X_1^T Z_1$ 可以表示为$X_1^T Z_1 = \|X_1\|_2\|Z_1\|_2\cos\varphi_{X_1Z_1}$，其中 $\varphi_{X_1Z_1}$ 表示X_1 与Z_1之间的夹角。另外，注意到 $\varphi_{X_1Z_1}$ 与 $\|X_1\|_2$ 和 $\|Z_1\|_2$在统计上相互独立。因此，经过简单的代数运算可得，

$$\mathcal{P}_{d,\rho} = \mathbb{P}_\rho\left(\bar{X}_1^T\bar{Y}_1 \geq \rho\right) \tag{22}$$

$$\geq \mathbb{P}_0\left[\cos(\varphi_{X_1Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}}\left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2}\right)\right] \tag{23}$$

$$\geq \mathbb{P}_0\left[\cos(\varphi_{X_1Z_1}) \geq \frac{\rho}{2\sqrt{1-\rho^2}}\left(\frac{\|Z_1\|_2}{\|X_1\|_2} - \frac{\|X_1\|_2}{\|Z_1\|_2}\right), \|X_1\|_2 \geq \|Z_1\|_2\right] \tag{24}$$

$$\geq \mathbb{P}_0\left(\cos(\varphi_{X_1Z_1}) \geq 0, \|X_1\|_2 \geq \|Z_1\|_2\right) \tag{25}$$

$$= \mathbb{P}_0\left(\cos(\varphi_{X_1Z_1}) \geq 0\right)\cdot\mathbb{P}_0\left(\|X_1\|_2 \geq \|Z_1\|_2\right) \tag{26}$$

$$= 1/4. \tag{27}$$

现在我们推导 $\mathcal{Q}_{d,\rho}$ 的上界。对于 d 维球面上的固定向量 $y_1 \in \mathbb{S}^{d-1}$，定义 $\mathcal{B}_\rho(y_1) \triangleq \{x_1 \in \mathbb{S}^{d-1} : x_1^T y_1 \geq \rho\}$。由于在 $\mathbb{P}_0$ 下，$\bar{X}_1$ 和 $\bar{Y}_1$ 独立且均匀地分布于 $\mathbb{S}^{d-1}$ 上，我们有

$$\mathcal{Q}_{d,\rho} = \int_{\mathbb{S}^{d-1}} \frac{\mathsf{Vol}(\mathcal{B}_\rho(y_1))}{\mathsf{Vol}(\mathbb{S}^{d-1})^2}\mathrm{d}y_1, \tag{28}$$

其中，集合 $\mathcal{A} \subset \mathbb{R}^d$ 的体积定义为 $\mathsf{Vol}(\mathcal{A}) = \int_{\mathcal{A}}\mathrm{d}x$，其中积分是相对于球面上的勒贝格测度进行的。下面我们从上方界定 $\mathsf{Vol}(\mathcal{B}_\rho)$。为此，

定义集合

$$\tilde{\mathcal{B}}_\rho(y_1) = \left\{ v \in \mathbb{R}^d : \frac{v}{\|v\|} \in B_\rho(y_1) \right\} = \left\{ v \in \mathbb{R}^d : v^T y_1 \geq \rho \, \|v\| \right\}. \tag{29}$$

令$f\colon (v_1, \ldots, v_d) \to (\varphi_1, \ldots, \varphi_{d-1}, r)$ 表示球坐标变换，并令P_{d-1} 表示将 $\mathbb{R}^d$ 中的元素投影到其前$d-1$ 个坐标上。注意到，由于 $\tilde{\mathcal{B}}_\rho(y_1)$ 是由球冠 $\mathcal{B}_\rho(y_1)$定义的锥体，因此可得，

$$f(\tilde{\mathcal{B}}_\rho(y_1)) = P\big(f(B_\rho(y_1))\big) \times [0, \infty]. \tag{30}$$

最后，令 μ 表示均值为y_1、协方差矩阵为 $\Sigma = \sigma^2 I_{d\times d}$的多元随机向量的高斯测度，其中 $\sigma^2 \geq 0$。于是，

$$\begin{aligned}
1 &= \int_{x\in\mathbb{R}^d} \mu(\mathrm{d}x) && (31)\\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v-y_1\|_2^2} dv && (32)\\
&\geq \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|v-y_1\|_2^2} dv && (33)\\
&\overset{(a)}{\geq} \int_{\tilde{\mathcal{B}}_\rho(y_1)} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}(\|v\|^2 - 2\rho\|v\| + 1)} dv && (34)\\
&\overset{(b)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \int_{f(\tilde{\mathcal{B}}_\rho(y_1))} e^{-\frac{1}{2\sigma^2}(r^2-2\rho r+1)} \left(r^{d-1} \prod_{i=1}^{d-1} \sin^{d-1-i}(\varphi_i) \right) d\varphi_1 \cdots d\varphi_{d-1} dr && (35)\\
&\overset{(c)}{=} \frac{1}{(2\pi\sigma^2)^{d/2}} \underbrace{\int_{P\left(f(\mathcal{B}_\rho(y_1))\right)} \left(\prod_{i=1}^{d-1} \sin^{d-1-i}(\varphi_i) \right) d\varphi_1 \cdots d\varphi_{d-1}}_{\mathsf{Vol}(\mathcal{B}_\rho(y_1))} \cdot \int_0^\infty r^{d-1} e^{-\frac{1}{2\sigma^2}(r^2-2\rho r+1)} dr && (36)\\
&\geq \frac{1}{(2\pi\sigma^2)^{d/2}} \mathsf{Vol}(\mathcal{B}_\rho(y_1)) \int_\rho^\infty \rho^{d-1} e^{-\frac{1}{2\sigma^2}(r^2-2\rho r+1)} dr && (37)\\
&= \frac{1}{(2\pi\sigma^2)^{(d-1)/2}} \mathsf{Vol}(\mathcal{B}_\rho(y_1)) e^{-\frac{1-\rho^2}{2\sigma^2}} \rho^{d-1} \int_\rho^\infty \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(r-\rho)^2}{2\sigma^2}} dr && (38)\\
&= \frac{\rho^{d-1}}{(2\pi\sigma^2)^{(d-1)/2}} e^{-\frac{1-\rho^2}{2\sigma^2}} \mathsf{Vol}(\mathcal{B}_\rho(y_1)) \frac{1}{2}, && (39)
\end{aligned}$$

其中 (a) 由 $\tilde{\mathcal{B}}_\rho(y_1)$ 的定义可得，(b) 由变量替换可得，而(c) 由（30）可得。因此，对每个 $y_1 \in \mathbb{S}^{d-1}$,

$$\mathsf{Vol}(\mathcal{B}_\rho(y_1)) \leq \min_{\sigma^2\geq 0} 2e^{\frac{1-\rho^2}{2\sigma^2} + \frac{d-1}{2}\log(2\pi\sigma^2)} \rho^{1-d} \tag{40}$$

$$= 2e^{\frac{d-1}{2}\log\left(2\pi e\frac{1-\rho^2}{d-1}\right)}\rho^{1-d}. \tag{41}$$

另一方面，众所周知，

$$\mathsf{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \geq \frac{2\pi^{d/2}}{(d/2)^{d/2-1}} = \frac{4}{d}e^{\frac{d}{2}\log(2\pi/d)}, \tag{42}$$

其中我们使用了这一事实：$\Gamma(x) < x^{x-1}$，对于$x > 1$（参见，例如 [AQ97]）。结合（42）、（41）和（28），我们得到

$$\mathcal{Q}_{d,\rho} \leq f(d)e^{\frac{d-1}{2}\log(1-\rho^2)}\rho^{1-d}, \tag{43}$$

其中$f(d) \triangleq \frac{d}{2}\left(\underline{ed}\,d-1\right)^{d/2}\left(\underline{2\pi}\,ed-1\right)^{1/2}$。最后，利用（27）和（43），我们可以看到，（12）可以进一步上界为

$$\mathbb{P}_{\mathcal{H}_0}(\phi_{\mathsf{count}} = 1) \leq 8nf(d)e^{\frac{d-1}{2}\log(1-\rho^2)}\rho^{1-d} = 8f(d)n(\rho^{-2}-1)^{\frac{d-1}{2}}, \tag{44}$$

而（18）可以上界为：

$$\mathbb{P}_{\mathcal{H}_1}(\phi_{\mathsf{count}} = 0) \leq \frac{16}{n}. \tag{45}$$

因此，对于固定的$d \geq 2$，显然有 $\mathbb{P}_{\mathcal{H}_1}(\phi_{\mathsf{count}} = 0) \to 0$，当$n \to \infty$，且有$\mathbb{P}_{\mathcal{H}_0}(\phi_{\mathsf{count}} = 1) \to 0$，如果$n(\rho^{-2}-1)^{\frac{d-1}{2}} = o(1)$。后一个结论在 $\rho^{-2} = 1 + o(n^{-2/(d-1)})$ 时成立，这意味着 $\rho^2 = 1 - o(n^{-2/(d-1)})$，如前所述。

4 下界

与许多检测问题一样，评估极小极大风险函数面临着一项重大挑战：需要针对 $\mathbb{S}_n$ 中的所有置换最大化误差，从而产生误差项。克服这一固有障碍的一个著名策略，是考虑该问题较为宽松的平均情形版本。令 π 为 $\mathbb{S}_n$ 上的均匀测度，并用 $\mathbb{P}_{\mathcal{H}_1}$ 表示对 $\mathbb{P}_{\mathcal{H}_1|\pi}$ 关于 π 求平均后得到的概率测度。对于检验 ϕ，我们考虑如下贝叶斯风险函数：

$$\bar{\mathsf{R}}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y}) = 1] + \mathbb{E}_{\sigma\sim\pi}\left[\mathbb{P}_{\mathcal{H}_1|\sigma}[\phi(\mathsf{X},\mathsf{Y}) = 0]\right], \tag{46}$$

以及我们问题的贝叶斯风险：

$$\bar{\mathsf{R}}^\star \triangleq \inf_{\phi}\bar{\mathsf{R}}(\phi). \tag{47}$$

显然，任意检验 ϕ 都满足 $\mathsf{R}(\phi) \geq \bar{\mathsf{R}}(\phi)$，因此 $\mathsf{R}^\star \geq \bar{\mathsf{R}}^\star$。由此可知，要证明定理 3，只需证明在给定假设下 $\bar{\mathsf{R}}^\star =$

$1+o(1)$。利用贝叶斯风险函数的一种广为人知的等价刻画（通过全变差距离给出），以及柯西-施瓦茨不等式，可以证明

$$\mathsf{R}^\star \geq \bar{\mathsf{R}}^\star = 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0}, \mathbb{P}_{\mathcal{H}_1}) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0\left[L^2\right] - 1}, \tag{48}$$

其中 L $\frac{\mathbb{P}_{\mathcal{H}_1}}{\mathbb{P}_{\mathcal{H}_0}}$ 为似然比，且期望是相对于 $\mathbb{P}_{\mathcal{H}_0}$取的。利用（48）中的界，只需证明在定理 3 的假设下， $\mathbb{E}_0\left[L^2\right] \leq 1+$ 为(1)的

受 [WXY20], Zeynep 和 Nazer [KN22b] 的启发，给出了$\mathbb{E}_0\left[L^2\right]$ 基于均匀随机抽取的渗透中循环分布的精确描述。为了证明定理 3，我们将仔细分析 $\mathbb{E}_0[L^2]$, , 并改进 [KN22b]中证明的界。为使本文内容完整，在展开我们的改进分析之前，我们先概述 Nazer 和 Zeynep 计算$\mathbb{E}_0\left[L^2\right]$ 的主要思路。

计算的第一步需要运用 Ingster-Suslina 方法，根据富比尼定理， $\mathbb{E}_0[L^2]$ 也可以等价地写成

$$\mathbb{E}_0[L^2] = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right], \tag{49}$$

其中，期望是针对 π 和 π' 的独立耦合取的，它们是 $\mathbb{S}_n$ 上均匀测度的两个副本。对于固定的排列 σ 和 σ'，注意到 $\mathbb{P}_{\mathcal{H}_1|\sigma}$、$\mathbb{P}_{\mathcal{H}_1|\sigma'}$ 和 $\mathbb{P}_0$ 关于 $\mathbb{R}^{2\times d\times n}$ 上的勒贝格测度绝对连续，因此有

$$\frac{\mathbb{P}_{\mathcal{H}_1|\sigma}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma}}{f_{\mathcal{H}_0}}, \quad \text{and} \quad \frac{\mathbb{P}_{\mathcal{H}_1|\sigma'}}{\mathbb{P}_{\mathcal{H}_0}} = \frac{f_{\mathcal{H}_1|\sigma'}}{f_{\mathcal{H}_0}}, \tag{50}$$

其中 $f^{\mathcal{H}_i}$ 表示 $\mathbb{P}_{\mathcal{H}_i}$ 关于勒贝格测度的拉东-尼科迪姆导数，即相应假设下的密度函数 (X, Y) 。令 $\mathcal{N}_\rho : \mathbb{R}^{2\times d} :\to \mathbb{R}_+$ 表示一对随机向量的密度函数，这对随机向量的分布为

$$\mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

注意到

$$\frac{f_{\mathcal{H}_1|\sigma}(\mathsf{X},\mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}\frac{f_{\mathcal{H}_1|\sigma'}(\mathsf{X},\mathsf{Y})}{f_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} = \prod_{i=1}^{n}\frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})}\frac{\mathcal{N}_\rho(X_i, Y_{\sigma'_i})}{\mathcal{N}_0(X_i, Y_{\sigma'_i})}. \tag{51}$$

为了继续计算，我们先指出两个关键事实。首先，零假设下的分布（$\mathcal{H}_0$）在坐标重新排序后保持不变。同样， $\mathbb{S}_n$上的均匀测度在与固定置换复合后也保持不变。因此，

$$\mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right] = \mathbb{E}_{\pi}\left[\mathbb{E}_0\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\,\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}}\right]\right]. \tag{52}$$

对于固定的 和 Id，我们考虑式（51）中给出的乘积，并将其记为 Z_σ。第二个关键事实是，Z_σ 可以分解为相互独立的项，这些项分别对应于渗透 σ的各个循环。回顾一下，置换σ 的一个循环是由 $[n]$ 中的元素组成的序列 $(i_0, i_2, \ldots\ldots, i_{|C|-1})$，满足 $\sigma(ij) = i_{j+1}$ 模 $|C|$，对所有 j均成立。若$|C| = k$，则称 C 为一个k循环。对于固定的循环C，我们记

$$Z_C \triangleq \prod_{i \in C} \frac{\mathcal{N}_\rho(X_i, Y_{\sigma_i})}{\mathcal{N}_0(X_i, Y_{\sigma_i})} \frac{\mathcal{N}_\rho(X_i, Y_i)}{\mathcal{N}_0(X_i, Y_i)}. \tag{53}$$

由于一个置换的循环集合构成了 $[n]$, 随机变量$\{Z_C\}_C$的一个划分，因此对应于 σ 所有循环的随机变量彼此独立（相对于 $\mathbb{P}_{\mathcal{H}_0}$），并且

$$Z_\sigma = \prod_C Z_C. \tag{54}$$

下面的引理指出，对于一个固定的循环 C， $\mathbb{E}_0[Z_C]$ 取决于 ρ 和 $|C|$。引理的证明基于高斯随机向量的性质。更多细节请参阅 [KN22b，引理 10]。

引理 1. 对于置换 σ 的一个固定循环 C,

$$\mathbb{E}_0[Z_C] = \frac{1}{(1-\rho^{2|C|})^d}. \tag{55}$$

对于一个固定置换 $\sigma \in \mathbb{S}_n$ 和 $k \in [n]$, 令 $N_k(\sigma)$ 表示 k-环的数量，其中置换为 σ。结合式（49）、（52）、（54）以及引理 1，我们得到

$$\mathbb{E}_0[L^2] = \mathbb{E}_\pi \left[\prod_C Z_C \right] = \mathbb{E}_\pi \left[\prod_{k=1}^n \frac{1}{(1-\rho^{2k})^{dN_k}} \right]. \tag{56}$$

通过分析式（56），Zeynep 和 Nazer 在 [KN22b, 引理 3] 中证明了 $\mathrm{E}_0[\mathrm{L}^2] \leq (1-\rho^2)^{-dn}$，其等于 $1 + \mathrm{o}(1)$，条件是 $\rho^2 = \mathrm{o}((\mathrm{nd})^{-1})$。受 [WXY20, 命题 2], 中所做计算的启发，我们利用均匀分布的随机置换中 k-环的统计性质，仔细地对式（56）进行上界估计。经过改进的分析使我们能够证明，只需假设 $\rho^2 = \mathrm{o}(\mathrm{d}^{-1})$，便有 $\mathbb{E}_0[\mathrm{L}^2] \leq 1 + \mathrm{o}(1)$。下面的命题构成了定理 3 和定理 4 中下界证明的主要论证。

命题 1。 设 N_k 为均匀分布的置换 π中 k-环的数量，且$1 \leq k \leq n$。则：

1. 对于所有 (ρ^2, d, n)，均成立：

$$\mathbb{E}_\pi \left[\prod_{k=1}^n \left(\frac{1}{1-\rho^{2k}} \right)^{dN_k} \right] \leq \exp\left(\frac{nd\rho^2}{1-\rho^2} \right). \tag{57}$$

2. 如果 n,d 中至少有一个趋于 ∞，且 $\rho^2 = o(d^{-1})$，那么：

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le 1 + o(1). \tag{58}$$

3. 若 n、d 均趋于 ∞ 且 $\rho^2 < (1-\varepsilon)d^{-1}$，其中 $\varepsilon > 0$，则

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right). \tag{59}$$

4. 若 d 和 ρ^2 为常数，满足 $d < \log(\rho^2)\log(1-\rho^2)$ 且 $n \to \infty$

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right), \tag{60}$$

其中 $c(d,\rho^2) = d(d+1)\,2(1-\rho^2)^{d+2}$。

证明此命题需要若干技术性结果。下面的引理讨论了将k-循环的联合分布近似为相互独立的泊松随机变量。

引理 2. *[AT92，定理 2]* 设 $1 \le k \le n$ 为整数，且 $Z_1 \ldots, Z_k$ 是相互独立的随机变量，并且对所有 $1 \le i \le k$，$Z_i \sim$ 服从泊松分布 (i^{-1})。那么，$N_1, \ldots, N_k$ 与 $Z_1, \ldots, Z_k$ 的分布律之间的全变差满足

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(N_1, N_2, \ldots, N_k\right), \mathcal{L}\left(Z_1, Z_2, \ldots, Z_k\right)\right) \le F\left(\frac{n}{k}\right), \tag{61}$$

其中 $F(x)$ 是一个单调递减函数，满足 $\log F(x) = -x\log x(1+o(1))$，当 $x \to \infty$ 时。

引理 3. 设 $1 \le m \le n$ 为一个整数，并设 $Z_1 \ldots, Z_m$ 是相互独立的随机变量，使得对所有 $1 \le i \le m$，$Z_i \sim$ 泊松分布 (i^{-1})。那么，

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \le \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right), \tag{62}$$

其中 $c(d,\rho^2) = d(d+1)\,2(1-\rho^2)^{d+2}$，因此，如果 $\rho^2 = o(d^{-1})$，

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \le 1 + o(1). \tag{63}$$

证明。这个引理的证明很基础，只需要泊松随机变量的矩母函数，以及对初等函数的一些线性近似。通过重新排列期望中的表达式并利用独立性，我们得到

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right]=\prod_{k=1}^{m}\mathbb{E}_\pi\left[\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \tag{64}$$

$$=\prod_{k=1}^{m}\mathbb{E}_\pi\left[\exp\left(-dZ_k\log\left(1-\rho^{2k}\right)\right)\right] \tag{65}$$

$$\overset{(a)}{=}\prod_{k=1}^{m}\exp\left(\frac{1}{k}\left(e^{-d\log(1-\rho^{2k})}-1\right)\right) \tag{66}$$

$$=\prod_{k=1}^{m}\exp\left(\frac{1}{k}\left(\frac{1}{(1-\rho^{2k})^d}-1\right)\right). \tag{67}$$

其中 (a) 后面给出了泊松分布随机变量的矩母函数定义。

现在我们从上方估计项 $(1-\rho^{2k})^{-d}-1$。对函数 $f(x)=(1-x)^{-d}$ 进行泰勒展开，可直接得到：对于 $x\in(0,1)$，有

$$\frac{1}{(1-x)^d}=\sum_{m=0}^{\infty}\binom{m+d-1}{d-1}x^m.$$

利用拉格朗日余项定理可得，对于所有 $x>(0,1$),

$$\frac{1}{(1-x)^d}=1+dx+\frac{d(d+1)}{2}\cdot\frac{1}{(1-c)^{(d+2)}}x^2 \tag{68}$$

$$\leq 1+dx+\frac{d(d+1)}{2(1-x)^{(d+2)}}x^2. \tag{69}$$

其中，c 是 $[0,x]$ 中的一个点。取 $x=\rho^{2k}$，可得对于任意 $k>0$

$$\frac{1}{(1-\rho^{2k})^d}\leq 1+d\rho^{2k}+\frac{d(d+1)}{2(1-\rho^{2k})^{(d+2)}}\rho^{4k} \tag{70}$$

$$\leq 1+d\rho^{2k}+\frac{d(d+1)}{2(1-\rho^{2})^{(d+2)}}\rho^{4k}. \tag{71}$$

在接下来的分析中，我们记

$$c(d,\rho^2)\triangleq\frac{d(d+1)}{2(1-\rho^2)^{(d+2)}}\rho^{4k}, \tag{72}$$

并且我们得到

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right]=\prod_{k=1}^{m}\mathbb{E}_\pi\left[\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \tag{73}$$

$$= \prod_{k=1}^{m} \exp\left(\frac{1}{k}\left(\frac{1}{(1-\rho^{2k})^d} - 1\right)\right) \tag{74}$$

$$\leq \prod_{k=1}^{m} \exp\left(\frac{1}{k}\left(1 + \left(d\rho^{2k} + c(d,\rho^2)\rho^{4k}\right) - 1\right)\right) \tag{75}$$

$$= \exp\left(d\sum_{k=1}^{m}\frac{\rho^{2k}}{k} + c(d,\rho^2)\sum_{k=1}^{m}\frac{\rho^{4k}}{k}\right) \tag{76}$$

$$\leq \exp\left(d\sum_{k=1}^{\infty}\frac{\rho^{2k}}{k} + c(d,\rho^2)\sum_{k=1}^{\infty}\frac{\rho^{4k}}{k}\right) \tag{77}$$

$$= \exp\left(-d\log(1-\rho^2) - c(d,\rho^2)\log(1-\rho^4)\right) \tag{78}$$

$$\overset{(a)}{\leq} \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right), \tag{79}$$

其中 (a) 由众所周知的不等式 $\log(1+x) \geq x/(1+x)$ 推出，对于 $x > -1$。在 $\rho^2 = o(d^{-1})$ 的情况下，显然 $d\rho^2 = o(1)$。此外，

$$c(d,\rho^2)\rho^4 = \frac{d(d+1)}{2(1-\rho^2)^{d+2}}\rho^4 \leq \frac{1}{2}\exp((d+2)\rho^2)d(d+1)\rho^4 = \exp(o(1))\cdot o(1) = o(1).$$

特别地，我们得到

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) \tag{80}$$

$$= \exp\left(\frac{o(1)}{1-o(1)} + \frac{o(1)}{1-o(1)}\right) = \exp(o(1)) = 1 + o(1). \tag{81}$$

□

现在我们可以证明命题 1。证明思路如下：考虑式（58）中给出的乘积的期望。当 $n \to \infty$ 时，我们证明最后 $n - \alpha\log n$ 项的乘积总是有上界 $1 + o(1)$，其中 α 取适当值。对于前 $\alpha\log n$ 项乘积的期望，利用引理 $\{N_k\}_k$ 中给出的 2 的泊松近似，以及引理 3 中给出的相应估计。另一种情况，即 n 为常数时，可用初等论证解决。

命题 *1* 的证明。我们根据 n 和 d 的渐近情形，将证明分为三部分。首先注意一个简单事实：只要 n 固定，就有 $\sum_{k=1}^{n} kN_K = n$，这意味着对于任意 $1 \leq m \leq n$，都有

$$\prod_{k=m}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \leq \prod_{k=m}^{n}\left(\frac{1}{1-\rho^{2m}}\right)^{dN_k} = \left(\frac{1}{1-\rho^{2m}}\right)^{d\sum_{k=m}^{n}N_k} \tag{82}$$

$$\leq \left(\frac{1}{1-\rho^{2m}}\right)^{dn} = \left(1 + \frac{\rho^{2m}}{1-\rho^{2m}}\right)^{dn} \leq \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right). \tag{83}$$

这立即证明了（57）。

当 n 和 d 趋于 ∞ 时： 我们假设 $n,d \to \infty$ 且 $\rho^2 = o(d^{-1})$ 或$\rho^2 < (1-\varepsilon)d^{-1}$。我们选择 $m = \lceil \log n \rceil$，得到

$$dn\rho^{2m} \le (d\rho^2) \cdot n(\rho^2)^{\log(n)-1} = (d\rho^2)n(\rho^2)^{\log\left(\frac{n}{e}\right)} = e(d\rho^2)\left(\frac{n}{e}\right)^{1+\log\rho^2}. \tag{84}$$

由于 $\rho^2 < d^{-1}$ (当 d 足够大时，这在以下特殊情形下显然成立：$\rho^2 = o(d^{-1})$)，且 $d \to \infty$，我们有 $\log\rho^2 \to -\infty$ as $n \to \infty$。将选定的 m 代入（83）可得：

$$\prod_{k=\log n}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \le \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right) = \exp(o(1)) = 1 + o(1). \tag{85}$$

对于固定的整数 m，我们考虑由以下条件给出的集合 $S_{n,m} \subseteq \mathbb{N}^m$

$$S_{n,m} = \left\{(n_1,\dots,n_m) \in \mathbb{N}^d; \sum_{k=1}^{m} n_k \le n\right\}, \tag{86}$$

以及由以下条件给出的函数 $f_{n,m}\colon \mathbb{N}^m \to [0,\infty]$

$$f_{n,m}(n_1,n_2,\dots,n_m) = \prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dn_k} \cdot \mathbb{I}^{S_{n,m}}(n_1,\dots,n_m). \tag{87}$$

注意，对于所有 $n_1,\dots,n_m \in \mathbb{N}$,

$$f_{n,m}(n_1,n_2,\dots,n_m) \le \prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2}}\right)^{dn_k} \cdot \mathbb{I}^{S_{n,m}}(n_1,\dots,n_m) \le \left(\frac{1}{1-\rho^{2}}\right)^{dn}. \tag{88}$$

令 $m = \lceil \log n \rceil$，并令 $\{Z_k\}_k$ 为相互独立的 泊松分布 (k^{-1}) 个随机变量，如引理2所述。由于 $\sum_{k=1}^{m} N_k \le n$ 以概率 1 成立，我们有

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] = \mathbb{E}_\pi\left[f_{n,m}(N_1,\dots,N_m)\right] \tag{89}$$

$$\le \mathbb{E}[f_{n,m}(Z_1,\dots,Z_m)] + d_{\mathsf{TV}}\left(\mathcal{L}\left(N_1^m\right),\mathcal{L}\left(Z_1^m\right)\right)\cdot\|f_{n,m}\|_\infty \tag{90}$$

$$\le \mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] + d_{\mathsf{TV}}\left(\mathcal{L}\left(N_1^m\right),\mathcal{L}\left(Z_1^m\right)\right)\cdot\|f_{n,m}\|_\infty \tag{91}$$

$$\overset{(a)}{\le} \mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] + F\left(\frac{n}{m}\right)\cdot\left(\frac{1}{1-\rho^{2}}\right)^{dn} \tag{92}$$

$$\overset{(b)}{\le} \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + F\left(\frac{n}{m}\right)\cdot\left(\frac{1}{1-\rho^{2}}\right)^{dn} \tag{93}$$

$$= \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + F\left(\frac{n}{\lceil\log n\rceil}\right)\cdot\left(\frac{1}{1-\rho^{2}}\right)^{dn}, \tag{94}$$

其中，(a) 由式（88）和引理2得出，(b) 由引理 3得出。根据引理 2，我们还有

$$\log\left(F\left(\frac{n}{\lceil\log n\rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn}\right) \le \log\left(F\left(\frac{n}{\log n}\right)\left(\frac{1}{1-\rho^2}\right)^{dn}\right) \tag{95}$$

$$= \log\left(F\left(\frac{n}{\log n}\right)\right) - nd\log\left(1-\rho^2\right) \tag{96}$$

$$\le -(1+o(1))\frac{n}{\log n}\log\left(\frac{n}{\log n}\right) - nd\log\left(1-\rho^2\right) \tag{97}$$

$$\overset{(a)}{\le} -(1+o(1))\frac{n}{\log n}\log\left(\frac{n}{\log n}\right) + nd\rho^2(1+o(1)) \tag{98}$$

$$= n\left(-(1+o(1))\left(1-\frac{\log\log n}{\log n}\right) + d\rho^2(1+o(1))\right) \tag{99}$$

$$= -n(1-d\rho^2+o(1)) \xrightarrow[n\to\infty]{(b)} -\infty, \tag{100}$$

其中，(a) 由函数 $\log(1-x)$ 的泰勒展开以及 $\rho^2 = o(1)$得出，而 (b) 则由假设 $\rho^2 < (1-\varepsilon)d^{-1}$ 得出。这意味着

$$F\left(\frac{n}{\lceil\log n\rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn} = o(1), \tag{101}$$

因此，

$$\mathbb{E}_\pi\left[\prod_{k=1}^{\lceil\log n\rceil}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + o(1). \tag{102}$$

结合式 (85) 和式 (102)，我们得出：

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le \mathbb{E}_\pi\left[\prod_{k=1}^{\lceil\log n\rceil}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\prod_{k=\lceil\log n\rceil}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \tag{103}$$

$$= (1+o(1))\mathbb{E}_\pi\left[\prod_{k=1}^{\lceil\log n\rceil}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \tag{104}$$

$$= (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + o(1). \tag{105}$$

我们现在已经证明了（59）。注意，进一步假设 $\rho^2 = o(d^{-1})$，根据引理 3的第二部分，我们得到

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le (1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + o(1) = 1+o(1). \tag{106}$$

我们现在已经证明了（58）。

当 n 为常数且 d 趋于 ∞ 时：我们假设 $\rho^2 = o(d^{-1})$。由于 n 为常数，因此有 $dn\rho^2 = o(1)$ 因此由（83）可得

$$E_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le \exp\left(\frac{dn\rho^2}{1-\rho^2}\right) = 1 + o(1). \tag{107}$$

当 d 为常数且 n 趋于 ∞ 时：我们还假设 ρ^2 是一个常数，使得 $d < \log_{(\rho^2)} \log_{(1-\rho^2)}$。我们重复与 $n \to \infty$ 且 $\rho^2 = o(d^{-1})$ 时相同的步骤，只做一处小改动。我们不再近似乘积中前 $m = \lceil \log n \rceil$ 项的乘积，而是取$m' = \lceil \alpha \log(n) \rceil$，其中 $\alpha = -\ ^1\log_{(\rho^2)} + \varepsilon$，其中 ε 足够小，从而

$$d\frac{\log(1-\rho^2)}{\log(\rho^2)} < \frac{1}{1-\varepsilon\log(\rho^2)}, \tag{108}$$

（根据假设 $d < d^\star(\rho^2) = \log_{(\rho^2)} \log_{(1-\rho^2)}$，这样的值存在）。重复上一部分中的相同步骤，我们得到

$$\prod_{k=m'}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \le \exp\left(\frac{dn\rho^{2m'}}{1-\rho^{2m'}}\right). \tag{109}$$

注意到

$$nd\rho^{2m'} \le nd(\rho^2)^{\alpha\log n} = ndn^{\alpha\log(\rho^2)} = dn^{1+\alpha\log(\rho^2)} = dn^{\varepsilon\log(\rho^2)} = o(1), \tag{110}$$

这意味着

$$\prod_{k=m'}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k} \le \exp\left(\frac{dn\rho^{2m'}}{1-\rho^{2m'}}\right) = 1 + o(1). \tag{111}$$

现在我们来计算前 m' 项的乘积。与第一部分类似，根据引理 2 和引理 3，

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m'}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \le \mathbb{E}_\pi\left[\prod_{k=1}^{m'}\left(\frac{1}{1-\rho^{2k}}\right)^{dZ_k}\right] + F\left(\frac{n}{m'}\right)\cdot\left(\frac{1}{1-\rho^2}\right)^{dn} \tag{112}$$

$$\le \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + F\left(\frac{n}{m'}\right)\cdot\left(\frac{1}{1-\rho^2}\right)^{dn} \tag{113}$$

$$= \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) + F\left(\frac{n}{\lceil\alpha\log n\rceil}\right)\cdot\left(\frac{1}{1-\rho^2}\right)^{dn}. \tag{114}$$

再次使用引理2，我们得到

$$\log\left(F\left(\frac{n}{\lceil\alpha\log n\rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn}\right)\leq-(1+o(1))\frac{n}{\alpha\log n}\log\left(\frac{n}{\alpha\log n}\right)-nd\log(1-\rho^2) \tag{115}$$

$$=-n\left((1+o(1))\frac{1}{\alpha}\left(1-\frac{\log(\alpha\log n)}{\log n}\right)+d\log(1-\rho^2)\right) \tag{116}$$

$$=-n(1+o(1))\left(\frac{1}{\alpha}+d\log(1-\rho^2)\right) \tag{117}$$

$$=-n(1+o(1))\left(\frac{\log(\rho^2)}{\varepsilon\log(\rho^2)-1}+d\log(1-\rho^2)\right). \tag{118}$$

由（108）可得

$$\frac{\log(\rho^2)}{\varepsilon\log(\rho^2)-1}+d\log(1-\rho^2)>0, \tag{119}$$

这意味着

$$F\left(\frac{n}{m'}\right)\cdot\left(\frac{1}{1-\rho^2}\right)^{dn}=o(1). \tag{120}$$

现在我们得出结论

$$\mathbb{E}_\pi\left[\prod_{k=1}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right]=\mathbb{E}_\pi\left[\prod_{k=1}^{m'}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\prod_{k=m'}^{n}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \tag{121}$$

$$=\mathbb{E}_\pi\left[\prod_{k=1}^{m'}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}(1+o(1))\right] \tag{122}$$

$$\leq(1+o(1))\left(\exp\left(\frac{d\rho^2}{1-\rho^2}+\frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)+o(1)\right) \tag{123}$$

$$=(1+o(1))\exp\left(\frac{d\rho^2}{1-\rho^2}+\frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right). \tag{124}$$

□

定理 3 和定理 4 的证明如下：

定理3的证明。 结合（49）、（56）和命题1，得到：

$$\mathsf{R}^\star\geq1-\sqrt{\mathbb{E}_0[L^2]-1} \tag{125}$$

$$=1-\sqrt{\mathbb{E}_\pi\left[\prod_{k=1}^{n}\frac{1}{(1-\rho^{2k})^{dN_k}}\right]-1} \tag{126}$$

$$\geq 1 - \sqrt{1 + o(1) - 1} \tag{127}$$
$$= 1 + o(1). \tag{128}$$

□

定理4的证明。 令(ρ,d,n) 为一个满足定理4的假设的序列。我们先回顾一个重要且广为人知的事实（例如，参见 [Tsy04、引理2.6和2.7]）：对于任意测度序列$(\mathbb{P}_{0,k})_k,(\mathbb{P}_{1,k})_k$,

$$\mathbb{E}_{\mathbb{P}_{0,k}}\left[\left(\frac{\mathbb{P}_{1,k}}{\mathbb{P}_{0,k}}\right)^2\right] = O(1) \implies d_{\mathsf{TV}}(\mathbb{P}_{0,k},\mathbb{P}_{1,k}) = 1 - \Omega(1).$$

因此，由（48），若 $\mathbb{E}_0[L^2] = O(1)$，则有

$$\mathsf{R}^\star \geq 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1}) = \Omega(1), \tag{129}$$

这意味着 (ρ,d,n) 不适用于强检测（因为（129）意味着 $\limsup \mathsf{R}^\star > 0$）。

事实上，由命题 1 和（56）可知：若 d 和 ρ 为常数，满足 $d < d^\star(\rho^2)$，且$n \to \infty$，则

$$\mathbb{E}_0[L^2] \leq (1 + o(1)) \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) = O(1).$$

另一方面，若 $d,n \to \infty$ 且 $\rho^2 < (1-\varepsilon)d^{-1}$，则有 $\rho^2 = o(1)$。因此，由命题 1 还可推出

$$\mathbb{E}_0[L^2] \leq (1 + o(1)) \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) \tag{130}$$
$$\leq (1 + o(1)) \exp\left((1 + o(1))\left(d\rho^2 + \frac{d(d+1)}{2(1-\rho^2)^{d+2}}\rho^4\right)\right) \tag{131}$$
$$\leq (1 + o(1)) \exp\left((1 + o(1))\left(1 + \frac{1}{2e^{-2}(1+o(1))}\right)\right) = O(1). \tag{132}$$

剩下的情形是 $d \to \infty$、n 为常数，且 $\rho^2 = O(d^{-1})$。由（57）可得，即使不对(ρ,d,n)作任何假设，也成立：

$$\mathbb{E}_0[L^2] \leq \exp\left(\frac{nd\rho^2}{1-\rho^2}\right) = O(1). \tag{133}$$

证明至此完成。 □

5 结论

本文研究了高斯相关数据库检测问题中弱检测和强检测的渐近阈值。我们的结果总结于表1中。具体而言，当d 趋于 ∞时，定理1 和定理3 证明了d^{-1} 是一个尖锐阈值。也就是说，当 $\rho^2 \ll d^{-1}$时，弱检测和强检测均不可能；而当 $\rho^2 \gg d^{-1}$时，强检测则是可能的。

参考文献

[AQ97] G Anderson 和 S-L Qiu。伽马函数的一个单调性性质。《美国数学会会刊》，125(11):3355–3362，1997。[AT92] Richard Arratia 和 Simon Tavar´e。随机排列的循环结构。《概率年刊》，第 1567–1591 页，1992 年。[BBM05] A.C. Berg、T.L. Berg 和 J. Malik。利用低失真对应关系进行形状匹配和目标识别。载于计算机视觉与模式识别会议论文集，2005 年。[BE21] Serhat Bakirtas 和 Elza Erkip。列删除条件下的数据库匹配。*2021 IEEE*国际信息论研讨会（*ISIT*），第 2720–2725 页，2021 年。[BE22] Serhat Bakirtas 和 Elza Erkip。列重复条件下的数据库匹配。*ArXiv*，abs/2202.01730，2022 年。[CMK18] Daniel Cullina、Prateek Mittal 和 Negar Kiyavash。数据库对齐的基本极限。载于*2018 IEEE*国际信息论研讨会（*ISIT*），第 651–655 页。IEEE Press，2018 年。[CSS06] Timothee Cour、Praveen Srinivasan 和 Jianbo Shi。平衡图匹配。载于第 *19* 届神经信息处理系统国际会议论文集，NIPS'06，第 313–320 页，美国马萨诸塞州剑桥，2006 年。MIT出版社。[DCK19] Osman E. Dai、Daniel Cullina 和 Negar Kiyavash。利用高斯特征进行数据库对齐。载于 Kamalika Chaudhuri 和 Masashi Sugiyama 编，《第二十二届人工智能与统计国际会议论文集》，第 89 卷《机器学习研究论文集》，第 3225–3233 页。PMLR，2019 年 4 月 16–18 日。[DCK20] 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 IEEE*国际信息论研讨会（*ISIT*），第 1230–1235 页，2020 年。[DMWX18] Jian Ding、Zongming Ma、Yihong Wu 和 Jiaming Xu。通过度分布实现高效的随机图匹配。《概率论及相关领域》，179:29–115，2018 年。[DWXY21] Jian Ding、Yihong Wu、Jiaming Xu 和 Dana Yang。植入匹配问题：尖锐阈值与无穷阶相变。*ArXiv*，abs/2103.09383，2021 年。[Gan20] Luca Ganassali。具有高斯权重的图数据库对齐的尖锐阈值。载于*MSML*，2020 年。

[KHP12] U. Kang、M. Hebert 和 S. Park。用于对应问题的快速且可扩展的近似谱图匹配。*Information Sciences*，2012。[KN22a] Zeynep K 和 Bobak Nazer。检测相关高斯数据库。载于*2022*年*IEEE*国际信息论研讨会 *(ISIT)*，第 2064–2069 页，2022。[KN22b] Zeynep K 和 Bobak Nazer。检测相关高斯数据库。*arXiv* 预印本 *arXiv:2206.12011*，2022。[MMX21] Mehrdad Moharrami、Cristopher Moore 和 Jiaming Xu。植入匹配问题：相变与精确结果。*The Annals of AppliedProbability*，31(6):2663 – 2720，2021。[MWXY21] Cheng Mao、Yihong Wu、Jiaming Xu 和 Sophie H. Yu。通过树计数高效检验网络相关性。2021。[NS08] Arvind Narayanan 和 Vitaly Shmatikov。大型稀疏数据集的鲁棒去匿名化。载于*2008*年*IEEE*安全与隐私研讨会 *(sp 2008)*，第 111–125 页，2008。[NS09] Arvind Narayanan 和 Vitaly Shmatikov。社交网络去匿名化。载于*2009*年第*30*届*IEEE*安全与隐私研讨会，第 173–187 页，2009。[PG11] Pedram Pedarsani 和 Matthias Grossglauser。匿名化网络的隐私问题。载于*Knowledge Discoveryand Data Mining*，2011。[SGE19] Farhad Shirani、Siddharth Garg 和 Elza Erkip。基于测度集中的数据库去匿名化方法。载于*2019*年*IEEE*国际信息论研讨会 *(ISIT)*，第 2748–2752 页。IEEE Press，2019。[SXB08] Rohit Singh、Jinbo Xu 和 Bonnie Berger。多个蛋白质相互作用网络的全局比对及其在功能直系同源检测中的应用。美国国家科学院院刊，105(35):12763–8，2008年9月。[Tam22] Ran Tamir。高斯数据库的联合相关性检测与对齐。2022。[Tsy04] Alexandre B Tsybakov。《非参数估计导论》，2009。网址 *https://doi. org/10.1007/b13794*。根据以下内容修订并扩展，9(10)，2004。[WXY20] Yihong Wu、Jiaming Xu 和 Sophie H Yu。无标签随机图的相关性检验。*arXiv* 预印本 *arXiv:2008.10097*，2020。[WXY22] Yihong Wu、Jiaming Xu 和 Sophie H. Yu。确定随机图匹配的尖锐重构阈值。*IEEE*信息论汇刊，68(8):5391–5417，2022。