Appl. Comput. Harmon. Anal. 36 (2014) 302–315

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Cover-based bounds on the numerical rank of Gaussian kernels

Amit Bermanis [a], Guy Wolf [b], Amir Averbuch [b,*]

[a] Department of Applied Mathematics, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[b] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 25 July 2012
Received in revised form 20 January 2013
Accepted 27 May 2013
Available online 31 May 2013
Communicated by M.V. Wickerhauser

Keywords:
Spectral analysis
Kernel method
Numerical rank
Gaussian convolution operator

ABSTRACT

A popular approach for analyzing high-dimensional datasets is to perform dimensionality reduction by applying non-parametric affinity kernels. Usually, it is assumed that the represented affinities are related to an underlying low-dimensional manifold from which the data is sampled. This approach works under the assumption that, due to the low-dimensionality of the underlying manifold, the kernel has a low numerical rank. Essentially, this means that the kernel can be represented by a small set of numerically-significant eigenvalues and their corresponding eigenvectors.

We present an upper bound for the numerical rank of Gaussian convolution operators, which are commonly used as kernels by spectral manifold-learning methods. The achieved bound is based on the underlying geometry that is provided by the manifold from which the dataset is assumed to be sampled. The bound can be used to determine the number of significant eigenvalues/eigenvectors that are needed for spectral analysis purposes. Furthermore, the results in this paper provide a relation between the underlying geometry of the manifold (or dataset) and the numerical rank of its Gaussian affinities.

The term cover-based bound is used because the computations of this bound are done by using a finite set of small constant-volume boxes that cover the underlying manifold (or the dataset). We present bounds for finite Gaussian kernel matrices as well as for the continuous Gaussian convolution operator. We explore and demonstrate the relations between the bounds that are achieved for finite and continuous cases. The cover-oriented methodology is also used to provide a relation between the geodesic length of a curve and the numerical rank of Gaussian kernel of datasets that are sampled from it.

1. Introduction

The rapid development of data collection techniques together with high availability of data and storage space introduce increasingly big high-dimensional datasets that fit data analysis tasks. In many cases the quantity of data does not reflect on its quality. Usually, it contains many redundancies that do not add important information over a limited set of representatives. Furthermore, more often than not, the distribution of samples (also called data points) is significantly affected by the sampling techniques that are used. These problems affect both the massive size of the sampled datasets and their high-dimensionality, which in turn prevent classical statistical methods from being effective tools to analyze these datasets due to the "curse of dimensionality" phenomenon.

* Corresponding author. Fax: +972 3 6422020.
E-mail address: amir@math.tau.ac.il (A. Averbuch).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.05.004

Due to the vast number of observable quantities that can be measured/sensed and used as parameters or features, the raw representation of the data is usually high-dimensional. Recent dimensionality reduction methods use manifolds to cope with this problem. Under this manifold existence assumption, a dataset is assumed to be sampled from an Euclidean submanifold that has a relatively small intrinsic dimension. The ambient high-dimensional Euclidean space of the manifold is defined by the raw parameters (or features) of the dataset. These parameters are mapped via non-linear functions to low-dimensional coordinates of the manifold, which represent the independent factors that control the behaviors of the analyzed phenomenon.

Several methods have been suggested to provide a low-dimensional representation of data points by preserving the intrinsic structure of their underlying manifold. Kernel methods such as k-PCA [13,17], LLE [16], Isomaps [19], Laplacian Eigenmaps [2], Hessian Eigenmaps [9], Local Tangent Space Alignment [22,23] and Diffusion Maps [5] have been used for this task. These methods extend the classical PCA [11,10] and MDS [8,12] methods that project the data on a low-dimensional hyperplane that preserves most of the variance in the dataset. Kernel methods substitute the linear relations (i.e., inner-products) that are preserved by PCA and MDS with a kernel construction that introduces the synonymous notion of similarity, proximity, or affinity between data points. Spectral analysis of this kernel is used to obtain an embedding of the data points into an Euclidean space while preserving the kernel's qualities, which are based on non-linear local qualities of the underlying manifold.

Beside the high-dimensionality of the data, its size (i.e., number of sampled data points) is usually very big. The massive size of the dataset is mostly due to the ease of obtaining data points. For example, most systems nowadays collect detailed logs of every action, event and operation that occur with high frequency over long periods of time. However, most of the collected data points are redundant, either because they are near-duplicates of other already-measured data points, or because their properties can be interpolated by suitable subsets of representatives. Therefore, a combination of subsampling and out-of-sample extension techniques can alleviate performance issues that massive datasets entail, and provide a more suitable representation of the analyzed data. Optimally, such a representation would not be affected by the availability of the data or by a sampling method but only rely on the behavior of the observed and analyzed phenomena.

The kernel approach, which is used for dimensionality reduction, has been applied for the described out-of-sample extension tasks. A classical kernel-based technique is the Nyström extension [14,1]. More recent methods are Geometric Harmonics [6] and the Multiscale Extension in [3]. These methods use the spectral decomposition of the kernel (i.e., its eigenvalues and eigenvectors) as a basis of its range. The eigenfunctions are shown to be easily extended to new data points, thus any function in its range, which can be expressed as a linear combination of these eigenfunctions, is also easily extended. Functions that are not in the range of the kernel are extended by projecting them on the kernel's range and using the resulting function (and extension) as an approximation of the original function.

Kernel methods work under the assumption that the used kernel has a small set of significant eigenvalues that should be considered for the analysis, and the rest are negligible in the sense that they are numerically zero. This can be phrased as a low numerical rank assumption, where the numerical rank is the number of numerically nonzero eigenvalues or singular values (see Definition 2.1 for an explicit formulation). While in practice this assumption is usually satisfied, most papers do not present rigorous mathematical support (beyond intuition) for it.

In this paper, we present upper bounds for the numerical rank of affinity kernels. We focus on Gaussian kernels, which are popular in many spectral kernel methods (e.g. [5,2]). Such an upper bound was achieved in [3] based on a bounding box volume of the analyzed dataset in the observable ambient space. We refine this bound by considering the underlying geometry that is provided by the underlying manifold from which the dataset is assumed to be sampled. Instead of using a single large bounding box, we use a finite set of small constant-volume boxes that cover the dataset (or its underlying manifold), and use the minimal cover to provide a cover-based bound. When the constant size of the boxes is large enough to cover the whole dataset with one box, this bound converges to the one in [3]. Thus, it is at least as tight as this already established one.

The paper has the following structure. The problem setup and a previously-established bound are described in Section 2. The refined cover-based bounds are established in Section 3. Section 4 demonstrates various nuances and concepts of cover-based bounds, as well as their theoretical application for proving relations between the geodesic length of curves and the numerical rank of datasets that are sampled from these curves.

2. Problem setup

Let $\mathcal{M}$ be a low-dimensional compact manifold that lies in the high-dimensional ambient space $\mathbb{R}^m$ that has an Euclidean metric $\|\cdot\|$. In addition, let β be the Borel σ-algebra on $\mathcal{M}$ and let μ be a probability measure on $(\mathcal{M}, \beta)$. Finally, let $M \subseteq \mathbb{R}^m$ be a set of n data points (i.e., $n = |M|$) sampled from the manifold $\mathcal{M}$.

Define the affinity between two data points $x, y \in M$ to be $g_\varepsilon(x, y) = e^{-\|x-y\|^2/\varepsilon}$ where ε is a positive parameter. Let G_ε^M be an $n \times n$ affinity kernel between the data points in M, where each row and each column of G_ε^M corresponds to a single data point in the dataset M, and each cell contains the affinity $g_\varepsilon(x, y)$ between the row's data point $x \in M$ and the column's data point $y \in M$.

The matrix G_ε^M is called the Gaussian kernel over the dataset M. This kernel introduces the notion of affinities and local neighborhoods of data points in the dataset M (or on the manifold $\mathcal{M}$) due to the exponential decay of it's values in

relation to the distances between data points. The Gaussian kernel with its spectral analysis and its spectral decomposition are utilized for dimensionality reduction in [5,7,2] and for out-of-sample function extension in [6,3].

We denote the rank of the Gaussian kernel G_ε^M (i.e., the dimension of its range or, equivalently, the number of its nonzero eigenvalues) by $\rho(G_\varepsilon^M)$. Usually, the kernel will not have strictly nonzero eigenvalues. However, since its spectrum decays rapidly (i.e., exponentially), most of its eigenvalues will have negligible (albeit nonzero) values, and it will only have a limited number of numerically-significant eigenvalues from a practical analysis point of view. Therefore, the algebraic rank is insufficient to characterize the spectral properties of G_ε^M. A more desirable characteristic needs to consider the number of numerically nonzero (i.e., numerically-significant) eigenvalues based on a predetermined significance threshold δ. Definition 2.1 introduces the *numerical rank* of G_ε^M for this purpose.[1] This definition is standard in many papers that use spectral analysis.

Definition 2.1. The *numerical rank* of the Gaussian kernel G_ε^M up to precision $\delta \geqslant 0$ is

$$\rho_\delta\big(G_\varepsilon^M\big) \triangleq \#\left\{ j\colon \frac{\sigma_j(G_\varepsilon^M)}{\sigma_1(G_\varepsilon^M)} \geqslant \delta \right\},$$

where $\sigma_j(G_\varepsilon^M)$ denotes the j-th largest singular value of the matrix G_ε^M.

The numerical rank $\rho_\delta(G_\varepsilon^M)$ determines the dimension of the embedded space that is achieved by dimensionality reduction methods such as Diffusion Maps [5,7] and Laplacian Eigenmaps [2], and the number of harmonics or sampled representatives that are used by out-of-sample methods such as Geometric Harmonics [6] and the Multiscale Extension in [3]. We notice that when the significance threshold δ is zero then the numerical rank converges to the algebraic rank $\rho(G_\varepsilon^M) = \rho_0(G_\varepsilon^M)$. For the rest of the paper, unless mentioned otherwise, we consider the parameters ε and δ to be predetermined and constant, such that $\varepsilon > 0$ and $0 \leqslant \delta < 1$. For clarity, we will refer to the numerical rank $\rho_\delta(G_\varepsilon^M)$ of the Gaussian kernel over the dataset M as the *Gaussian numerical rank* of the dataset M.

2.1. Ambient box-based bounds

The relation between the numerical rank of G_ε^M and the observable ambient space $\mathbb{R}^m$ of the manifold $\mathcal{M}$, from which the dataset M was sampled, is shown in [3]. This relation was expressed by an upper bound on the numerical rank, which was expressed by the volume of a bounding box of the dataset in the ambient space. However, the geometry of the manifold $\mathcal{M}$ is ignored by this bound. In this paper, we refine the bounds achieved in [3] by considering a small set of boxes that cover the manifold and any dataset that is sampled from it. First, we reiterate the results from [3], then, in Section 3, we use these results to prove the new manifold-related bound.

Let $Q \subset \mathbb{R}^m$ be a box in the observable space, where $q_1 \geqslant \cdots \geqslant q_m \in \mathbb{R}$ are the lengths of its sides (listed, without loss of generality, in a descending order). Thus, the volume of Q is $\prod_{i=1}^m q_i$. Let $X \subseteq Q$ be a finite dataset that is contained within the box Q, and let G_ε^X be the Gaussian kernel over this dataset. Then, according to [3], the numerical rank of G_ε^X is bounded from above by

$$\rho_\delta\big(G_\varepsilon^X\big) \leqslant \prod_{j=1}^m \big\{ \lfloor \kappa q_j \rfloor + 1 \big\}, \tag{2.1}$$

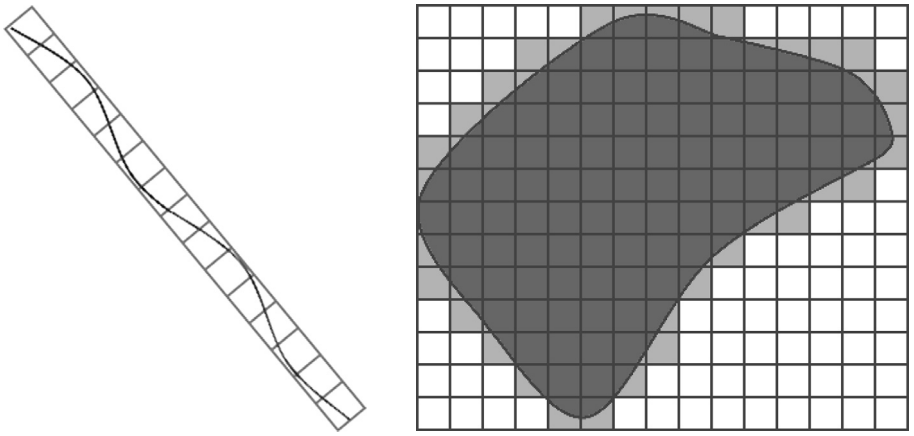
where

$$\kappa \triangleq \frac{2}{\pi}\sqrt{\varepsilon^{-1}\ln\big(\delta^{-1}\big)}. \tag{2.2}$$

We examine now an arbitrary side length q_j $(j = 1, \ldots, m)$ of the box Q. If $q_j < \frac{1}{\kappa}$, then the j-th term of the product in Eq. (2.1) is $\lfloor \kappa q_j \rfloor + 1 = 1$ and the side length q_j does not affect the bound in this equation. A side of Q whose length $q_j < \frac{1}{\kappa}$ $(j = 1, \ldots, m)$, which is too short to affect the bound in Eq. (2.1), is called a *short side*. A side whose length $q_j \geqslant \frac{1}{\kappa}$ does affect this bound is called a *long side*. We call Q a *d-box* if it has exactly $d \leqslant m$ long sides and its other $m - d$ sides are short sides. Since we assumed (without loss of generality) that the side lengths of Q are listed in descending order, then for a d-box we have $q_1 \geqslant \cdots \geqslant q_d \geqslant 1/\kappa > q_{d+1} \geqslant \cdots \geqslant q_m$. Since, in this case, $q_{d+1}, \ldots, q_m$ are short-side lengths that do not affect the bound in Eq. (2.1), then for d-box Q (and any finite dataset $X \subset Q$) the following bound is satisfied:

$$\rho_\delta\big(G_\varepsilon^X\big) \leqslant \prod_{j=1}^d \big\{ \lfloor \kappa q_j \rfloor + 1 \big\}. \tag{2.3}$$

[1] Specifically, for the discussed Gaussian kernel, its singular values and eigenvalues are the same and the definition can be based on either of them equivalently. The presented definition (using singular values) is also valid for any general matrix, and not just for the Gaussian kernel.



(a) A single bounding box is sufficient for a relatively flat manifold.

(b) For a non-flat manifold, a single bounding box is unnecessarily large.

Fig. 3.1. Covering a manifold (or a compact set) using a set of small (e.g., unit-size) boxes vs. a single box with discretized (e.g., integer) side lengths.

3. Cover-based bounds

The bound in Eq. (2.1) is based on a single box that covers the whole dataset (or the whole manifold). The volume (or, more accurately, the product of the discretized side lengths) of this box determines the value of this upper bound. If the dataset is sampled from a flat manifold (e.g., a hyperplane), the long sides of the bounding box can be set on the principal direction of this manifold while the remaining short sides on other directions (see Fig. 3.1(a)). In this case, the bound in Eq. (2.1) considers the intrinsic geometry of the data and measures the volume on the approximately linear area of the manifold from which the data is sampled. However, when the manifold is not flat and contains curved areas, a single box, which contains the whole dataset, is expected to be unnecessarily large (see Fig. 3.1(b)).

Instead of covering the whole dataset (or its underlying manifold) with a single large box, we use a set of small boxes to obtain a cover. Since each box covers a small area on the manifold, and due to the locally low-dimensional nature of the manifolds, each box is expected to have a small number $d \ll m$ of long sides. It is convenient to have all the boxes of approximately the same size by setting a constant length ℓ to their long sides. This way, the size of the cover can be easily determined by ℓ, d and the number of boxes in the cover. Definition 3.1 introduces the type of boxes that will be used to cover a manifold or a dataset that is sampled from it.

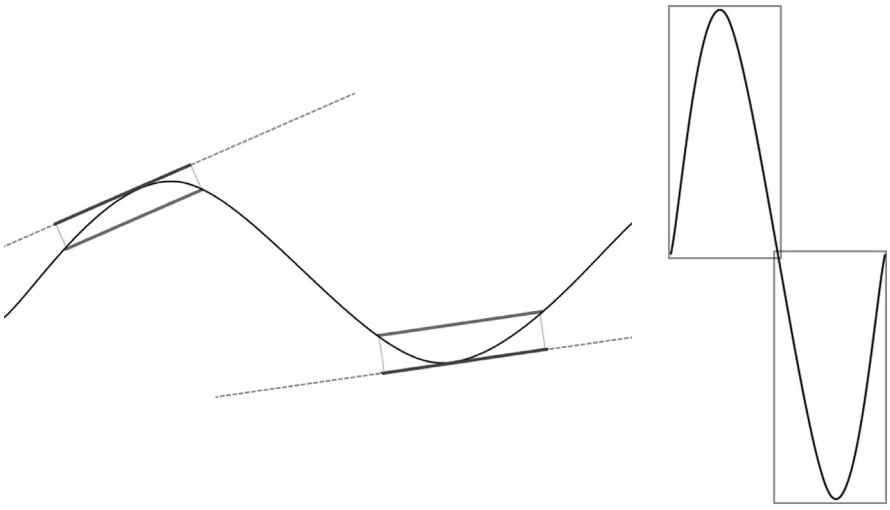
Definition 3.1 *((ℓ, d)-box).* Let $\ell \geqslant \frac{1}{\kappa}$ (where κ is defined in Eq. (2.2)) be a real number and $1 \leqslant d \leqslant m$ be a positive integer. An (ℓ, d)-box in $\mathbb{R}^m$ is a d-box whose length of each of its d long sides is ℓ.

The boxes from Definition 3.1 are the building blocks for the cover that will be used to set a bound on the numerical rank of Gaussian kernels of manifolds and datasets. Definition 3.2 presents this cover for any subset in the ambient space $\mathbb{R}^m$. In particular, it defines the cover for a manifold that lies in this ambient space and for any dataset that is sampled from such a manifold.

Definition 3.2 *((ℓ, d)-cover).* Let C be a finite set of (ℓ, d)-boxes in $\mathbb{R}^m$, $\ell \geqslant \frac{1}{\kappa}$, and a positive integer $1 \leqslant d \leqslant m$. Denote the number of boxes in it by $\#(C)$. The set C is called an (ℓ, d)-cover of an arbitrary set $\mathcal{X} \subset \mathbb{R}^m$ if for every data point $x \in \mathcal{X}$ there is at least one (ℓ, d)-box $Q \in C$ such that $x \in Q$. The size of an (ℓ, d)-cover C is the number $\#(C)$ of boxes in C.

We will use the notation $\mathcal{C}_{(\ell,d)}(\mathcal{X})$ for the set of all (ℓ, d)-covers, $\ell \geqslant \frac{1}{\kappa}$ and a positive integer $1 \leqslant d \leqslant m$, of a subset $\mathcal{X} \subset \mathbb{R}^m$ of the ambient space. Specifically, $\mathcal{C}_{(\ell,d)}(\mathcal{M})$ is the set of all (ℓ, d)-covers of a manifold $\mathcal{M}$ that lies in this ambient space, and $\mathcal{C}_{(\ell,d)}(M)$ is the set of all (ℓ, d)-covers of the dataset M that is sampled from this manifold. When the exact values of ℓ and d are irrelevant, we will use the term *box-cover* for an (ℓ, d)-cover with arbitrary values of the length $\ell \geqslant \frac{1}{\kappa}$ and the integer $1 \leqslant d \leqslant m$. The exact values of ℓ and d will be referred to as the scale of the box-cover. The sets of all box-covers of $\mathcal{X}$, $\mathcal{M}$ and M will be denoted by $\mathcal{C}(\mathcal{X})$, $\mathcal{C}(\mathcal{M})$ and $\mathcal{C}(M)$, respectively.

Definition 3.2 specifies the conditions that define a box-cover of a set in the ambient space. Not all the sets have a box-cover since by definition only a finite number of boxes can be used in it. In this paper, we are only interested in their existence for compact manifolds and finite datasets that are sampled from such manifolds. The existence of box-covers for finite datasets is immediate. Proposition 3.1 shows the existence of box-covers (for every scale) for a compact manifold in $\mathbb{R}^m$.



(a) Low curvatures (b) High curvatures

Fig. 3.2. An illustration of the flexibility of (ℓ, d)-covers: (a) in low-curvature areas, the long sides can be set on tangent directions; (b) in high-curvature areas, they can be set on normal (i.e., orthogonal to the tangent) directions.

Proposition 3.1. *Let $\mathcal{M} \subset \mathbb{R}^m$ be a compact manifold in the ambient space. Then, $\mathcal{C}(\mathcal{M}) \neq \emptyset$, and for every length $\ell \geqslant \frac{1}{\kappa}$ and every integer $1 \leqslant d \leqslant m$, $\mathcal{C}_{(\ell,d)}(\mathcal{M}) \neq \emptyset$.*

Proof. Consider an arbitrary scale (ℓ, d). Surely, we can construct an open (ℓ, d)-box around every data point $x \in \mathcal{M}$ on the manifold. This infinite set of open boxes covers the entire manifold. Since the manifold is compact, there must be a finite subset of these boxes that is sufficient for covering the entire manifold. This set constitutes an (ℓ, d)-cover of $\mathcal{M}$. This argument is valid for every scale that proves the proposition. □

Let $C \in \mathcal{C}(\mathcal{M})$ be a box-cover of the manifold $\mathcal{M}$. Notice that there are no limitations or conditions set on the orientations and positions of the boxes in C. In low-curvature areas, it seems beneficial to set the long sides of the covering boxes to be tangent to the manifold (see Fig. 3.2(a)). However, in high-curvature areas, it might be more efficient (depending on the scale of the box-cover) to set the long sides along the normal of the tangent space (see Fig. 3.2(b)). The definition of the box-covers allows us to use this flexibility to consider efficient coverings of the manifold. Theorem 3.2 introduces an upper bound on the numerical rank of Gaussian kernels on datasets that are sampled from the manifold. Corollary 3.3 extends this result to set a mutual upper bound on *any* dataset that is sampled from the given manifold.

Theorem 3.2. *Let $\mathcal{M} \subset \mathbb{R}^m$ be a manifold in the ambient space $\mathbb{R}^m$ and let $M \subset \mathcal{M}$ be a dataset sampled from this manifold. The numerical rank of the Gaussian kernel G_ε^M over the dataset M is bounded by $\rho_\delta(G_\varepsilon^M) \leqslant \mathrm{r}(M)$, where*

$$\mathrm{r}(M) \triangleq \min\left\{\#(C) \cdot h(\ell, d) \;\middle|\; \ell \geqslant 1/\kappa,\ 1 \leqslant d \leqslant m,\ C \in \mathcal{C}_{(\ell,d)}(M)\right\}, \tag{3.1}$$

and $h(\ell, d) = \prod_{j=1}^{d} \{\lfloor \kappa \ell \rfloor + 1\}$.

Theorem 3.2 shows that any box-cover of a dataset provides an upper bound on the numerical rank of the Gaussian kernel over this dataset. We use the term *cover-based bound* for the upper bounds in the set $\{\#(C) \cdot h(\ell, d) \mid \ell \geqslant 1/\kappa,\ 1 \leqslant d \leqslant m,\ C \in \mathcal{C}_{(\ell,d)}(M)\}$ from Theorem 3.2. We call their minimum $\mathrm{r}(M)$ as the *tightest cover-based bound* of the dataset M. The proof of Theorem 3.2 is based on the subadditivity of the Gaussian numerical rank, which is shown in Section 3.1.

Proposition 3.1 shows that box-covers exist for any compact manifold. Thus, it is reasonable to consider cover-based bounds that are set by the underlying manifold for any dataset that is sampled from it. Such bounds are not dependent on any specific sampling, but rather on the geometry of the analyzed phenomena. Corollary 3.3 extends the result of Theorem 3.2 and introduces the cover-based bounds, as well as the tightest cover-based bound of any compact manifold.

Corollary 3.3. *Let $\mathcal{M} \subset \mathbb{R}^m$ be a manifold in the ambient space $\mathbb{R}^m$. The numerical rank of the Gaussian kernel G_ε^M over any sampled dataset $M \subset \mathcal{M}$ is bounded by $\rho_\delta(G_\varepsilon^M) \leqslant \mathrm{r}(\mathcal{M})$, where*

$$\mathrm{r}(\mathcal{M}) \triangleq \min\left\{\#(C) \cdot h(\ell, d) \;\middle|\; \ell \geqslant 1/\kappa,\ 1 \leqslant d \leqslant m,\ C \in \mathcal{C}_{(\ell,d)}(\mathcal{M})\right\},$$

and $h(\ell, d) = \prod_{j=1}^{d} \{\lfloor \kappa \ell \rfloor + 1\}$.

Proof. The existence of box-covers for the manifold $\mathcal{M}$ is established by Proposition 3.1 and by the definition $\#(C)\cdot h(\ell,d)$ for any $\ell \geqslant 1/\kappa$, $1 \leqslant d \leqslant m$ and $C \in \mathcal{C}_{(\ell,d)}(\mathcal{M})$, is a positive integer. Therefore, the minimum $\mathrm{r}(\mathcal{M})$ exists and is well defined.

Every box-cover of the manifold is also a box-cover of any dataset $M \subset \mathcal{M}$ that is sampled from the manifold. Thus, the set of bounds in the corollary is a subset of the set in Theorem 3.2 and thus $\rho_\delta(G_\varepsilon^M) \leqslant \mathrm{r}(M) \leqslant \mathrm{r}(\mathcal{M})$. □

We call the bound $\mathrm{r}(\mathcal{M})$ in Corollary 3.3 the tightest cover-based bound of the manifold $\mathcal{M}$, and any bound in the set $\{\#(C)\cdot h(\ell,d) \mid \ell \geqslant 1/\kappa,\ 1 \leqslant d \leqslant m,\ C \in \mathcal{C}_{(\ell,d)}(\mathcal{M})\}$ is called a cover-based bound of the manifold. Proposition 3.4 shows that the tightest cover-based bound of the manifold is indeed the tightest cover-based bound of some large enough dataset that is sampled from it. Therefore, no tighter cover-based bound can be set for every possible dataset that is sampled from $\mathcal{M}$.

Proposition 3.4. *Let $\mathcal{M} \subset \mathbb{R}^m$ be a manifold in the ambient space $\mathbb{R}^m$. The tightest cover-based bound of the manifold satisfies* $\mathrm{r}(\mathcal{M}) \geqslant \max\{\mathrm{r}(M) \mid M \subseteq \mathcal{M},\ |M| < \infty\}$.

Proof. Surely, every box-cover of the manifold $\mathcal{M}$ also covers any subset of the manifold. Specifically, it is true for every finite dataset that is sampled from it. Therefore, we must have $\mathrm{r}(\mathcal{M}) \geqslant \mathrm{r}(M)$ for every dataset $M \subseteq \mathcal{M}$ and the weak inequality in the proposition is proved. The existence of the maximum is due to the discreteness of the tightest cover-based bounds. □

Proposition 3.4 justifies the name 'tightest cover-based bound' that we used for $\mathrm{r}(\mathcal{M})$ by showing that it indeed serves as a maximal tightest cover-based bound for all the finite sampled datasets from the manifold. Section 4.1 provides examples for equality and strict inequality cases. In addition to examining finite datasets and defining Gaussian kernel matrices over them, we can also define a continuous Gaussian kernel operator $G_\varepsilon^{\mathcal{M}} : C(\mathcal{M}) \to C(\mathcal{M})$ over the whole manifold $\mathcal{M}$. This operator is defined by

$$G_\varepsilon^{\mathcal{M}} f(x) = \int_{\mathcal{M}} g_\varepsilon(x,y) f(y)\, d\mu(y), \quad f : \mathcal{M} \to \mathbb{R},\ x \in \mathcal{M}, \tag{3.2}$$

and it represents the affinities between all the data points on the manifold.

Due to the compactness of $\mathcal{M}$ and the continuity of g_ε, then according to the Hilbert–Schmidt theorem, the Gaussian kernel operator $G_\varepsilon^{\mathcal{M}}$ has a discrete set of real eigenvalues that forms a decaying spectrum [5,6], which is similar to the spectrum of Gaussian kernel matrices over datasets that are sampled from the manifold. Therefore, we can also examine the numerical rank of this operator that considers the manifold itself instead of considering a finite sampling of data points from it. Theorem 3.5 shows that the tightest cover-based bound $\mathrm{r}(\mathcal{M})$ also serves as an upper bound for the Gaussian numerical rank $\rho_\delta(G_\varepsilon^{\mathcal{M}})$ of the manifold $\mathcal{M}$ and not only as an upper bound on Gaussian numerical ranks of finite datasets that are sampled from it.

Theorem 3.5. *Let $\mathcal{M} \subset \mathbb{R}^m$ be a compact manifold in the ambient space $\mathbb{R}^m$. The numerical rank of the Gaussian kernel operator $G_\varepsilon^{\mathcal{M}}$ over the manifold $\mathcal{M}$ is bounded by $\rho_\delta(G_\varepsilon^{\mathcal{M}}) \leqslant \mathrm{r}(\mathcal{M})$, where $\mathrm{r}(\mathcal{M})$ is the tightest cover-based bound (from Corollary 3.3) of the manifold $\mathcal{M}$.*

Theorem 3.5, which will be proved in Section 3.2, shows that the achieved upper bound of the Gaussian numerical rank is a property of the manifold itself and not just a result of finite samplings of the manifold. In some sense, it also provides an insight for the finite datasets usage to represent properties of the manifold. Together with Proposition 3.4, it shows a relation between the maximal tightest cover-based bound that is achieved by a finite dataset and the upper bound on the Gaussian numerical rank of the continuous manifold itself.

Some implications and nuances of the results in this section are demonstrated on simple manifolds (i.e., curves and surfaces) in Section 4. The rest of this section deals first with proving the two main theorems. In Section 3.1, we prove Theorem 3.2 by showing that the Gaussian numerical rank is subadditive. In Section 3.2, we prove Theorem 3.5 by showing a series of finite matrices whose numerical ranks converge to the numerical rank of the continuous operator in Eq. (3.2).

3.1. Subadditivity of the Gaussian numerical rank

Theorem 3.2 is a result of the subadditivity of the numerical rank of Gaussian kernels. In this section, we will prove this property and then prove the theorem by using this result. Lemma 3.7 shows the relation between the numerical rank of Gaussian kernels of two sets and the numerical rank of their Gaussian kernel union. In order to prove it, we first show a technical result in Lemma 3.6, about the relations between the numerical rank and the algebraic rank of principal submatrices.

Lemma 3.6. *Let $G \in \mathbb{C}^{n\times n}$ be a nonsingular complex matrix and let $\tilde{G} \in \mathbb{C}^{q\times q}$ $(q < n)$ be a principal submatrix of G. If $\rho_\delta(G) = \rho(G)$ then $\rho_\delta(\tilde{G}) = \rho(\tilde{G})$.*

Proof. If $\rho_\delta(G)=\rho(G)$ then, by the definition of the numerical rank, $\frac{\sigma_n(G)}{\sigma_1(G)}\geqslant\delta$. From Cauchy's interlacing theorem [18] we get $\sigma_q(\tilde{G})\geqslant\sigma_n(G)$ and $\sigma_1(\tilde{G})\leqslant\sigma_1(G)$, thus, $\frac{\sigma_q(\tilde{G})}{\sigma_1(\tilde{G})}\geqslant\delta$. By using the definition of the numerical rank again, we finally get $\rho_\delta(\tilde{G})=\rho(\tilde{G})$. □

Lemma 3.7. *Let $X=\{x_1,x_2,\ldots,x_{p-1},x_p\}$ and $Y=\{y_1,y_2,\ldots,y_{q-1},y_q\}$ be two sets in $\mathbb{R}^m$. Then, for any $\varepsilon>0$ and $0\leqslant\delta<1$ $\rho_\delta(G_\varepsilon^{X\cup Y})\leqslant\rho_\delta(G_\varepsilon^X)+\rho_\delta(G_\varepsilon^Y)$.*

Proof. Suppose that $\rho_\delta(G_\varepsilon^{X\cup Y})=r$ and let $Z\subset X\cup Y$ be a subset of r data points such that $\rho_\delta(G_\varepsilon^Z)=r$. Additionally, let $\tilde{X}=X\cap Z$ and $\tilde{Y}=Y\cap Z$. According to Bochner's theorem [21], $\rho(G_\varepsilon^Z)=r$. Thus we get, $r=\rho(G_\varepsilon^Z)\leqslant|\tilde{X}|+|\tilde{Y}|$. According to Lemma 3.6, $\rho_\delta(G_\varepsilon^{\tilde{X}})=|\tilde{X}|$ and $\rho_\delta(G_\varepsilon^{\tilde{Y}})=|\tilde{Y}|$. Since $\tilde{X}\subset X$ and $\tilde{Y}\subset Y$, $\rho_\delta(G_\varepsilon^{\tilde{X}})\leqslant\rho_\delta(G_\varepsilon^X)$ and $\rho_\delta(G_\varepsilon^{\tilde{Y}})\leqslant\rho_\delta(G_\varepsilon^Y)$. As a consequence, $\rho_\delta(G_\varepsilon^{X\cup Y})\leqslant\rho_\delta(G_\varepsilon^X)+\rho_\delta(G_\varepsilon^Y)$. □

Lemma 3.7 shows that the Gaussian numerical rank of a union of two sets is at most the sum of their Gaussian numerical ranks. This result can be easily extended to unions of any number of sets by applying Lemma 3.7 as many times as needed. Therefore, we get Corollary 3.8 that states the subadditivity of the Gaussian numerical rank.

Corollary 3.8 *(Subadditivity of the Gaussian numerical rank). Let $X_1,X_2,\ldots,X_q$ be q finite subsets of $\mathbb{R}^m$, and let $M=\bigcup_{j=1}^q X_j$. Then, $\rho_\delta(G_\varepsilon^M)\leqslant\sum_{j=1}^q\rho_\delta(G_\varepsilon^{X_j})$.*

We are now ready to prove Theorem 3.2 by combining Corollary 3.8 and the results from [3]. In essence, each box provides an upper bound on the Gaussian numerical rank of a local subset according to the result from [3], and these bounds can be combined according to Corollary 3.8, thus achieving an upper bound on the Gaussian numerical rank of the whole dataset.

Proof of Theorem 3.2. Let $M\subseteq\mathcal{M}$ be a finite dataset that is sampled from the compact manifold $\mathcal{M}\subseteq\mathbb{R}^m$. Since the dataset is finite, there exists a box-cover $C\in\mathcal{C}_{(\ell,d)}(M)$ for some $\ell\geqslant1/\kappa$ and $1\leqslant d\leqslant m$. By Definition 3.2, the cover-based bound $\#(C)\cdot h(\ell,d)$ for any such box-cover is a positive integer. Therefore, the minimum

$$\mathrm{r}(M)\triangleq\min\left\{\#(C)\cdot h(\ell,d)\;\middle|\;\ell\geqslant1/\kappa,\ 1\leqslant d\leqslant m,\ C\in\mathcal{C}_{(\ell,d)}(M)\right\}\tag{3.3}$$

of a nonempty set of positive integers exists and it is well defined.

By Definition 3.2, any arbitrary (ℓ,d)-cover C of M (for appropriate values of ℓ and d) is a set of (ℓ,d)-boxes $Q_1,\ldots,Q_q$, $q=\#(C)$, such that $M\subseteq Q_1\cup\cdots\cup Q_q$. Therefore, we can define the q sets $M_j\triangleq Q_j\cap M$, $j=1,\ldots,q$, and get that $M=M_1\cup\cdots\cup M_q$ where each set M_j, $j=1,\ldots,q$, is bounded by the corresponding (ℓ,d)-box. Each of these boxes is a d-box where all its long sides have the length ℓ. Therefore, according to Eq. (2.3), the Gaussian numerical rank of every M_j, $j=1,\ldots,q$, is bounded by

$$\rho_\delta\left(G_\varepsilon^{M_j}\right)\leqslant\prod_{i=1}^d\left\{\lfloor\kappa\ell\rfloor+1\right\}=\left(\lfloor\kappa\ell\rfloor+1\right)^d=h(\ell,d),$$

thus, together with Corollary 3.8 we get that the Gaussian numerical rank of $M=M_1\cup\cdots\cup M_q$ is bounded by

$$\rho_\delta\left(G_\varepsilon^M\right)\leqslant\sum_{j=1}^q\rho_\delta\left(G_\varepsilon^{M_j}\right)\leqslant\sum_{j=1}^{\#(C)}h(\ell,d)=\#(C)\cdot h(\ell,d).$$

Therefore, each arbitrary (ℓ,d)-cover $C\in\mathcal{C}_{(\ell,d)}(M)$ provides a cover-based upper bound $\#(C)\cdot h(\ell,d)$ on the Gaussian numerical rank of M. In particular, the tightest (i.e., minimum) cover-based bound $\mathrm{r}(M)$ (see Eq. (3.3)) is indeed an upper bound for this numerical rank as the theorem states. □

Notice that, in fact, the proof of Lemma 3.7 does not rely on any specific inter-subset affinities values of $g_\varepsilon(x,y)$, $x\in X-Y$, $y\in Y-X$, in the context of Lemma 3.7. As a result, both Lemma 3.7 and Corollary 3.8, also apply when these inter-subset affinities are not directly determined. Therefore, one can measure the affinities in local areas on the manifolds, and either ignore (i.e., set to zero) or deduce (e.g., by random-walks or diffusion) the affinities between farther data points. The resulting kernel will still abide by Lemma 3.7 and Corollary 3.8. If the local neighborhoods, in which the affinities are directly measured, are determined by a box-cover that achieves the tightest cover-based bound $\mathrm{r}(M)$, then, according to the proof of Theorem 3.2, the numerical rank of the resulting locally-measured kernel will not exceed the bound $\mathrm{r}(M)$ of the Gaussian numerical rank.

3.2. The Gaussian convolution operator

In this section, we focus on the continuous kernel operator $G_\varepsilon^{\mathcal{M}}$ from Eq. (3.2). This kernel is in fact a Gaussian convolution operator that acts on the manifold $\mathcal{M}$ in $\mathbb{R}^m$. The main goal of this section is to bound its numerical rank (from above) and prove Theorem 3.5. The notations and the techniques in the rest of this section are similar to the ones that were presented in [20]. These notations are slightly different from the rest of this paper, but they are more suitable for the purposes of the following discussion.

Let $\mathcal{M}$ be the manifold defined in Section 2. Assume, without loss of generality, that $\int_{\mathcal{M}} d\mu(x) = 1$. Let $X = \{x_i\}_{i\in\mathbb{N}}$ be a discrete set of data points that are drawn independently from $\mathcal{M}$ according to the probability distribution μ. Let $X_n = \{x_i\}_{i=1}^n$ be the subset consisting of the first n data points in X. We define the empirical measure $\mu_n(\mathcal{M}) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where δ_x is the Dirac delta function centered at $x \in X$. Thus, for any function $f : \mathcal{M} \to \mathbb{R}$ we have $\int_{\mathcal{M}} f(x)\, d\mu_n(x) = \frac{1}{n}\sum_{i=1}^n f(x_i)$.

Let $(C(\mathcal{M}), \|\cdot\|_\infty)$ be the Banach space of all real continuous functions defined on $\mathcal{M}$ with the infinity norm, and B is the unit ball in this space. Let $g_\varepsilon : \mathcal{M}\times\mathcal{M} \to \mathbb{R}$ be the Gaussian affinity $g_\varepsilon(x,y) = \exp\{-\|x-y\|^2/\varepsilon\}$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^m$. Define the integral operator $G_\varepsilon : C(\mathcal{M}) \to C(\mathcal{M})$ to be the convolution operator $G_\varepsilon f(x) = \int_{\mathcal{M}} g_\varepsilon(x,y) f(y)\, d\mu(y)$. According to Hilbert–Schmidt theorem G_ε, as an operator from $L^2(\mathcal{M},\mu)$ to itself, is a compact operator. Additionally, G_ε is positive-definite, due to Bochner's theorem. Therefore, the spectrum of G_ε consists of isolated eigenvalues. For brevity, since ε is constant throughout this section, we omit the ε subscript. We will call the operator G the *full convolution operator*, as opposed to the partial convolution operators that will be defined later in this section. Notice that the defined operator G is the same as the operator $G_\varepsilon^{\mathcal{M}}$ from Eq. (3.2). We denote the spectrum of the operator G by $\sigma(G)$.

For every positive integer $n \in \mathbb{N}$ we define an $n\times n$ matrix $\bar{\mathcal{G}}_n \triangleq \frac{1}{n} G^{X_n}$, where G^{X_n} is the Gaussian kernel matrix over the dataset X_n. We also define for $n \in \mathbb{N}$ the partial convolution operator $G_n : C(\mathcal{M}) \to C(\mathcal{M})$ that computes the convolution over the data points in X_n instead of computing it over the whole manifold as done for $G_n f(x) \triangleq \int_{\mathcal{M}} g_\varepsilon(x,y) f(y)\, d\mu_n(y)$. Finally, we define the restriction operator $\mathcal{R}_n : C(\mathcal{M}) \to \mathbb{R}^n$ to be $\mathcal{R}_n(f) \triangleq (f(x_1), f(x_2), \ldots, f(x_n))^T$. We will use these constructions to show the relation between the Gaussian numerical rank of a manifold and the Gaussian numerical rank of finite datasets that are sampled from it. Proposition 3.9 shows the relations between the defined constructions.

Proposition 3.9. *The operators G and G_n, $n \in \mathbb{N}$, are compact, uniformly bounded in $(C(\mathcal{M}), \|\cdot\|_\infty)$, and $\bar{\mathcal{G}}_n \mathcal{R}_n = \mathcal{R}_n G_n$.*

Proof. Since the dimension of the range of G_n is finite then G_n is compact for any $n \in \mathbb{N}$. In order to prove that G is compact, we will prove that for any sequence of functions $\{f_n\}_{n\in\mathbb{N}} \subset B$, the sequence $\{Gf_n\}_{n\in\mathbb{N}}$ is relatively compact. Due to Arzela–Ascoli Theorem (e.g., Section I.6 in [15]), it suffices to prove that the set $\{Gf_n\}_{n\in\mathbb{N}}$ is pointwise bounded and equicontinuous. Since $\|g_\varepsilon\|_\infty = 1$, $\|f_n\|_\infty = 1$ and $\mu(\mathcal{M}) = 1$, we get $\|Gf_n\|_\infty = \|\int_{\mathcal{M}} g_\varepsilon(x,y) f_n(y)\, d\mu(y)\|_\infty \leqslant 1$, namely the set $\{Gf_n\}_{n\in\mathbb{N}}$ is pointwise bounded. In addition,

$$\begin{aligned}\left|Gf_n(x) - Gf_n(x')\right| &= \left|\int_{\mathcal{M}} \left(g_\varepsilon(x,y) - g_\varepsilon(x',y)\right) f_n(y)\, d\mu(y)\right| \\ &\leqslant \left\|g_\varepsilon(x,\cdot) - g_\varepsilon(x',\cdot)\right\|_\infty \\ &\leqslant \sqrt{\frac{2}{e\varepsilon}}\left\|x - x'\right\|.\end{aligned}$$

This proves the equicontinuity of $\{Gf_n\}$, which completes the proof of compactness of G.

It remains to show that G and G_n, $n \in \mathbb{N}$, are uniformly bounded;

$$\begin{aligned}\|G_n\|_\infty &= \sup_{f\in B} \|G_n f\|_\infty \\ &= \sup_{f\in B,\, x\in\mathcal{M}} \left|\frac{1}{n}\sum_{i=1}^n g_\varepsilon(x,x_i) f(x_i)\right| \\ &\leqslant 1.\end{aligned}$$

Due to the first part of the proof, $\|G\|_\infty \leqslant 1$. Therefore, G and G_n, $n \in \mathbb{N}$, are uniformly bounded by 1. The last part of the lemma is a direct result from the definitions of $\bar{\mathcal{G}}_n$, $\mathcal{R}_n$ and G_n. □

The definition of the numerical rank of a compact self-adjoint operator G is identical to the definition of the numerical rank on matrices (see Definition 2.1), where instead of singular values we use eigenvalues.[2] For diagonalizable operators, and

[2] The singular values and eigenvalues of Gaussian kernel matrices are anyway equal. Therefore, the results achieved for them in this paper are also valid when using this eigenvalue-based definition.

specifically for compact self-adjoint operators, the numerical rank is the dimensionality of the significant eigen-subspaces, namely, the subspaces that correspond to the significant eigenvalues. Therefore, Definition 3.3 is an equivalent definition of the numerical rank definition of a compact operator G. We use the term *Gaussian numerical rank of a manifold* $\mathcal{M}$ to denote the numerical rank of the Gaussian convolution operator that acts on that manifold.

Definition 3.3. Let G be a compact operator in a Banach space. The numerical rank of G up to precision $\delta \geqslant 0$ is

$$\rho_\delta(G) \triangleq \sum_{\lambda \geqslant \delta \lambda_{\max}} \dim(\mathrm{proj}_\lambda G), \tag{3.4}$$

where $\lambda_{\max}$ is the largest eigenvalue of G, $\mathrm{proj}_\lambda G$ is the projection operator on the eigenspace corresponding to λ, and $\dim(\mathrm{proj}_\lambda G)$ is the dimension of this eigenspace.

Our goal is to prove that the Gaussian numerical rank $\rho_\delta(G)$ of a manifold $\mathcal{M}$ is bounded by $\rho_\delta(G) \leqslant \mathrm{r}(\mathcal{M})$. For this purpose, we take a linear-operator approximation approach. First, in Section 3.2.1, we prove that $\rho_\delta(G_n) = \rho_\delta(\bar{\mathcal{G}}_n)$ for any $n \in \mathbb{N}$. Therefore, due to Proposition 3.4, the numerical rank of each partial convolution operator is bounded by $\rho_\delta(G_n) \leqslant \mathrm{r}(\mathcal{M})$. Then, in Section 3.2.2, we show that the full convolution operator G is the limit operator of the partial convolution operators $\{G_n\}_{n\in\mathbb{N}}$ and as a consequence $\rho_\delta(G_n) \to \rho_\delta(G)$, which completes the proof.

3.2.1. The numerical rank of G_n, $n \in \mathbb{N}$

Due to Bochner's theorem, the matrix $\bar{\mathcal{G}}_n$ is strictly positive-definite, hence all its eigenvalues are positive. Lemma 3.10 shows that $\bar{\mathcal{G}}_n$ and G_n have the same nonzero eigenvalues with the same geometric multiplicities.

Lemma 3.10. *The following relations between the eigen-systems of the matrix $\bar{\mathcal{G}}_n$ and the partial convolution operator G_n are satisfied*:

1. *Let $v = (v_1, v_2, \ldots, v_n)^t$ be an eigenvector of $\bar{\mathcal{G}}_n$ that corresponds to an eigenvalue λ. Then, the continuous function $f_v : \mathcal{M} \to \mathbb{R}$, defined by $f_v(x) = \frac{1}{n\lambda}\sum_{j=1}^n k(x, x_j) v_j$ is an eigenfunction of G_n, corresponding to the same eigenvalue λ.*
2. *If f is an eigenfunction of G_n that corresponds to an eigenvalue λ then $\mathcal{R}_n f$ is an eigenvector of g_n that corresponds to the same eigenvalue λ.*
3. *Let λ be an eigenvalue of $\bar{\mathcal{G}}_n$ with the geometric multiplicity m. Then, the geometric multiplicity of λ as an eigenvalue of G_n is m.*
4. *$\rho_\delta(G_n) = \rho_\delta(\bar{\mathcal{G}}_n)$ for any $n \in \mathbb{N}$.*

Proof. 1. Since $\bar{\mathcal{G}}_n v = \lambda v$, then $\lambda f_v(x_i) = \frac{1}{n}\sum_{j=1}^n g_\varepsilon(x_i, x_j) v_j = \lambda v_i$ for all $i = 1, 2, \ldots, n$. Therefore,

$$\begin{aligned} G_n f_v(x) &= \frac{1}{n}\sum_{i=1}^n g_\varepsilon(x, x_i) f_v(x_i) \\ &= \frac{1}{n}\sum_{i=1}^n \left[g_\varepsilon(x, x_i) \cdot \frac{1}{n\lambda}\sum_{j=1}^n g_\varepsilon(x_i, x_j) v_j \right] \\ &= \frac{1}{n}\sum_{i=1}^n g_\varepsilon(x, x_i) v_i = \lambda f_v(x). \end{aligned}$$

2. If $G_n f = \lambda f$ then, due to Proposition 3.9, $\bar{\mathcal{G}}_n \mathcal{R}_n f = \mathcal{R}_n G_n f = \lambda \mathcal{R}_n f$.

3. Let $v_1, \ldots, v_m$ be a basis for the eigenspace of $\bar{\mathcal{G}}_n$ that corresponds to the eigenvalue λ. Since $v_1, \ldots, v_m$ are linearly independent, then the functions $f_{v_1}, \ldots, f_{v_n}$ are linearly independent. Therefore, $\dim(\mathrm{proj}_\lambda G_n) \geqslant \dim(\mathrm{proj}_\lambda \bar{\mathcal{G}}_n)$. Since the ranges of G_n and $\bar{\mathcal{G}}_n$ are both of dimension n, we get $\dim(\mathrm{proj}_\lambda G_n) = \dim(\mathrm{proj}_\lambda \bar{\mathcal{G}}_n)$ for any nonzero eigenvalue λ.

4. The equality $\rho_\delta(G_n) = \rho_\delta(\bar{\mathcal{G}}_n)$ is a direct consequence of the above. □

Corollary 3.11 is an immediate result of Theorem 3.2 and Lemma 3.10. This proposition provides an upper bound for the numerical rank of the partial convolution operators G_n, $n \in \mathbb{N}$. This bound will be used in Section 3.2.2 to provide an upper bound for the numerical rank of the full convolution operator G.

Corollary 3.11. *The numerical rank of G_n, for any $n \in \mathbb{N}$, is bounded by $\rho_\delta(G_n) \leqslant r(\mathcal{M})$.*

3.2.2. The numerical rank of G

In this section, we prove that the sequence $\{G_n\}_{n\in\mathbb{N}}$ converges to G compactly as defined in Definition 3.4. Proposition 3.12 shows that this convergence also guarantees the convergence of the corresponding eigenspaces of the sequence $\{G_n\}_{n\in\mathbb{N}}$ to those of G.

Definition 3.4 *(Convergence of operators).* Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a Banach space, B its unit ball and $\{S_n\}_{n\in\mathbb{N}}$ is a sequence of bounded linear operators on $\mathcal{F}$:

- The set $\{S_n\}_{n\in\mathbb{N}}$ converges pointwise, denoted by $S_n \xrightarrow{p} S$, if $\|S_n f - Sf\|_{\mathcal{F}} \to 0$ for all $f \in \mathcal{F}$.
- The set $\{S_n\}_{n\in\mathbb{N}}$ converges compactly, denoted by $S_n \xrightarrow{c} S$, if $S_n \xrightarrow{p} S$ and if for every sequence $\{f_n\}_{n\in\mathbb{N}}$ in B, the sequence $\{(S - S_n)f_n\}_{n\in\mathbb{N}}$ is relatively compact (has a compact closure) in $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$.

Proposition 3.12. *(See Proposition 6 in [20].) Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a Banach space, and $\{S_n\}_{n\in\mathbb{N}}$ and S are bounded linear operators on $\mathcal{F}$ such that $S_n \xrightarrow{c} S$. Let $\lambda \in \sigma(S)$ be an isolated eigenvalue with finite multiplicity m, and $M \subset \mathbb{C}$ an open neighborhood of λ such that $\sigma(S) \cap M = \{\lambda\}$. Then:*

1. *Convergence of eigenvalues: There exists an $N \in \mathbb{N}$ such that, for all $n > N$ the set $\sigma(S_n) \cap M$ is an isolated part of $\sigma(S_n)$ that consists of at most m different eigenvalues, and their multiplicities sum up to m. Moreover, the sequence of the sets $\sigma(S_n) \cap M$ converges to the set $\{\lambda\}$ in the sense that every sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ with $\lambda_n \in \sigma(S_n) \cap M$ satisfies $\lim_{n\to\infty} \lambda_n = \lambda$.*
2. *Convergence of spectral projections: Let Pr be the spectral projection of S that corresponds to λ, and for $n > N$, let Pr_n be the spectral projection of S_n that corresponds to $\sigma(S_n) \cap M$. Then, $Pr_n \xrightarrow{p} Pr$.*

Lemma 3.13. *The full and partial convolution operators satisfy $G_n \xrightarrow{p} G$ in $(C(\mathcal{M}), \|\cdot\|_\infty)$.*

Proof. Let $f \in C(\mathcal{M})$; then

$$\begin{aligned}\|G_n f - Gf\|_\infty &= \sup_{x\in\mathcal{M}} \left| \int_{\mathcal{M}} g_\varepsilon(x,y) f(y)\, d\mu_n(y) - \int_{\mathcal{M}} g_\varepsilon(x,y) f(y)\, d\mu(y) \right| \\ &= \sup_{x\in\mathcal{M}} \left| \frac{1}{n}\sum_{i=1}^{n} g_\varepsilon(x,x_i) f(x_i) - \mathbb{E}\big(g_\varepsilon(x,\cdot) f(\cdot)\big) \right|,\end{aligned}$$

where $\mathbb{E}(g_\varepsilon(x,\cdot) f(\cdot))$ is the expected value of $g_\varepsilon(x,y) f(y)$ as a function of y for a fixed x. As $n \to \infty$, this expression converges to zero due to the uniform law of large numbers, and therefore the convergence in the lemma is proved. □

Lemma 3.14. *The partial and the full convolution operators satisfy $G_n \xrightarrow{c} G$ in $(C(\mathcal{M}), \|\cdot\|_\infty)$.*

Proof. Due to Lemma 3.13, we already have $G_n \xrightarrow{p} G$. It remains to show that for every sequence $\{f_n\}_{n\in\mathbb{N}}$ in the unit ball B in $C(\mathcal{M})$, the sequence $\{(G - G_n)f_n\}_{n\in\mathbb{N}}$ is relatively compact in $(C(\mathcal{M}), \|\cdot\|_\infty)$. Due to Arzela–Ascoli Theorem, it is suffices to show that $\{(G - G_n)f_n\}_{n\in\mathbb{N}}$ is pointwise bounded and equicontinuous. As for the first property, according to the proof of Proposition 3.9, $\|(G - G_n)f_n\|_\infty \leqslant \|Gf_n\|_\infty + \|G_n f_n\|_\infty \leqslant 2$. The second property is a result of the bounded derivative of the Gaussian function:

$$\begin{aligned}\big|(G - G_n)f_n(x) - (G - G_n)f_n(x')\big| &\leqslant \big|G\big(f_n(x) - f_n(x')\big)\big| + \big|G_n\big(f_n(x) - f_n(x')\big)\big| \\ &= \left| \int_{\mathcal{M}} \big(g_\varepsilon(x,y) - g_\varepsilon(x',y)\big) f_n(y)\, d\mu(y) \right| \\ &\quad + \left| \frac{1}{n}\sum_{i=1}^{n} \big(g_\varepsilon(x,x_i) - g_\varepsilon(x',x_i)\big) f_n(x_i) \right| \\ &\leqslant 2 \max_{y\in\mathcal{M}} \big|g_\varepsilon(x,y) - g_\varepsilon(x',y)\big| \\ &\leqslant 2\sqrt{\frac{2}{\varepsilon e}}\,\|x - x'\|. \qquad \square\end{aligned}$$

Proposition 3.9 shows that the full convolution operator G is compact. This operator is also strictly positive-definite due to Bochner's theorem. Therefore, all the eigenvalues of this operator are positive and isolated. Theorem 3.15 shows the relation between the numerical rank of G and the numerical ranks of the partial convolution operators G_n, $n \in \mathbb{N}$. This theorem is an immediate result of Corollary 3.11, Proposition 3.12 and Lemma 3.14.

Theorem 3.15. *The operators G_n, $n \in \mathbb{N}$, and G satisfy*

$$\lim_{n\to\infty} \rho_\delta(G_n) = \rho_\delta(G).$$

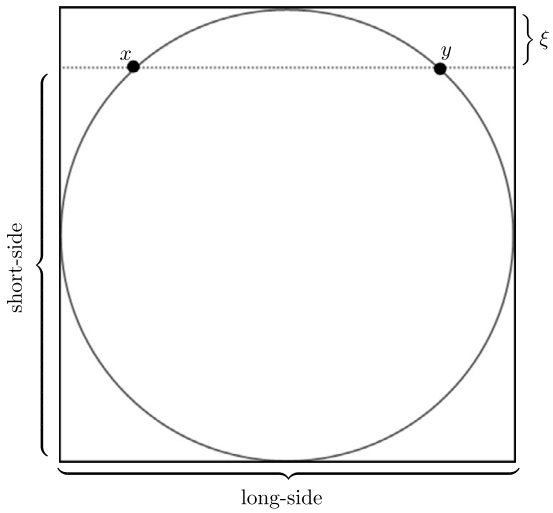


Fig. 4.1. Due to the curvature of the unit-circle, for two adjacent data points x and y in a finite dataset (sampled from the unit-diameter circle) there is a ξ-wide band that is not necessary when only covering the dataset, since there are no data points on the arc between them. This band *is* necessary when the entire (continuous) unit-diameter circle is covered.

Theorem 3.5 essentially states that $\rho_\delta(G) \leqslant r(\mathcal{M})$, which we proceed to prove in this section, is also a direct result of this discussion, and can be considered as a corollary of Theorem 3.15. Therefore, the tightest cover-based bound of the manifold bounds the numerical rank of the affinity kernel operator that considers all the data points on the manifold. This property of the tightest cover-based bound shows that it can be regarded as a property of the manifold itself, and not just a bound for the purpose of analyzing sampled datasets.

4. Examples and discussion

4.1. Strict inequality and equality in Proposition 3.4

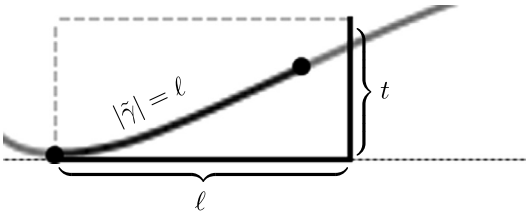
Example 1. *The unit-diameter circle curve* (*strict inequality*). Let the parameters δ and ε have values such that $\kappa = 1$, and consider a circle $\mathcal{M}$ (as a plain curve) with unit-diameter in $\mathbb{R}^2$. The (ℓ, d)-covers have two parameters (ℓ and d) that need to be considered. In this case, there are two possible values for d:

- If $d = 1$, then each box in the cover has one side (i.e., the long side) of length $\ell \geqslant 1$, and the other side (i.e., the short side) is of length $1 - \xi$ (for an arbitrarily small $0 < \xi < 1$) since it has to be strictly less than one. In any case, an $(\ell, 1)$-cover of $\mathcal{M}$ must consist of at least two $(\ell, 1)$-boxes, since the short side of a single box is shorter than the diameter of the circle (see Fig. 4.1). The resulting bound (from Theorem 3.2) in this case is $2 \cdot (\lfloor 1 \cdot \ell \rfloor + 1) \geqslant 4$. On the other hand, for any finite dataset $M \subset \mathcal{M}$, we can select two adjacent data points $x, y \in M$ and set the long side of the box to be parallel to the straight line between x and y as illustrated in Fig. 4.1. We can assume, without loss of generality, that $\ell = 1$ and that ξ is small enough for this single $(1, 1)$-box to cover M, and thus the bound in this case is $1 \cdot (\lfloor 1 \cdot 1 \rfloor + 1) = 2$.
- If $d = 2$, then clearly we can use a single $(1, 2)$-box to form a $(1, 2)$-cover of both M and $\mathcal{M}$, thus the resulting bound is $1 \cdot (\lfloor 1 \cdot 1 \rfloor + 1)^2 = 4$. Any larger value of ℓ will achieve the same (or larger) bound.

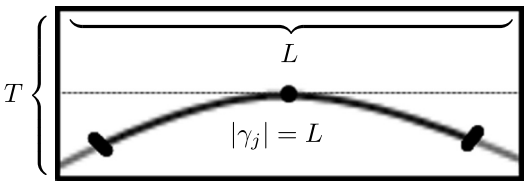
As a consequence of the above, we get $r(M) = 2 < 4 = r(\mathcal{M})$ for any finite dataset $M \subset \mathcal{M}$.

Example 2. *A two-dimensional unit square* (*equality*). Let $\mathcal{M}$ be the unit square curve[3] in $\mathbb{R}^2$. We use the same parameters δ and ε as in Example 1 such that $\kappa = 1$. Using arguments similar to the ones in the previous example, we need at least two $(\ell, 1)$-boxes to cover $\mathcal{M}$, or exactly one $(\ell, 2)$-box, for any $\ell \geqslant 1$. Both resulting bounds are again at least four, so in this case $r(\mathcal{M}) = 4$. Let $M \subset \mathcal{M}$ be the dataset that contains the four corners of the square. This dataset cannot be covered by a single $(\ell, 1)$-box, since its short side must be shorter than one, and therefore the $(\ell, 2)$-covers are anyway similar for the dataset and the manifold in this case, therefore, we can use the same arguments that we used for $\mathcal{M}$ and get $r(M) = 4 = r(\mathcal{M})$.

[3] The manifold in this example is not differentiable at the four corners of the square, but the corners of the square can be slightly rounded by conformal mapping to become smooth in a way that preserves the validity of the presented results.



(a) The relation from Proposition 4.1 between the arc-length ℓ and the bounding $\ell \times t$ box.



(b) A single local bounding box (of the curve section γ_j, $j = 1, \ldots, k$) from the box-cover in Corollary 4.2.

Fig. 4.2. Illustrations of the relations that are used to provide the length-based bound of the Gaussian numerical rank of plain curves.

4.2. Cover-based bounds of plain curves

In this section, we examine the curves (i.e., one-dimensional manifolds) in a two-dimensional ambient plane $\mathbb{R}^2$. We apply the cover-based methodology to introduce the relation between the Gaussian numerical rank of a curve (or datasets sampled from it) and its geodesic arc-length. Specifically, we show that the Gaussian numerical rank of datasets that are sampled from a finite-length curve is bounded by a function of its length.

Proposition 4.1, which is illustrated in Fig. 4.2(a), presents a relation between the geodesic length of a curvature-bounded curve section $\tilde{\gamma}$ and the dimensions of a tangent bounding box of that section. The presented relation provides a method to determine the size of the local boxes that can be used to construct a box-cover of the entire curve.

Proposition 4.1. *Let $\gamma \subseteq \mathbb{R}^2$ be a smooth plain curve and let t and r be positive constants such that $t \leqslant r$. Let $\tilde{\gamma}$ be a section of γ with arc-length*

$$|\tilde{\gamma}| = \ell = r \arccos\left(1 - \frac{t}{r}\right). \tag{4.1}$$

Let $\tilde{\gamma}(s)$, $0 \leqslant s \leqslant \ell$, be an arc-length parametrization of $\tilde{\gamma}$ and assume that the curvature $c(s)$ is bounded from above by $\frac{1}{r}$. Then, the section $\tilde{\gamma} \subseteq \mathbb{R}^2$ can be bounded in a two-dimensional box whose dimensions are $\ell \times t$.

Proof. Suppose that $\gamma : \mathbb{R} \to \mathbb{R}^2$ is parameterized by arc-length such that $\gamma(s) = \tilde{\gamma}(s)$ for $0 \leqslant s \leqslant \ell$. Let $\{e_1, e_2\}$ be the standard coordinates system for $\mathbb{R}^2$ such that $\gamma(0) = 0$ and the derivative $\gamma'(0) = e_1$. Let $\gamma(s) = (x(s), y(s))$ be the parametrization of γ in these coordinates, i.e., $x(s)$ and $y(s)$ are the orthogonal projections of $\gamma(s)$ on e_1 and e_2, respectively. Let $\theta : [0, \ell] \to [0, 2\pi)$, $\theta(s) = \arctan(\frac{y'(s)}{x'(s)})$ be the angle that $\gamma'(s)$ makes with e_1. Thus (see [4]), $\theta'(s) = c(s)$ and $y'(s) = \sin(\theta(s))$ or, equivalently, $y(s) = \int_0^s \sin(\theta(s))\, ds$ and $\theta(s) = \int_0^s c(z)\, dz \leqslant \frac{s}{r}$ for any $0 \leqslant s \leqslant \ell$. Thus, due to Eq. (4.1), we get

$$\begin{aligned} y(\ell) &= \int_0^\ell \sin\big(\theta(s)\big)\, ds \leqslant \int_0^\ell \sin\left(\frac{s}{r}\right) ds \\ &= r - r\cos\left(\frac{\ell}{r}\right) = t. \end{aligned}$$

Obviously, $x(\ell) \leqslant \ell$, therefore, γ can be bounded in an $\ell \times t$ box. □

Corollary 4.2, which is illustrated in Fig. 4.2(b), uses Proposition 4.1 to provide a relation between the geodesic length of a finite-length curve and its Gaussian numerical rank. Specifically, it shows that this Gaussian numerical rank is bounded in proportion to the arc-length of the curve.

Corollary 4.2. *Let t and r satisfy the conditions of Proposition 4.1, such that $t \leqslant \frac{1}{2\kappa}$ and let $\frac{1}{\kappa} \leqslant L = 2r\arccos(1-\frac{t}{r})$. Assume that γ is a plain curve of finite length $|\gamma|$ whose curvature is bounded from above by $\frac{1}{r}$. Then, for any finite configuration $X \subset \gamma$, $\rho_\delta(G_\varepsilon^X) \leqslant h(L,1)\cdot\lceil\frac{|\gamma|}{L}\rceil$.*

Proof. Divide γ to $k = \lceil\frac{|\gamma|}{L}\rceil$ sub-curves such that $\gamma = \bigcup_{j=1}^{k}\gamma_j$ where each is of length L except, perhaps, γ_k. Let $T = 2t$. For each sub-curve γ_j, construct an $L \times T$ bounding box B_j, whose center c_j is the midpoint of γ_j, such that its long side is parallel to $\gamma'(c_j)$. This construction is possible due to Proposition 4.1 since $t \leqslant \frac{1}{2\kappa}$ and $L \geqslant \frac{1}{\kappa}$, $\bigcup_{j=1}^{k} B_j$ constitutes an $(l,1)$-cover of X. Therefore, according to Theorem 3.2, for any finite configuration $X \subset \gamma$, $\rho_\delta(G_\varepsilon^X) \leqslant h(L,1)\cdot\lceil\frac{|\gamma|}{L}\rceil$. □

It should be noted that extending these results to volumes of higher-dimensional manifolds (e.g., geodesic areas of surfaces) is not trivial. This type of analysis depends on the exact volume form of the manifold and is beyond the scope of this paper. However, in practical cases, manifold characterizations in general, and its volume form specifically, are anyway not known. From a practical data point of view, the box-covers used in this paper provide a sufficient volume metric that incorporates the low-dimensional locality nature of the manifold together with possible high-curvature singularities and noisy sampling techniques.

4.3. Discussion

In many cases, although not in all of them, the subadditivity of the Gaussian numerical rank, which is presented in Corollary 3.8, enables to provide a much tighter bound than the one presented in [3]. This bound considers the intrinsic dimensionality of the data, rather than its extrinsic dimensionality.

For example, consider a dataset that was sampled from a one-dimensional square-shaped manifold, whose side length is q, embedded in the real plane. Then, the bound on the Gaussian numerical rank provided by [3] is, due to Eq. (2.1), quadratic in q (i.e., $(\lfloor\kappa q\rfloor+1)^2$). On the other hand, by covering the data with four $(q,1)$-boxes, a linear bound is provided by Corollary 3.8 (i.e., $4(\lfloor\kappa q\rfloor+1)$). This bound is tighter than the quadratic one for sufficiently large q (i.e., $q > 4/\kappa$).

In any case, the definition of the proposed bound $r(M)$ (Eq. (3.1)) considers all the (ℓ,d)-covers of the data, including single-box covers. As such, this bound is at least as tight as the bound presented in [3].

5. Conclusion

In this paper we presented a relation between the numerical rank of Gaussian affinity kernels of low-dimensional manifolds (and datasets that are sampled from them) and the local-geometry of these manifolds. Specifically, we introduced an upper bound for this numerical rank based on the properties of a box-cover of the manifold. The used cover is based on a set of small boxes that contain local areas of the manifold. Together, this set of boxes incorporates the non-linear nature of the manifold while coping with varying curvatures and possible sampling noise.

The presented relation validates one of the fundamental assumptions in kernel-based manifold-learning techniques that local low-dimensionality of the underlying geometry yields a low numerical rank of the used affinities, thus, spectral analysis of these affinities provides a dimensionality reduction of the analyzed data. The results in this paper support this assumption by showing that, in the Gaussian affinity case, its numerical rank is indeed bounded by properties of the underlying manifold geometry.

Acknowledgments

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10) and by the Israeli Ministry of Science & Technology 3-9096. The second author was also supported by the Eshkol Fellowship from the Israeli Ministry of Science & Technology.

References

[1] C.T.H. Baker, The Numerical Treatment of Integral Equations, Clarendon Press, Oxford, 1977.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[3] A. Bermanis, A. Averbuch, R.R. Coifman, Multiscale data sampling and function extension, Appl. Comput. Harmon. Anal. 34 (1) (2013) 15–29.
[4] M.P. Do Carmo, Differential Geometry of Curves and Surfaces, Prentice Hall, 1976.
[5] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30, Diffusion Maps and Wavelets.
[6] R.R. Coifman, S. Lafon, Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions, Appl. Comput. Harmon. Anal. 21 (1) (2006) 31–52, Diffusion Maps and Wavelets.
[7] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proc. Natl. Acad. Sci. USA 102 (21) (2005) 7426–7431.
[8] T. Cox, M. Cox, Multidimensional Scaling, Chapman and Hall, London, UK, 1994.
[9] D.L. Donoho, C. Grimes, Hessian eigenmaps: New locally linear embedding techniques for high dimensional data, Proc. Natl. Acad. Sci. USA 100 (2003) 5591–5596.
[10] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933).

[11] I.T. Jolliffe, Principal Component Analysis, Springer, New York, NY, 1986.
[12] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–27.
[13] S. Mika, B. Schölkopf, A. Smola, K. -R. Müller, M. Scholz, G. Rätsch, Kernel pca and de-noising in feature spaces, in: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, MIT Press, Cambridge, MA, USA, 1999, pp. 536–542.
[14] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, 2nd edition, Cambridge Univ. Press, 1992.
[15] M. Reed, B. Simon, Functional Analysis, vol. I, 2nd edition, Academic Press, New York, 1980.
[16] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[17] B. Schölkopf, A.J. Smola, K. -R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[18] G.W. Stewart, J. Sun, Matrix Perturbation Theory, Academic Press, 1990.
[19] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[20] U. Von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, Ann. Statist. 36 (2008) 555–586.
[21] H. Wendland, Scattered Data Approximation, Cambridge Univ. Press, 2005.
[22] G. Yang, X. Xu, J. Zhang, Manifold alignment via local tangent space alignment, in: International Conference on Computer Science and Software Engineering, December 2008.
[23] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, Technical Report CSE-02-019, Department of Computer Science and Engineering, Pennsylvania State University, 2002.