使用高斯核对持久同调进行降维

Jean-Daniel Boissonnat [*1] 和 Kunal Dutta [†2]

[1]法国索菲亚-安提波利斯，INRIA，蔚蓝海岸大学，电子邮件：jean-daniel.boissonnat@inria.fr[2]波兰华沙大学数学、信息学与力学学院。电子邮件：K.Dutta@mimuw.edu.pl

arXiv:2301.03321v1 [cs.CG] 9 Jan 2023

摘要

Phillips、Wang 和 Zheng [SoCG2015]指出，使用高斯核计算持久同调在拓扑数据分析和机器学习领域很有用。然而，与使用欧几里得距离甚至 k-距离计算持久同调的情形不同，目前尚不清楚如何使用高斯核计算高维数据的持久同调。本文考虑 Phillips、Wang 和 Zheng 提出的高斯核距离（GKPD）的幂距离版本，并证明使用 GKPD 计算的 P 切赫滤子的持久同调近似保持不变。对于 $\mathbb{R}^D$中的数据集，在相对误差界为 $\varepsilon \in (0,1]$, 的条件下，对于包含 n 个点的数据集，我们得到的维数为 (i) $O(\varepsilon^{-2}\log^2 n)$；对于直径为 r （可缩放）的数据集，维数为 (ii) $O(D\varepsilon^{-2}\log(Dr/\varepsilon))$。

我们采用两个主要的构成要素。第一项是利用核幂距离，以顶点之间的两两广义核功率距离表示 Čech 单形平方半径的一种新分解，我们将对其进行陈述并证明。第二项是 Rahimi 和 Recht 提出的随机傅里叶特征（RFF）映射 [NeurIPS 2007], , Chen 和 Phillips 在 [ALT 2017] 中使用了该映射。

1 引言

持久同调（PH）是拓扑数据分析中从数据提取信息的主要工具之一。给定环境空间中的一个点云数据集，其思路是从点云构造一系列拓扑空间的过滤序列，并从该序列中提取拓扑信息。

需要面对两个主要问题。首先，数据点通常位于非常高维的空间中，而计算持久同调（PH）对环境维数呈指数级或更糟的依赖。因此，一旦维数增长到几十以上，PH 很快就会变得无法使用——这确实是许多应用中的情况，例如图像处理、神经生物网络和数据挖掘（例如参见 Giraud [19]）。这一现象通常被称为维数灾难。第二个主要困难来自数据通常会受到噪声和离群值的污染。事实上，尽管 PH（使用相对于距离函数的偏移集计算）对豪斯多夫噪声具有很强的鲁棒性，但不难看出，即使存在单个离群值，也可能显著影响 PH（例如参见 [10]）。

*由法国政府通过“未来投资”项目 3IA Côte d’Azur 提供支持，该项目由法国国家研究署（ANR）管理，项目编号为 ANR-19-P3IA-0002。

†获波兰国家科学中心 SONATA 项目资助，项目编号 2019/35/D/ST6/04525 *(Probabilistic tools for high-dimensional geometric inference, topological data analysis and large-scale networks)*。

超越欧几里得距离的持久同调 规避离群值问题的一种方法，是使用对离群值更具鲁棒性的距离函数，例如测度距离（DTM）以及相关的 k-距离（针对有限数据集）；该方法由 Chazal 等人近期提出 [7]。这种方法还具有兼容去噪技术的优点，例如 [8]。然而，尽管 DTM 是一个很有前景的方向，但精确实现可能会产生显著的运行时间成本。为克服这一困难，近期有人提出了 k-距离的近似方法，从而得到经过认证的（尽管相当粗略的）PH 近似（Guibas 等人 [20]; Buchet 等人 [7]）。最近，有报道称使用该方法获得了改进的近似结果 [3]。

另一种规避离群值问题的方法是使用核（Phillips et al. [32]）。核是环境空间中点对上的相似性函数。核方法是机器学习和数据分析的中流砥柱。核方法广受欢迎的一个重要原因是所谓的核技巧，其核心思想是利用映射 ϕ 将数据映射到某个（通常为无限维的）希尔伯特空间 $\mathcal{H}$，使得点对 x、$y \in \mathbb{R}^D$ 的核函数等于其映像的内积 $\langle \phi(x)\ \phi(y) \rangle_{\mathcal{H}}$。如 [32], 所述，与欧几里得距离或 k-距离相比，使用核距离计算持久同调具有一定优势，尤其适用于机器学习。这些优势包括存在 ε-coresets [23, 30]，以及核距离函数的一些性质，例如，其子水平集映射到核密度估计的上水平集（核密度估计常用于机器学习）；它关于平滑参数 σ 是 Lipschitz 连续的，而该参数可针对固定输入进行调整；当 σ 趋于无穷时，两个测度之间的核距离有一个由 2-瓦瑟斯坦距离给出的上界。高斯核距离的这些性质已在 [32], 中得到证明；该文使用核距离的一种近似幂距离形式——我们称之为高斯核幂距离（GKPD）——计算了一些数据集的持久同调，并与使用现有距离函数计算的持久同调进行了比较。近期， [37] 运用这一方法，在构造稳健持久图方面取得了进一步进展。

降维与持久同调 面对高维问题，缓解维数灾难最简单且最常用的方法之一是使用随机投影，正如著名的 Johnson-Lindenstrauss 引理 [22] （简称 JL 引理）所应用的那样。JL 引理指出，对于任意 $\varepsilon \in (0, 1)$，欧几里得空间中的任意一组 n个点，都可以嵌入到维数为 $O(\varepsilon^{-2} \log\ n)$ 的空间中，且畸变为 $(1 \pm \varepsilon)$。自 Johnson 和 Lindenstrauss 于 1984年 提出这一事实的非构造性证明以来，许多作者陆续对其进行了改进，例如 [21, 14, 1, 2, 27, 24], 解决高效构造与实现的问题，并使用支持快速乘法的稀疏随机矩阵。最近，Narayanan 和 Nelson [29], 在 [17, 26], 的基础上证明，对于给定的一组点或终端点，采用非线性映射可以实现降维，同时保留从任意终端点到环境空间中任意点的距离。

JL引理也被 Sheehy [35] 和 Lotz [25] 用于降低计算有界基数或高斯宽度点集持久同调的复杂度。Lotz 的结果还意味着，对于有界加倍维数的集合，可以根据扩散度（最大点间距离与最小点间距离之比）进行降维。不过，他们的方法只涉及通常的点集距离，因此对离群值和噪声的存在非常敏感，正如前文所述。因此，一个自然的问题是，如何调整随机投影方法，以降低使用 k距离计算持久同调的复杂度。Arya等人在 [4] 对此进行了研究，并证明在随机投影下，使用 k距离保持持久同调的界与保持成对距离的界相同。

使用核距离进行持久同调降维 在涉及核距离的计算中，正如前文所述，有许多技术可用于应对数据量过大的问题。然而，尤其是在计算持久同调时，已知能够解决环境维数过高问题的技术并不多。例如，JL引理在这种情况下并不适用，因为核距离函数涉及欧氏距离平方的高次幂。JL引理确实可以将核幂距离近似嵌入为欧几里得距离。不过，这种构造需要进行 $n \times n$ Cholesky分解，以及其他计算开销较大的操作。此外，这种嵌入会依赖输入数据点，因此不再是数据无关的。[1]

在另一方向上，Chen和Phillips [13] 利用Rahimi和Recht的随机傅里叶特征（RFF） [33], 给出的映射，可以将成对核距离保持在一个（$1 \pm \varepsilon$）倍因子以内。然而，该映射无法保持测度之间的距离，而在 [32]的框架下计算持久同调时，这一点似乎是必要的。类似地，其他不同的方法（例如 [38, 28, 5, 9]）在保持点集之间或一般测度之间的距离方面并不高效。Phillips和Tai [31] 得到的另一种嵌入，则能以较小的加性误差对点集之间的核距离给出相对近似。即使有这样一个映射（即保持点分布之间的核距离），也不清楚它是否能够保持持久同调，因为这涉及在幂距离下保持多个球的交集（例如参见 [4, 32]）。在GKPD下，一个关键问题是，与数据点相关联的权重并不只是点本身的函数，而是数据集中所有点的成对核距离的函数。这意味着，对于将数据点映射到低维空间的任意映射，都必须在新空间中重新计算这些点的权重。在使用 k距离计算持久同调时，也会出现类似问题。 [35], 将解决这一问题作为一个开放问题提出，Arya等人 [4]最近给出了答案。然而，他们的解决方案关键性地利用了降维映射的线性性质，以及他们所证明的最小包围球在 k距离下的某些性质，因此并不适用于GKPD。正如 [4], 的结论中所述，通过在欧几里得距离取较小值时用欧几里得距离近似核距离，可以得到持久同调的常数因子近似。然而，寻找GKPD的（$1+\varepsilon$）倍因子近似仍然是一个开放问题。

1.1 我们的贡献

本文证明，对于任意 $\varepsilon \in (0, 1]$, ，我们都可以将维数降至 $O\left(\varepsilon^{-2} \log\ n\right)$，同时以 $(1+\varepsilon)$ 倍因子近似一个含有 n 个点的数据集的持久同调；该持久同调是通过高斯核距离的幂距离版本计算得到的（Phillips 等人 [32] 使用了这一方法）。我们的结果类似于 Sheehy [35] 和 Lotz [25] 关于使用欧氏距离计算持久同调的结果，以及 Arya 等人 [4] 关于欧几里得 k- 距离的结果。此外，由于我们的目标维数与 Johnson-Lindenstrauss 界的目标维数同阶，而后者对于欧几里得距离而言在常数因子范围内是最优的；并且高斯核距离是欧几里得距离的 Lipschitz 函数，因此这意味着我们的目标维数也是最优的（在常数因子范围内）。

非正式地说，我们的主定理指出，给定高维空间中的一个点集，存在一种可高效计算的映射，可将这些点映射到一个低维目标空间，并以任意精度近似使用高斯核计算得到的持久同调。此外，目标空间的维数仅与所需精度参数的倒数平方成正比，并仅以对数形式依赖于（i）点的数量或（ii）数据集的直径。

[1]概率降维技术并非完全独立于数据，因为需要通过比较原始数据点与投影后的数据点来验证算法是否成功。

正式表述见第5节中的定理 13。该定理给出了一种映射，使我们能够使用 GKPD 近似计算高维空间中一个点集的持久同调，同时实际在低维空间中使用欧几里得距离进行计算。因此，它肯定地回答了 [4]结论中提出的问题。

我们的主要工具是一条新的分解定理，它表明，使用 GKPD 计算一组加权点的最小包围球的半径平方，可以表示为这些点之间成对幂距离的线性组合。我们将证明，尽管 GKPD 是非线性的，但将其提升到某个希尔伯特空间后，它具有若干良好性质，我们随后利用这些性质证明分解定理。

论文结构 本文其余部分安排如下。第2节介绍一些必要的背景和预备知识。第 3 节证明希尔伯特空间中加权点的最小包围球的一些性质。第 4 节研究使用 GKPD 构造的切赫滤子在低失真映射下的稳定性，并证明定理 12。第5节证明主定理，展示 Rahimi 和 Recht 的随机傅里叶特征映射如何结合我们新的分解结果（定理 12）给出定理 13 的证明。最后，我们在第6节提出一些看法和开放性问题。

2 背景

我们简要介绍一些证明结果所需的定义和工具。若想深入了解，参考文献 [7, 10] 将对读者大有裨益。我们也推荐感兴趣的读者阅读 [4, 32] 以获取更多信息。

2.1 持久同调

单纯复形与滤子 设 V 为一个有限集。顶点集为 V 的一个（抽象）单纯复形，是由 V 的有限子集构成的集合 K，并且若 $A \in K$ 且 $B \subseteq A$，则 $B \in K$。K 中的集合称为 K的单形。若单形 $F \in K$ 严格包含于单形 $A \in K$中，则称 A是 $A \in K$的一个面。

若单纯复形 K配有一个函数 $f\colon K \to \mathbb{R}$，且只要 σ是 τ的一个面，就有 $f(\sigma) \leq f(\tau)$，则称其为过滤单纯复形。 f在 $r \in \mathbb{R}$处的子水平集，即 $f^{-1}(-\infty, r]$，是 K的一个子复形。通过取 r的不同值，我们得到 K的一个嵌套子复形序列（称为滤子），
$\emptyset = K^0 \subseteq K^1 \subseteq \ldots \subseteq K^m = K$，其中 K^i 是取值为r_i时的子水平集。由 $\mathbb{R}^D$ 中的有限点集 P 关联得到的切赫过滤，在拓扑数据分析中发挥着重要作用。

定义 1（˘Cech 复形）。˘ $C_\alpha(P)$ ˘Cech 复形是由满足 $\mathrm{rad}(\sigma) \leq \alpha$ 的单形 $\sigma \subset P$ 构成的集合，其中 $\mathrm{rad}(\sigma)$ 是 σ 的最小外接球的半径，即:

$$\mathrm{rad}(\sigma) \leq \alpha \Leftrightarrow \exists x \in \mathbb{R}^D,\ \forall p_i \in \sigma,\ \|x - p_i\| \leq \alpha.$$

当 α 从 0 变化到 $+\infty$ 时，我们得到 P 的 ˘Cech 滤子。˘ $C_\alpha(P)$ 也可以等价地定义为闭球 $\overline{B}(p, \alpha)$ 的神经，这些闭球以 P 中的点为中心，半径为 α:

$$\check{C}_\alpha(P) = \{\sigma \subset P | \cap_{p \in \sigma} \overline{B}(p, \alpha) \neq \emptyset\}.$$

根据神经引理（例如 [18, 6]），我们知道，球的并集 $U_\alpha = \cup_{p \in P} \overline{B}(p, \alpha)$、 $p \in P$ 和 ˘ $C_\alpha(P)$ 具有相同的同伦型。此外，由于规则形状 X 的良好采样˘P 的球并集能够捕捉 X 的同伦型，计算 P 的 Čech 复形即可得到 X 的同伦型。我们还要指出，一个称为 P 的 α-复形的更简单复形（参见例如 [15]）也能捕捉 $\bigcup_{p \in P} \overline{B}(p, \alpha)$ 的同伦型。我们的结果适用于这两种复形。

持久图。持久同调是一种计算并记录过滤复形的拓扑随参数 α 从零增至无穷大时所发生变化的方法。Edelsbrunner、Letscher 和 Zomorodian [16] 给出了一种计算持久同调的算法，它以过滤单纯复形为输入，并输出由实数对 $(\alpha_{birth}, \alpha_{death})$ 构成的序列。每一对数都对应一个拓扑特征，并分别记录该特征在滤子中出现和消失时的 α 值。因此，滤子的拓扑特征可以用这一对数序列表示；该序列既可以表示为扩展平面 $\bar{\mathbb{R}}^2 = (\mathbb{R} \cup \{-\infty, \infty\})^2$，中的点（称为持久图），也可以表示为条形码序列（持久条形码）（参见例如 [15]）。分别对应于滤子 (G_α) 和 (H_α) 的一对持久图 $\mathbb{G}$ 和 $\mathbb{H}$，若对所有 α 都有 $G_{\alpha/\beta} \subseteq H_\alpha \subseteq G_{\alpha\beta}$，则称它们是乘法 β-交错的（$\beta \geq 1$）。我们将关键地依赖这样一个事实：如果两个持久图是乘法 c-交错的，且 c 接近 1，那么一个持久图可以由另一个持久图很好地近似（参见例如 [11]）。

持久稳定性定理 [12]表明，

对称滤子的持久同调与距离函数的次水平集滤子的同调相同。˘Delaunay 过滤 [12]也有同样的结果。

2.2 测度距离；基于幂距离的持久同调

拓扑数据分析中最常见的方法，是考虑由到 V中某个点的最短距离给出的距离函数，即 d_V: $\mathbb{R}^D \to \mathbb{R}_+$ is $d_V(x) = \inf_{y \in V} d(x, y)$。˘ 给定这个距离函数后，可以将 $d_V(.)$ 的 α-偏移集（即子水平集 $\{x \in \mathbb{R}^D \mid d_V(x) \leq \alpha\}$）视为球的并集，并计算这些并集的神经，从而构造切赫滤子。然而，正如引言中所述，由这种距离函数得到的持久同调对离群值十分敏感，甚至单个离群值也可能使结果发生显著变化。为了解决这个问题，Chazal 等人 [7] 提出了测度距离（测度距离）的概念。原则上，正如 Chazal 等人所证明的，可以利用测度距离函数从 P计算切赫滤子。然而在实践中，计算α-偏移集的神经需要测量空间中每个点的距离，因此需要对测度距离函数进行近似。为此，可以考虑这种距离的一种有限形式，称为 k-距离；它对应于原始点云 [7, 4]的 k-重心集合上的幂距离。一般而言，幂距离常用于近似难以处理的距离函数，以便计算持久同调。其思路是：在点 $x \in \mathbb{R}^D$ 处，将到 P 的距离的平方近似为到点 $p \in P$的易于计算的距离平方与 p的权重平方之和： $d_P(x)^2 := d'(x, p)^2 + w(p)^2$, 其中，$d'(x, p)$ 选为比 $d_P(x)$ 更简单、更易计算的距离函数； $w(p)$ 是 p的权重，设定为距离函数 $d_P(.)$ 在 p邻域内各点处的局部近似。接下来的段落将讨论如何利用幂距离计算持久同调。

给定集合 X 和距离函数 $d: X \times X \to \mathbb{R}$，若˜距离函数 $d(., .)$ 满足自反性、对称性和三角不等式，则二元组 (X, d) 是一个度量空间。设 P 是度量空间 $(\mathcal{M}, d)$ 中的加权点集合 $\widehat{p} = (p, w(p))$。在度量空间 $(\mathcal{M}, d)$ 中，两个加权点 $\widehat{p}$ 和 $\widehat{q}$ 之间的幂距离定义为

$$D(\widehat{p}, \widehat{q}) = d(p, q)^2 - w(p) - w(q).$$

˘ 因此，我们需要将 Čech 复形的定义扩展到加权点集合。

定义 2（加权切赫复形）。设 ˆ $P = \{\widehat{p}_1, ..., \widehat{p}_n\}$ 为加权点集，其中$\widehat{p}_i = (p_i, w_i) \in \mathbb{R}^D \times \mathbb{R}$。ˆ$P$ 的 α-˘Cech 复形 ˘C_α(ˆP), 是满足以下条件的所有单形 σ 的集合:

$$\exists x,\ \forall p_i \in \sigma,\ d(x, p_i)^2 \leq w_i + \alpha^2 \quad \Leftrightarrow \quad \exists x,\ \forall p_i \in \sigma,\ D(x, \widehat{p}_i) \leq \alpha^2.$$

（这里， $D(x,\widehat{p}_i)$ 表示无权重点 x （即 $w(x)=0$）与加权点 p 之间的幂距离。）换言之，ˆP 的 α-ˇCech 复形是闭球$\overline{B}(p_i,r_i^2=w_i+\alpha^2)$的神经，这些闭球以 p_i 为中心，半径平方为 $w_i+\alpha^2$ （若其为负，则$\overline{B}(p_i,r_i^2)$为虚数）。

ˇ 加权切赫滤子及其持久同调的概念由此自然得出。

ˇ 在欧几里得情形中，我们将 α-切赫复形定义为这样一类单形的集合：其最小外接球的半径至多为 α。在加权情形中，我们也可以采用类似的方法。ˆ ˆ ˆ设 $X\subseteq P$。我们将 X 的半径定义为

$$\mathrm{rad}^2(\widehat{X}) \;=\; \min_{x\in\mathbb{R}^D}\max_{\widehat{p}_i\in\widehat{X}} D(x,\widehat{p}_i), \tag{1}$$

ˆ ˆ并将 X 的加权中心，或简称中心，定义为达到该最小值的点，记作 $c(X)$，即

$$c \;=\; c(\widehat{X})=\arg\min_{x\in\mathbb{R}^D}\max_{\widehat{p}_i\in\widehat{X}} D(x,\widehat{p}_i). \tag{2}$$

2.3 核；高斯核幂距离

核 $K\colon \mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}$ 是定义在 $\mathbb{R}^D$中各点上的相似性函数，满足对于所有 $x\in\mathbb{R}^D$， $K(x,x)=1$。再生核是一大类核，具有如下性质：给定一个再生核 K，存在一个到希尔伯特空间 $\mathcal{H}_K$ 的提升映射 ϕ，使得核函数提升为 $\mathcal{H}_K$上的内积，即对于所有 $x,y\in\mathbb{R}^D$， $K(x,y)=\langle\phi(x)\ \phi(y)\rangle_{\mathcal{H}_K}$。

由希尔伯特空间 $\mathcal{H}_K$ 上的范数所诱导的自然距离函数，可利用 $\mathbb{R}^D$ 上的核定义距离，如下所示。

$$\begin{aligned}\|\phi(x)-\phi(y)\| &= \sqrt{\langle\phi(x)-\phi(y),\phi(x)-\phi(y)\rangle_{\mathcal{H}_K}}\\ &= \sqrt{\langle\phi(x),\phi(x)\rangle_{\mathcal{H}_K}+\langle\phi(y),\phi(y)\rangle_{\mathcal{H}_K}-2\langle\phi(x),\phi(y)\rangle_{\mathcal{H}_K}}\\ &= \sqrt{K(x,x)+K(y,y)-2K(x,y)} \;=\; \sqrt{2(1-K(x,y))},\end{aligned}$$

其中，最后一步是因为对所有 $x\in\mathbb{R}^D$， $K(x,x)=1$。对于特征核（再生核的一个略小子集），该距离函数是一个度量 [36]。本文主要关注的核是高斯核，其定义为

$$K(x,y)=\exp\left(-\|x-y\|^2/2\sigma^2\right).$$

对于 $x,y\in\mathbb{R}^D$，高斯核的核距离 $D_K(.,.)$ 因此为，

$$D_K(x,y)^2 \;=\; 2(1-e^{-\|x-y\|^2/2\sigma^2}). \tag{3}$$

核到测度距离 核作为 $\mathbb{R}^D$ 上点对相似性函数的定义，可以自然地扩展为 $\mathbb{R}^D$ 上测度对 μ,ν 的相似性函数。

$$\kappa(\mu,\nu):=\int_{x\in\mathbb{R}^D}\int_{y\in\mathbb{R}^D}K(x,y)d\mu(x)d\nu(y).$$

当 μ,ν 分别是 P 和 Q 上的经验测度时，我们得到

$$\kappa(\mu,\nu):=\frac{1}{|P||Q|}\sum_{p,q\in\mathbb{R}^D}K(p,q).$$

现在令 μ 为定义在 P 上的经验测度，定义为 $\mu=\frac{1}{|P|}\sum_{p\in P}\delta_p$，其中 δ_p 是 P 上的狄拉克δ测度，而 ν 是固定点 $p\in P$ 处的 δ_p。如上所述，我们可以将

点质量 δ_p 到测度 μ 的核距离视为 x 的函数，记作 $d_\mu^K(x) := D_K(\mu, x)$。在 [32], 中，Phillips、Wang 和 Zheng 使用 $d_\mu^K(.)$ 研究点集的持久同调，其中 K 是高斯核。他们证明（[32], 定理 4.1 和 4.2），只要给定阈值范围内没有 $d_\mu^K(x)$ 的临界点，利用距离函数 $d_\mu^K(x)$ 的子水平集得到的 P 的偏移集在同伦意义下等价；并且当输入发生扰动且豪斯多夫距离有界时，这些偏移集保持稳定。因此，可以利用 d_μ^K 的偏移集来估计点云 P 的拓扑性质。

高斯核幂距离 如引言和第2.2节所述，若要精确地使用 d_μ^K 计算持久同调，就需要在 $\mathbb{R}^D$ 中处处计算 d_μ^K。因此，为避免这项计算开销，Phillips、Wang 和 Zheng [32] 使用 P 中各点的权重，以幂距离 $f_\mu^K(x)$ 近似了 d_μ^K： $f_\mu^K(x)^2 := \min_{p\in P}\left(D_K^2(x,p) - w(p)\right)$，其中 $w: P \to \mathbb{R}$ 是在 p 处的高斯核权重函数，定义如下

$$w(p) := -D_K^2(\mu, p) = -\left(\frac{1}{|P|}\sum_{y\in P} D_K^2(p,y) - \frac{1}{2|P|^2}\sum_{x,y\in P} D_K^2(x,y)\right). \tag{4}$$

注 1。请注意，权重 $w(p)$ 不仅取决于点 p，还取决于整个集合 P 的成对高斯核距离平方。这一事实对任何嵌入或降维过程都提出了一项关键要求：必须在像空间中重新计算核权重函数。

因此，根据式 (4)，点 $x \in \mathbb{R}^D$ 与点集 P 之间的高斯核幂距离（GKPD）可表示为

$$\begin{aligned} f_\mu^K(x)^2 &:= \min_{p\in P}\left(D_K^2(x,p) - w(p)\right) \\ &= \min_{p\in P}\left(D_K^2(x,p) + \frac{1}{|P|}\sum_{y\in P} D_K^2(y,p) - \frac{1}{2|P|^2}\sum_{y,z\in P} D_K^2(y,z)\right). \end{aligned} \tag{5}$$

可以观察到，核幂距离的水平集是若干球的并集（因为高斯核距离是欧几里得距离的径向函数）。此外，Phillips、Wang 和 Zheng [32][定理 3.1、引理 3.1] 证明，对于任意 $x \in \mathbb{R}^D$,

$$d_\mu^K(x)^2 \le 2f_\mu^K(x)^2 \le 4d_\mu^K(x)^2 + 6D_K^2(p,x).$$

这表明，在相差常数因子的意义下，GKPD f_μ^K近似于相对于均匀测度 d_μ^K 的高斯核距离所定义的高斯幂距离。

3 最小包围幂球

正如引言中所述，我们主要结果背后的核心技术思想，是针对使用高斯核幂距离计算的一组加权点的最小包围球半径平方所建立的一个分解定理。要证明这样的分解定理，我们需要了解在幂距离下，加权点集合的最小包围球的一些性质。然而，由于高斯核距离是非线性的，为了证明我们关于 GKPD 下加权点最小包围球的新结果，关键是将这些点提升到希尔伯特空间中，使高斯核距离对应于一个内积的范数。因此，本节将研究希尔伯特空间中的这类最小包围球，其中范数由内积给出，并

证明一些新性质，这些性质对于下一节中技术结果的证明至关重要。请注意，由于欧几里得空间也是希尔伯特空间，而且通常的 ℓ_2 距离可表示为一个内积的范数，因此我们的所有结果也适用于通常的欧几里得情形。

˘ 回顾第 2.2 节中利用幂距离定义的加权切赫复形。给定希尔伯特空间 $\mathcal{H}$ 中的一组加权点 $\{\widehat{p}_1, \ldots, \widehat{p}_k\}$，其中对于所有 $i \in [k]$,$\widehat{p}_i := (p_i, w(p_i))$ 表示点 $p_i \in \mathcal{H}$，其权重为 $w(p_i)$。点对 x, $y \in \mathcal{H}$ 之间的幂距离定义为

$$D(\hat{x}, \hat{y}) := \|x - y\|_{\mathcal{H}}^2 - w(x) - w(y),$$

其中，$\|.\|_{\mathcal{H}}$ 是由希尔伯特空间 $\mathcal{H}$ 的内积诱导的范数，且当 $x \notin P$ 时，$w(x) = 0$ 。令 $\widehat{\sigma}$ 表示由这些点构成的单形。利用上文定义的 $D(\widehat{x}, \widehat{y})$，我们分别按照公式 (2) 和 (1) 定义中心 $c(\widehat{\sigma})$ 和半径 $\mathrm{rad}(\widehat{\sigma})$。

引理 3. 中心 $c(\widehat{\sigma})$ 和半径 $\mathrm{rad}(\widehat{\sigma})$ 是唯一的。

引理 3 的证明。假设存在不同的中心 $c_0 \neq c_1$。令 r 表示 $\mathrm{rad}(\widehat{\sigma})$。根据中心的定义，我们有

$$\exists \widehat{p}_0 \in \widehat{X}: \ \forall \widehat{p}_i \in \widehat{X}: \ D(c_0, \widehat{p}_i) \leq D(c_0, \widehat{p}_0) = \|c_0 - p_0\|_{\mathcal{H}}^2 - w(p_0) = r^2.$$

$$\exists \widehat{p}_1 \in \widehat{X}: \ \forall \widehat{p}_i \in \widehat{X}: \ D(c_1, \widehat{p}_i) \leq D(c_1, \widehat{p}_1) = \|c_1 - p_1\|_{\mathcal{H}}^2 - w(p_1) = r^2.$$

对于任意 $\lambda \in (0, 1)$，定义

$$D_\lambda(\widehat{p}_i) := (1 - \lambda) D(c_0, \widehat{p}_i) + \lambda D(c_1, \widehat{p}_i),$$

and

$$c_\lambda := (1 - \lambda) c_0 + \lambda c_1.$$

于是,

$$\begin{aligned} D_\lambda(\widehat{p}_i) &= (1 - \lambda) D(c_0, \widehat{p}_i) + \lambda D(c_1, \widehat{p}_i) \\ &= (1 - \lambda) \|c_0 - p_i\|_{\mathcal{H}}^2 - (1 - \lambda) w(p_i) + \lambda \|c_1 - p_i\|_{\mathcal{H}}^2 - \lambda w(p_i) \\ &= (1 - \lambda) \|c_0 - p_i\|_{\mathcal{H}}^2 + \lambda \|c_1 - p_i\|_{\mathcal{H}}^2 - w(p_i) \\ &> \|(1 - \lambda) c_0 + \lambda c_1 - p_i\|_{\mathcal{H}}^2 - w(p_i) \qquad (6) \\ &= D(c_\lambda, p_i), \end{aligned}$$

其中，第 (6) 步依据的是：由范数定义的距离函数是严格凸函数。上面最后一行与我们的假设矛盾，即 c_0 和 c_1 是 X的两个不同$\widehat{\ }$ $\widehat{\ }$中心，因此可知 X 只能有一个中心。

□

令 I 为 $p_j \in \widehat{\ }X$的指标集合，使得 $\mathrm{rad}^2(\widehat{X}) = D(c, \widehat{p}_j)$，并令 $\widehat{\ }$ X_I 为相应的加权点集合。下一个引理表明，中心 c 可表示为 X_I中各点的凸组合。

引理 4. 存在一组非负实数 $(\lambda_i)_{i \in I}$，使得 $\sum_{i \in I} \lambda_i = 1$且$c = \sum_{i \in I} \lambda_i p_i$。

引理 4 的证明。设 $c \notin \mathrm{conv}(X_I)$。根据希尔伯特投影定理 [34], 存在唯一的 $c' \neq c$使得 $c' = \arg\inf_{x \in \mathrm{conv}(X_I)} \|c - x\|$。令 $\tilde{c} = \lambda c + (1 - \lambda) c'$，其中 $\lambda \in [0, 1]$。那么对于任意 p_i、$i \in I$，距离 $\|\tilde{c} - p_i\|$ 满足

$$\begin{aligned} \|\tilde{c} - p_i\|_{\mathcal{H}} &= \|\lambda c + (1 - \lambda) c' - p_i\|_{\mathcal{H}} \\ &= \|\lambda (c - p_i) + (1 - \lambda)(c' - p_i)\|_{\mathcal{H}} \\ &\leq \lambda \|c - p_i\|_{\mathcal{H}} + (1 - \lambda) \|c' - p_i\|_{\mathcal{H}} \ \leq \ \|c - p_i\|_{\mathcal{H}}, \end{aligned}$$

其中，最后一步是因为对于任意 $i \in I$都有 $\|c' - p_i\|_{\mathcal{H}} \leq \|c - p_i\|_{\mathcal{H}}$。将距离平方后，再从两边减去 p_i 的权重，可得对于所有 $i \in I$，

$$D(\tilde{c}, \hat{p}_i) \leq D(c, \hat{p}_i).$$

当 λ 充分接近 1 时， $\tilde{c}$ 到加权点 $\hat{}p_k, k \notin I$ 的距离仍小于到 $\hat{}p_i, i \in I$ 的距离。因此我们得到

$$D(\tilde{c}, \hat{p}_k) < D(\tilde{c}, \hat{p}_i) < D(c, \hat{p}_i) = \mathrm{rad}^2(X_I),$$

这与 c 是 X_I 的中心这一事实矛盾。

□

下一个引理表明，$\hat{}X_I$ 中各点到中心 c 的平方距离的凸组合，可以用 X_I 中各点之间的成对距离 $\hat{}$ 的线性组合来表示。证明这个引理时，我们需要假设这些点位于希尔伯特空间 $\mathcal{H}$ 中，且其范数由内积给出。

引理 5。设 $c = \sum_{i \in I} \lambda_i p_i$，其中 $\lambda_i \geq 0$ 对于 $i \in I$，且 $\sum_{i \in I} \lambda_i = 1$。那么

$$\sum_{i \in I} \lambda_i \|c - p_i\|_{\mathcal{H}}^2 = \frac{1}{2} \sum_{i,j \in I} \lambda_i \lambda_j \|p_i - p_j\|_{\mathcal{H}}^2. \tag{7}$$

引理 5 的证明。虽然可以用基本线性代数证明该引理，但我们将用概率语言陈述证明；我们认为这样更简单、更直观。令 X_1, X_2 为从 $\{p_i \mid i \in I\}$ 中独立随机选取的两个点，其中点 p_i 对每个 $i \in I$ 以概率 λ_i 选取。根据定义，我们有 $\|X_1 - X_2\|_{\mathcal{H}}^2 = \langle X_1 - X_2, X_1 - X_2 \rangle$。利用期望的线性性可得

$$\begin{aligned} \mathbb{E}\left[\|X_1 - X_2\|_{\mathcal{H}}^2\right] &= \mathbb{E}\left[\langle X_1 - X_2, X_1 - X_2 \rangle\right] \\ &= \mathbb{E}\left[\langle X_1, X_1 \rangle\right] + \mathbb{E}\left[\langle X_2, X_2 \rangle\right] - \mathbb{E}\left[\langle X_2, X_1 \rangle\right] - \mathbb{E}\left[\langle X_1, X_2 \rangle\right] \\ &= 2\left(\mathbb{E}\left[\langle X_1, X_1 \rangle\right] - \mathbb{E}\left[\langle X_2, X_1 \rangle\right]\right). \end{aligned}$$

上文中，最后一步是由 X_1 和 X_2 是相互独立且同分布的随机变量这一事实得出的。项 $\mathbb{E}\left[\langle X_2, X_1 \rangle\right]$ 可先对 X_2 求期望，利用内积对第一个变量的线性性质，然后再对 X_1 求期望来计算。因此，我们得到

$$\mathbb{E}\left[\langle X_2, X_1 \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}\left[X_2\right], X_1 \rangle\right] = \mathbb{E}\left[\langle \mathbb{E}\left[X_1\right], X_1 \rangle\right] = \langle \mathbb{E}\left[X_1\right], \mathbb{E}\left[X_1\right] \rangle.$$

因此，$\mathbb{E}\left[\|X_1 - X_2\|_{\mathcal{H}}^2\right]$ 的表达式可写为

$$\begin{aligned} \mathbb{E}\left[\|X_1 - X_2\|_{\mathcal{H}}^2\right] &= 2\left(\mathbb{E}\left[\langle X_1, X_1 \rangle\right] - \langle \mathbb{E}\left[X_1\right], \mathbb{E}\left[X_1\right] \rangle\right) \\ &= 2\left(\mathbb{E}\left[\langle X_1 - \mathbb{E}\left[X_1\right], X_1 - \mathbb{E}\left[X_1\right] \rangle\right]\right) = 2\mathbb{E}\left[\|X_1 - \mathbb{E}\left[X_1\right]\|_{\mathcal{H}}^2\right]. \end{aligned}$$

计算这些期望，得到 $\mathbb{E}\left[X_1\right] = \sum_{i \in I} \lambda_i p_i = c$，于是

$$\mathbb{E}\left[\|X_1 - X_2\|_{\mathcal{H}}^2\right] = \sum_{i,j \in I} \lambda_i \lambda_j \|p_i - p_j\|_{\mathcal{H}}^2,$$

and

$$\mathbb{E}\left[\|X_1 - \mathbb{E}\left[X_1\right]\|_{\mathcal{H}}^2\right] = \sum_{i \in I} \lambda_i \|p_i - c\|_{\mathcal{H}}^2.$$

因此，最终表达式为

$$\begin{aligned} \sum_{i,j \in I} \lambda_i \lambda_j \|p_i - p_j\|_{\mathcal{H}}^2 &= \mathbb{E}\left[\|X_1 - X_2\|_{\mathcal{H}}^2\right] \\ &= 2\mathbb{E}\left[\|X_1 - \mathbb{E}\left[X_1\right]\|_{\mathcal{H}}^2\right] = 2\sum_{i \in I} \lambda_i \|c - p_i\|_{\mathcal{H}}^2. \end{aligned}$$

□

下面的分解引理表明，加权点集 X 的最小包围球的半径平方 ^可以表示为其中点两两之间的幂^距离的组合。

引理 6（分解引理）。设 I 是引理 4 中定义的指标集合。 n

$$\mathrm{rad}^2(\widehat{X}) = \frac{1}{2}\sum_{i\in I}\sum_{j\in I}\lambda_i\lambda_j D(\widehat{p}_i,\widehat{p}_j).$$

引理 6 的证明。对于所有 $i \in I$，有

$$\mathrm{rad}^2(\widehat{X}) \;=\; \sum_{i\in I}\lambda_i(\|c-p_i\|_{\mathcal{H}}^2 - w(p_i)), \tag{8}$$

因此，根据引理 5，有

$$\sum_{i\in I}\lambda_i\|c-p_i\|_{\mathcal{H}}^2 \;=\; \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\|p_i-p_j\|_{\mathcal{H}}^2.$$

代入方程 (8)，

$$\begin{aligned}
\mathrm{rad}^2(\widehat{X}) &= \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\|p_i-p_j\|_{\mathcal{H}}^2 - \frac{1}{2}\sum_{i\in I}2\lambda_i w(p_i) = \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\|p_i-p_j\|_{\mathcal{H}}^2 - \frac{1}{2}\sum_{i,j\in I}2\lambda_i\lambda_j w(p_i)\\
&= \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\|p_i-p_j\|_{\mathcal{H}}^2 - \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j(w(p_i)+w(p_j))\\
&= \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j\left(\|p_i-p_j\|_{\mathcal{H}}^2 - w(p_i) - w(p_j)\right) = \frac{1}{2}\sum_{i,j\in I}\lambda_i\lambda_j D(\widehat{p}_i,\widehat{p}_j).
\end{aligned}$$

□

最后，我们提及 GKPD 的以下有趣推论。考虑一组点 $p_1,\ldots,p_k \in \mathbb{R}^D$，根据 (4) 中定义的 GKPD 为其赋予权重，并令 $\widehat{\sigma}$ 表示由 $\{\widehat{p}_1,\ldots,\widehat{p}_k\}$ 构成的相应抽象单形。根据第 4.2 节，存在一个提升映射 $\phi:\ \mathbb{R}^D \to \mathcal{H}_K$，其中 $\mathcal{H}_K$ 是希尔伯特空间，使得任意一对点之间的高斯核距离平方对应于其在 $\mathcal{H}_K$ 中的位置向量之差的内积范数。将引理 3 和引理 6 应用于 ^ p_i $i \in [k]$ 在提升映射 ϕ 下的像，并利用 GKPD 下原始点成对距离的相应结论，我们推导出 GKPD 下最小包围球的以下新性质：

推论 7. 1. $\widehat{\sigma}$ 的中心和半径是唯一的。

2. 存在一个非负实数集 $(\lambda_i)_{i\in[k]}$，使得 $\sum_{i\in[k]}\lambda_i = 1$，并且

$$\mathrm{rad}^2(\widehat{\sigma}) = \frac{1}{2}\sum_{i\in[k]}\sum_{j\in[k]}\lambda_i\lambda_j D(\widehat{p}_i,\widehat{p}_j).$$

4 幂距离的低失真映射

本节将探讨幂距离的低失真映射。首先，我们需要定义度量空间之间幂距离的 ε-失真映射。

定义 8. 给定度量空间 (X, d_X) 和 (Y, d_Y)、点集 $P \subset X$ 以及 $\varepsilon \in (0, 1)$，映射 $f : X \to Y$ 是一个关于 P 中成对距离的 ε-失真映射，若：

$$\forall x, y \in P : \ (1-\varepsilon) d_X(x,y)^2 \leq d_Y(f(x), f(y))^2 \leq (1+\varepsilon) d_X(x,y)^2.$$

进一步地，给定一对权重函数 $w_X : P \to \mathbb{R}$ 和 $w_Y : P \to \mathbb{R}$, f 是一个关于 w_X 的 ε-畸变映射，当且仅当：

$$\forall x \in P : \ |w_Y(f(x)) - w_X(x)| \leq \varepsilon w_X(x).$$

因此，给定一个定义为 $D_X(\hat{x}, \hat{y}) = d_X(x,y)^2 - w_X(x) - w_X(y)$ 的幂距离，其中 x、$y \in P$ 和 $\hat{x} = (x, w_X(x))$，$x \in P$ 是加权点，映射 $f : X \to Y$ 是一个关于幂距离 D_X 的 ε-畸变映射，当且仅当 f 对成对距离以及权重函数 w_X, w_Y 都是一个 ε-畸变映射。

4.1 再生核希尔伯特空间与降维

在高斯核的背景下，最广为人知的 ε-畸变映射实例或许是 Rahimi 和 Recht 的 RFF 映射 [33], , Chen 和 Phillips [13] 已证明它是高斯核距离的一个 ε-畸变映射。

对于 $\mathbb{R}^D$ 中的点，存在一个映射到 $\mathbb{R}^t$ 的映射，其中 $t = O\left(\varepsilon^{-2} \log n\right)$ 可对 $\mathbb{R}^D$ 上的核距离给出相对近似，因为 $\mathbb{R}^t$ 上的自然内积。他们的 RFF 映射如下：对于 $i = 1, \ldots, d/2$，给定 $\sigma \geq 0$，令 $\omega_i \sim \mathcal{N}_D(0, \sigma^{-2})$ 为 D 维相互独立的高斯随机变量。对于每个 i，定义随机映射 $f_i : \mathbb{R}^D \to \mathbb{R}^2$，如下所示：

$$f_i(x) = (\cos(\langle \omega_i, x \rangle), \sin(\langle \omega_i, x \rangle)).$$

最后，定义映射 $f : \mathbb{R}^D \to \mathbb{R}^t$ 如下：

$$f(x) \ = \ \bigotimes_{i=1}^{d/2} f_i(x). \tag{9}$$

定理 9（Chen、Phillips [13]）。给定任意 $\varepsilon, \delta \in (0, 1)$，对于任意集合 $P \subset \mathbb{R}^D$，其中包含 (i) 个 n 点，或 (ii) 任意数量的点，使得对于所有 $x, y \in P$, $\|x - y\|/\sigma \leq r$，其中 $r > 0$ 是给定参数，$f : \mathbb{R}^D \to \mathbb{R}^t$ 如公式（9）所定义，且具有 (i) $t := \Omega\left(\varepsilon^{-2} \log(n/\delta)\right)$ 维度，或 (ii) $t := \Omega\left(\varepsilon^{-2} D \log(rD/\varepsilon\delta)\right)$ 维度，分别是高斯核距离的 ε-失真映射，即 f 对所有点对 $x, y \in P$ 均满足 $\frac{\|f_{(x-fy)}\|^2)()(,)}{D_K^2 \ x \ y} \in (1-\varepsilon, 1+\varepsilon)$，其概率至少为 $1 - \delta$。

注 2. 对于高斯核函数，存在一个提升映射 ϕ，将 $\mathbb{R}^D$ 映射到希尔伯特空间 $\mathcal{H}_K$，使得核函数提升为 $\mathcal{H}_K$ 上的内积（对于$\mathbb{R}^D$中的任意一对点）。与再生核 $K(.,.)$ 相对应的希尔伯特空间 $\mathcal{H}_K$ 通常是无限维的。然而，Rahimi 和 Recht 提出的 RFF 映射 *[33],* ，以及 Chen 和 Phillips *[13]*表明，可以用有限维空间上的欧几里得内积近似 $\mathcal{H}_K$ 上的内积。

根据注 1，我们知道，高斯权重函数需要在像空间中重新计算。对于每个 $p \in P$，定义

$$w_t(p) := -\left(\frac{1}{|P|} \sum_{y \in P} \|f(p) - f(y)\|^2 - \frac{1}{2|P|^2} \sum_{x,y \in P} \|f(x) - f(y)\|^2 \right). \tag{10}$$

现在比较高斯核权重函数的定义式 (4) 和式 (10)，并应用定理 9，可以看出，f 也是高斯核权重函数的 ε-畸变映射，因此也是 GKPD 的 ε-畸变映射。

推论 10。定理 9 中的映射 f 是 GKPD 的 ε-畸变映射。

4.2 ε-畸变映射与 GKPD 下的单形半径

现在，我们可以证明：对高维空间中的点样本应用 ε-畸变映射，切赫过滤中各单形的最小包围球半径至多会改变一个 $(1 \pm \varepsilon)$ 倍的因子。这就是单形畸变引理，下面将陈述并证明该引理。

引理 11（单形畸变引理）。设 $\widehat{\sigma} \subset \widehat{P}$ 是使用高斯核幂距离 f_μ^K: $\mathbb{R}^D \to \mathbb{R}$ 的加权 ˇCech 复形 $\check{C}_\alpha(\widehat{P})$ 中的一个单形，且 $G: (\mathbb{R}^D, D_K) \to (\mathbb{R}^t, \|.\|)$ 是一个 ε-畸变映射，对应由 $D(\hat{p}, \hat{q}) = D_K^2(p,q) - w(p) - w(q), \hat{p}, \hat{q} \in P$ 给出的成对幂距离。那么

$$(1-\varepsilon)\mathrm{rad}^2(\widehat{\sigma}) \leq \mathrm{rad}^2(\widehat{G(\sigma)}) \leq (1+\varepsilon)\mathrm{rad}^2(\widehat{\sigma}),$$

其中 $G(\sigma)$ 表示单形 σ 在 $\mathbb{R}^t$ 中的像，权重在 $\mathbb{R}^t$ 中按照式（10）重新计算。

单形畸变引理 11 的证明。设单形为 $\widehat{\sigma} = \{\widehat{p}_1, \ldots, \widehat{p}_k\}$，其中对所有 $i \in [k], \widehat{p}_i = (p_i, w(p_i))$ 是一个加权点；并设 $c(\widehat{\sigma})$ 和 $\mathrm{rad}(\widehat{\sigma})$ 分别表示它的中心和半径。

由于高斯核是特征核，因此存在一个希尔伯特空间 $\mathcal{H}_K$ 和一个提升映射 ϕ: $\mathbb{R}^D \to \mathcal{H}_K$，使得对于所有 $x, y \in \mathbb{R}^D$, $D_K(x,y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}_K}$（例如参见 [13, 31]）。根据等式 (4)，点 $p \in P$ 的权重是核距离平方的加权和：

$$w(p) := -D_K^2(\mu, p) = -\left(\frac{1}{|P|}\sum_{y\in P} D_K^2(p,y) - \frac{1}{2|P|^2}\sum_{x,y\in P} D_K^2(x,y)\right).$$

因此，提升映射 ϕ 可自然地扩展到权重 $w(p_i)$, $i \in [k]$，如下所示：

$$\phi(w(p_i)) = -\left(\frac{1}{|P|}\sum_{y\in P} \|\phi(p)-\phi(y)\|_{\mathcal{H}_K}^2 - \frac{1}{2|P|^2}\sum_{x,y\in P} \|\phi(x)-\phi(y)\|_{\mathcal{H}_K}^2\right),$$

这使我们能够将提升空间中的权重定义为 $w(\phi(p)) := \phi(w(p))$.

对 $\mathcal{H} = \mathcal{H}_K$ 应用分解引理 6，并使用由 $\phi(\hat{p}_i) := (\phi(p_i), \phi(w(p_i)))$ 给出的提升后的加权点，可得

$$\mathrm{rad}^2(\widehat{\sigma}) = \frac{1}{2}\sum_{i,j\in[k]} \lambda_i\lambda_j D(\widehat{p}_i, \widehat{p}_j). \tag{11}$$

由于 G 是一个 ε-畸变映射，对于每一对 $\widehat{p}_i, \widehat{p}_j \in \widehat{\sigma}$，有

$$(1-\varepsilon)\|\phi(p_i)-\phi(p_j)\|_{\mathcal{H}_K}^2 \leq \|G(p_i)-G(p_j)\|^2 \leq (1+\varepsilon)\|\phi(p_i)-\phi(p_j)\|_{\mathcal{H}_K}^2 \tag{12}$$

由于 G 对权重函数也是一个 ε-畸变映射，因此对于每个 $\widehat{p}_i \in \widehat{\sigma}$，我们有

$$(1-\varepsilon)w(\widehat{p}_i) \leq w(\widehat{G(p_i)}) \leq (1+\varepsilon)w(\widehat{p}_i). \tag{13}$$

从平方距离 $\|G(p_i) - G(p_j)\|^2$ 中减去权重 $w(\widehat{G(p_i)})$ $w(\widehat{G(p_j)})$，并利用 $D(\hat{p}_i, \hat{p}_j) = D_K^2(p_i, p_j) - w(p_i) - w(p_j) = \|\phi(p_i) - \phi(p_j)\|_{\mathcal{H}_K} - w(p_i) - w(p_j)$，得到

$$(1-\varepsilon)D(\widehat{p}_i, \widehat{p}_j) \leq D(\widehat{G(p_i)}, \widehat{G(p_j)}) \leq (1+\varepsilon)D(\widehat{p}_i, \widehat{p}_j). \tag{14}$$

令 $G(\sigma)$ 表示单形 $\widehat{\sigma}$ 在映射 G 下的像，$c(G(\sigma))$ 表示其中心。在空间 $(\mathbb{R}^t, \|.\|)$ 上应用引理 4，可得 $c(G(\sigma))$ 是 $G(\sigma)$ 的顶点的凸组合，即

$$c(\widehat{G(\sigma)}) = \sum_{i\in[k]} \mu_i G(\widehat{p}_i),$$

其中 $\forall i \in [k]; \mu_i \geq 0$ 和 $\sum_{i\in[k]} \mu_i = 1$[2]。由于 G 是一个 ε-畸变映射，利用分解引理 6 可得

$$\begin{aligned} \mathrm{rad}^2(\widehat{G(\sigma)}) &= \sum_{i,j\in[k]} \mu_i\mu_j \left(\frac{1}{2} D(\widehat{G(p_i)}, \widehat{G(p_j)})\right) \\ &\geq \sum_{i,j\in[k]} \frac{\mu_i\mu_j}{2} \left((1-\varepsilon) D(\widehat{p}_i, \widehat{p}_j)\right) \end{aligned} \tag{15}$$

$$\text{i.e.} \sum_{i,j\in[k]} \frac{\mu_i\mu_j}{2} \left(D(\widehat{p}_i, \widehat{p}_j)\right) \leq \frac{\mathrm{rad}^2(\widehat{G(\sigma)})}{1-\varepsilon}. \tag{16}$$

此外，根据加权单形半径平方定义中的极小性，我们有

$$\mathrm{rad}^2(\widehat{\sigma}) = \frac{1}{2} \sum_{i,j\in[k]} \lambda_i\lambda_j D(\widehat{p}_i, \widehat{p}_j) \leq \frac{1}{2} \sum_{i,j\in[k]} \mu_i\mu_j D(\widehat{p}_i, \widehat{p}_j), \quad \text{and} \tag{17}$$

$$\begin{aligned} \mathrm{rad}^2(\widehat{G(\sigma)}) &= \frac{1}{2} \sum_{i,j\in[k]} \mu_i\mu_j D(\widehat{G(p_i)}, \widehat{G(p_j)}) \leq \frac{1}{2} \sum_{i,j\in[k]} \lambda_i\lambda_j D(\widehat{G(p_i)}, \widehat{G(p_j)}), \\ &\leq \sum_{i,j\in[k]} \frac{\lambda_i\lambda_j}{2} \left((1+\varepsilon) D(\widehat{p}_i, \widehat{p}_j)\right), \qquad (18) \\ &= (1+\varepsilon)\mathrm{rad}^2(\widehat{\sigma}) \qquad (19) \end{aligned}$$

其中在步骤 (18) 中，我们再次利用了 G 是一个 ε-畸变映射这一事实，最后一步则由分解引理 6 得出。结合方程 (16)、(17) 和 (19)，可得

$$(1-\varepsilon)\mathrm{rad}^2(\widehat{\sigma}) \leq \mathrm{rad}^2(\widehat{G(\sigma)}) \leq (1+\varepsilon)\mathrm{rad}^2(\widehat{\sigma}), \tag{20}$$

由此完成引理的证明。 □

我们以陈述主要的分解定理作为本节的结尾。

定理 12. 设 $P = \{p_1, \ldots, p_k\}$ 是一组加权点，使得 $p = (p, w(p))$，其中 $p \in \mathbb{R}^D$ 和 $w : \mathbb{R}^D \to \mathbb{R}$ 为 GKPD 给出的权重函数（方程 (4)）。设 $G: (\mathbb{R}^D, D_K) \to (\mathbb{R}^t, \|.\|)$ 是一个 ε-畸变映射，对应于由 $D(p_i, p_j) = D_K^2(p_i, p_j) - w(p_i) - w(p_j)$ 给出的幂距离。那么，使用 GKPD $f_\mu^K: \mathbb{R}^D \to \mathbb{R}$ 计算得到的切赫滤子 $\check{C}_\alpha(\widehat{P})$，以及在映射 G 下 $\mathbb{R}^t$ 中相应的加权过滤（权重按式 (10) 在 $\mathbb{R}^t$ 中重新计算），是乘法意义下 $((1-\varepsilon)^{-1} + o(1))$-交错的。

将单形畸变引理应用于 $\check{C}_\alpha(\widehat{P})$ 中的每个单形，即可直接得出定理 12 的证明。

[2]请注意，对于 $c(\widehat{\sigma})$，凸组合 $(\mu_i)_{i\in[k]}$ 与原空间中的组合 $(\lambda_i)_{i\in[k]}$ 不一定相同。

5 主要结果的证明

本节将证明我们的主要结果，陈述如下。

定理 13。给定 $\sigma > 0$、 ε、$\delta \in (0,1)$、 $\sigma > 0$，以及一个由以下元素组成的有限集 $P \subset \mathbb{R}^D$

(i) n个点，或

(ii) 任意数量的点，其欧几里得直径至多为 $r\sigma$，其中 $r > 0$ 是给定参数，

分别将随机傅里叶特征投影映射到 (i) $t := \Omega\left(\varepsilon^{-2}\log(n/\delta)\right)$ 维，或 (ii) $t := \Omega\left(D\varepsilon^{-2}\log(Dr/\varepsilon\delta)\right)$ 维，使得使用 GKPD 计算的切赫滤子与 $\mathbb{R}^t$ 中对应的加权过滤以至少 $1-\delta$ 的概率满足 $((1-\varepsilon)^{-1}+o(1))$-交错。此外，随机傅里叶特征映射 $f\colon \mathbb{R}^D \to \mathbb{R}^t$ 可在时间 $O(nt)$ 内计算。

定理 13 的证明。为了证明映射到 t 个维度的 RFF 映射能得到一个数据集，使其加权切赫过滤（其中权重在像空间中重新计算）与原始滤子交错，只需证明：对于任意加权单形 σ，在 GKPD 下，σ 的半径经 RFF 映射后以高概率至多发生 $(1 \pm \varepsilon)$ 倍的失真。因此，我们的总体思路是应用定理 12，并使用由定理 9 得到的 ε-失真映射。以下事实——已在第 3 节（尤其是推论 7）和第 4 节中证明——将是证明的关键。

1. 在 GKPD 下，$\check{C}_\alpha(\widehat{P})$ 中原始加权单形的半径平方，在 $\mathbb{R}^D$ 中可表示为 $\hat{}\sigma$ 的顶点两两广义核功率距离的线性组合。 2. $\check{C}_\alpha(\widehat{G\ P\ (}))$ 中的加权单形在欧几里得距离下的半径平方（权重在 $\mathbb{R}^t$ 中重新计算，如式 (10) 所示）可表示为 $G(\sigma)$ 中顶点的加权欧氏距离的线性组合。 3. RFF 映射近似保持 $\hat{}\sigma$ 顶点之间的两两广义核功率距离，因此也近似保持这些功率距离的每个线性组合，尤其是与 $\widehat{\sigma}$ 的原始半径和新半径相对应的线性组合；由此即可对二者进行比较。

因此，为了应用定理 12，我们来验证：给定一个 ε-畸变映射，定理的条件是否成立。应用定理 9，我们得到函数 $f\colon \mathbb{R}^D \to \mathbb{R}^t$，使得

$$\forall x, y \in P:\ \ (1-\varepsilon)D_K(x,y) \le \|f(x) - f(y)\| \le (1+\varepsilon)D_K(x,y).$$

因此，我们得到 f 是关于核距离 D_K 的 ε-畸变映射。对于权重函数$w : P \to \mathbb{R}$，由式 (4) 可知，点 $p \in P$ 的权重是核距离平方的加权和。现在，将 $\mathbb{R}^t$ 中的权重按式 (10) 重新定义后，推论 10 表明这些权重是 $(1 \pm \varepsilon)$-保持的。

因此，定理 12 的所有条件都满足。

现在，我们可以应用定理 12，得到使用切赫核距离函数构造的加权切赫滤子 $\check{C}_\alpha(\widehat{P})$，与使用欧几里得距离在加权点集 $\widehat{P}$ 经 G 映射后的像上构造的切赫滤子交错，即 $\check{C}_\alpha(\widehat{G\ P\ (}))$，如下所示，

$$\check{C}_{\alpha_-}(\widehat{P}) \subseteq \check{C}_\alpha(\widehat{G(P)}) \subseteq \check{C}_{\alpha_+}(\widehat{P}).$$

其中 $\alpha_- := \alpha\sqrt{1-\varepsilon}$ 且 $\alpha_+ := \alpha\sqrt{1+\varepsilon}$。最后，由于对于 $\varepsilon \in (0,1)$，我们有 $1 < (1+\varepsilon) < (1-\varepsilon)^{-1}$，因此可以应用第 2.1 节中 β-交错的定义，得到定理的结论。

□

6 结论

我们已经证明，随机傅里叶特征映射可用于降低构建持久图所需输入数据的维数。需要注意的是，和 Johnson-Lindenstrauss 引理的应用类似，降维后的维数带有一个 ε^{-2} 因子，因此当允许误差 ε 非常小时，维数实际上无法降得很低。尽管如此，它仍能将维数从几百万（人工智能应用中的典型规模）降至几百。如果能将这种技术与其他降维技术结合起来，例如主成分分析或基于梯度下降的技术，将会很有意思。

参考文献

[1] Dimitris Achlioptas。《数据库友好的随机投影：使用二进制随机变量的 Johnson–Lindenstrauss 变换》。《计算机与系统科学杂志》，66(4):671–687，2003。PODS 2001 专刊。

[2] Nir Ailon 和 Bernard Chazelle。《近似最近邻与快速 Johnson–Lindenstrauss 变换》。载于第三十八届 ACM 计算理论研讨会论文集（STOC '06），第 557–563 页，纽约，纽约州，美国，2006。ACM。

[3]Hirokazu Anai、Fr´ed´eric Chazal、Marc Glisse、Yuichi Ike、Hiroya Inakoshi、Rapha¨el Tinarrage 和 Yuhei Umeda。《基于 DTM 的滤过》。载于 Gill Barequet 和 Yusu Wang 编，第 35 届国际计算几何研讨会（SoCG 2019），2019 年 6 月 18–21 日，美国俄勒冈州波特兰，LIPIcs 第 129 卷，第 58:1–58:15 页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2019。网址：https://doi.org/10.4230/LIPIcs.SoCG.2019.58，doi:10.4230/LIPIcs.SoCG.2019.58。

[4]Shreya Arya、Jean-Daniel Boissonnat、Kunal Dutta 和 Martin Lotz。《应用于持久同调的 k-距离降维》。J. Appl. Comput. Topol., 5(4):671–691, 2021。网址：https://doi.org/10.1007/s41468-021-00079-x，doi:10.1007/s41468-021-00079-x。[5] Haim Avron、Michael Kapralov、Cameron Musco、Christopher Musco、Ameya Velingker 和 Amir Zandieh。《用于核岭回归的随机傅里叶特征：近似界与统计保证》。载于第 34 届国际机器学习会议论文集——第 70 卷，ICML'17，第 253–262 页。JMLR.org, 2017。[6] Karol Borsuk。《关于将紧空间系嵌入单纯复形》。Fundamenta Mathematicae, 35(1):217–234, 1948。网址：http://eudml.org/doc/213158。[7]Micka¨el Buchet、Fr´ed´eric Chazal、Steve Y. Oudot 和 Donald R. Sheehy。《面向测度的高效且稳健的持久同调》。Comput. Geom., 58:70–96, 2016。网址：https://doi.org/10.1016/j.comgeo.2016.07.001，doi:10.1016/j.comgeo.2016.07.001。[8] Micka¨el Buchet、Tamal K. Dey、Jiayuan Wang 和 Yusu Wang。《去杂与重采样：迈向无参数去噪》。JoCG, 9(2):21–46, 2018。网址：https://doi.org/10.20382/jocg.v9i2a3，doi:10.20382/jocg.v9i2a3。[9] Moses Charikar 和 Paris Siminelakis。《基于哈希的高维核密度估计器》。载于 2017 年 IEEE 第 58 届计算机科学基础年度研讨会（FOCS），第 1032–1043 页，2017。doi:10.1109/FOCS.2017.99。

[10]Fr´ed´eric Chazal、David Cohen-Steiner 和 Quentin M´erigot。《概率测度的几何推断》。Found. Comput. Math., 11(6):733–751, 2011。网址：https://doi.org/10.1007/s10208-011-9098-0, doi:10.1007/s10208-011-9098-0。

[11] Fr´ed´eric Chazal、Vin de Silva、Marc Glisse 和 Steve Y. Oudot。《持久性模的结构与稳定性》。施普林格数学简明读本。施普林格，2016。URL：https://doi.org/10.1007/978-3-319-42545-0, doi:10.1007/978-3-319-42545-0。

[12] Fr´ed´eric Chazal 和 Steve Oudot。《面向基于持久性的欧几里得空间重建》。载于 Monique Teillaud（编），《第24届 ACM 计算几何研讨会论文集》，美国马里兰州学院公园，2008年6月9日至11日，第232–241页。ACM，2008。URL：https://doi.org/10.1145/1377676.1377719, doi:10.1145/1377676.1377719。

[13]Di Chen 和 Jeff M. Phillips。高斯核距离的相对误差嵌入。载于《国际算法学习理论会议》，ALT 2017，2017 年 10 月 15—17 日，日本京都京都大学，第 560–576 页，2017 年。网址：http://proceedings.mlr.press/v76/chen17a.html。

[14]Sanjoy Dasgupta 和 Anupam Gupta。John- son 和 Lindenstrauss 定理的初等证明。《随机结构与算法》，22(1)：60–65，2003 年。网址：https://doi.org/10.1002/rsa.10073, doi:10.1002/rsa.10073。

[15]Herbert Edelsbrunner 和 John Harer。《计算拓扑学——引言》。美国数学学会，2010 年。网址：http://www.ams.org/bookstore-getitem/item=MBK-69。

[16]Herbert Edelsbrunner、David Letscher 和 Afra Zomorodian。拓扑持久性与简化。离散与计算几何，28(4):511–533，2002。网址：https://doi.org/10.1007/s00454-002-2885-2, doi:10.1007/s00454-002-2885-2。[17]Michael Elkin、Arnold Filtser 和 Ofer Neiman。终端嵌入。理论计算机科学，697:1–36，2017。网址：https://doi.org/10.1016/j.tcs.2017.06.021, doi:10.1016/j.tcs.2017.06.021。[18]Robert Ghrist。应用拓扑学基础。Createspace，2014年9月。网址：https://www2.math.upenn.edu/~ghrist/notes.html。[19]C. Giraud。《高维统计学导论》。Chapman & Hall/CRC统计学与应用概率专著。Taylor & Francis，2014。网址：https://www.crcpress.com/Introduction-to-High-Dimensional-Statistics/Giraud

/p/book/9781

[20]Leonidas J. Guibas、Dmitriy Morozov 和 Quentin M´erigot。《见证 k-距离》。离散与计算几何，49(1):22–45，2013年。网址：https://doi.org/10.1007/s00454-012-9465-x, doi:10.1007/s00454-012-9465-x。[21] Piotr Indyk、Rajeev Motwani、Prabhakar Raghavan 和 Santosh S. Vempala。《多维空间中的局部敏感哈希》。载于 Frank Thomson Leighton 和 Peter W. Shor 编，《第二十九届 ACM 计算理论年会论文集》，美国得克萨斯州埃尔帕索，1997年5月4日至6日，第618–625页。ACM，1997年。网址：https://doi.org/10.1145/258533.258656，doi:10.1145/258533.258656。[22]William B Johnson 和 Joram Lindenstrauss。《利普希茨映射到希尔伯特空间的扩张》。《当代数学》，26(189-206):1，1984年。网址：https://doi.org/10.1007/BF02764938, doi:10.1007/BF02764938。

[23] Sarang C. Joshi、Raj Varma Kommaraju、Jeff M. Phillips 和 SureshVenkatasubramanian。《使用核距离比较分布和形状》。载于 Ferran Hurtado 和 Marc J. van Kreveld 编，《第27届 ACM计算几何研讨会论文集》，法国巴黎，2011年6月13日至15日，第47–56页。ACM，2011年。

[24] Daniel M. Kane 和 Jelani Nelson。《更稀疏的 Johnson-Lindenstrauss 变换》。《ACM期刊》，61(1):4:1–4:23，2014年。网址：https://doi.org/10.1145/2559902，doi:10.1145/2559902。

[25] Martin Lotz。《低复杂度模型的持久同调》。《皇家学会学报A辑：数学、物理与工程科学》，475(2230):20190081，2019年。网址：https://doi.org/10.1098/rspa.2019.0081，doi:10.1098/rspa.2019.0081。

[26] Sepideh Mahabadi、Konstantin Makarychev、Yury Makarychev 和 Ilya P. Razenshteyn。《通过外双Lipschitz扩张实现非线性降维》。载于 Ilias Diakonikolas、David Kempe 和 Monika Henzinger 编，《第50届 ACM SIGACT计算理论研讨会论文集》，STOC 2018，美国加利福尼亚州洛杉矶，2018年6月25日至29日，第1088–1101页。ACM，2018年。网址：https://doi.org/10.1145/3188745.3188828，doi:10.1145/3188745.3188828。

[27]Jiˇr´ı Matouˇsek。《Johnson–Lindenstrauss 引理的若干变体》。Random Structures & Algorithms，33(2)：142–156，2008年。网址：https://doi.org/10.1002/rsa.20218，doi:10.1002/rsa.20218。

[28] Cameron Musco 和 David P. Woodruff。《核低秩近似能否实现输入稀疏度时间？》。载于第31届神经信息处理系统国际会议论文集，NIPS'17，第4438–4448页，美国纽约州雷德胡克，2017年。Curran Associates Inc.

[29]Shyam Narayanan 和 Jelani Nelson。《欧几里得空间中的最优终端降维》。载于 Moses Charikar 和 Edith Cohen 编，第51届 ACM SIGACT计算理论研讨会论文集，STOC 2019，美国亚利桑那州菲尼克斯，2019年6月23日至26日，第1064–1069页。ACM，2019年。网址：https://doi.org/10.1145/3313276.3316307，doi:10.1145/3313276.3316307。

[30] Jeff M. Phillips。适用于核的 ε-样本。载于 Sandeep Khanna（编），《第二十四届 ACM-SIAM 离散算法年度研讨会论文集》，SODA 2013，美国路易斯安那州新奥尔良，2013年1月6日至8日，第1622–1632页。SIAM，2013年。

[31]Jeff M. Phillips 和 Wai Ming Tai。用于近似相对误差核距离的高斯草图。载于 Jarosław Byrka 和 Raghu Meka 编，《近似、随机化与组合优化：算法与技术》，APPROX/RANDOM 2020，2020年8月17日至19日，线上会议，LIPIcs 第 176 卷，页码 12:1–12:20。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2020。网址：https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2020.12，doi:10.4230/LIPIcs.APPROX/RANDOM.2020.12。[32]Jeff M. Phillips、Bei Wang 和 Yan Zheng。核密度估计的几何推断。载于 Lars Arge 和 J´anos Pach 编，《第 31 届国际计算几何研讨会》，SoCG 2015，2015年6月22日至25日，荷兰埃因霍温，LIPIcs 第 34 卷，页码 857–871。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2015。网址：https://doi.org/10.4230/LIPIcs.SOCG.2015.857，doi:10.4230/LIPIcs.SOCG.2015.857。[33] Ali Rahimi 和 Benjamin Recht。面向大规模核机器的随机特征。载于 John C. Platt、Daphne Koller、Yoram Singer 和 Sam T. Roweis 编，《神经信息处理系统进展》

《神经信息处理系统 20：第二十一届神经信息处理系统年度会议论文集》，加拿大不列颠哥伦比亚省温哥华，2007年12月3日至6日，页码 1177–1184。Curran Associates, Inc.，2007。网址：https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstrac

[34] Walter Rudin。《实分析与复分析》，第3版。McGraw-Hill, Inc.，美国，1987年。

[35] Donald R. Sheehy。随机投影下距离函数的持久同调。载于Siu-Wing Cheng和Olivier Devillers编，《第30届计算几何年度研讨会》，SOCG'14，日本京都，2014年6月8日至11日，第328页。ACM，2014年。网址：https://doi.org/10.1145/2582112.2582126，doi:10.1145/2582112.2582126。

[36] Bharath K. Sriperumbudur、Kenji Fukumizu和Gert R. G. Lanckriet。普适性、特征核与测度的再生核希尔伯特空间嵌入。《机器学习研究杂志》，12：2389–2410，2011年7月。

[37] Siddharth Vishwanath、Kenji Fukumizu、Satoshi Kuriki 和 Bharath Sriperumbudur。使用再生核构建鲁棒持久图。载于《第 34 届神经信息处理系统国际会议论文集》，NIPS'20，美国纽约州雷德胡克，2020 年。Curran Associates Inc.

[38] Shusen Wang、Alex Gittens 和 Michael W. Mahoney。使用 nystr¨om 近似的可扩展核 k 均值聚类：相对误差界。《机器学习研究杂志》，20(1):431–479，2019 年 1 月。