Algorithms, Geometry and Learning

Reading group
Paris Syminelakis

October 11, 2016

Contents

Chapter 1

Local Dimensionality Reduction

Introduction

Many algorithms operate on data (inputs) that are explicitly (e.g. $\mathbb{R}^d$) or implicitly (e.g. weighted graph) embedded in a metric space. The *curse of dimensionality* refers to the phenomenon that many problems or algorithms require exponential amount of resources (time, space, samples) in the "dimension" of the input. Researchers aiming to go beyond this barrier either make assumptions about the "intrinsic dimension" of the data (impose some further structural restriction) or settle for relaxing the guarantees that the algorithms enjoy.

In the context of problems defined on metric spaces, researchers search for either embeddings into simpler metric spaces with more structure (ℓ_p spaces, ultrametrics etc.) and/or reduced dimension while preserving distances to some extent. This is formalizing by the notion of *distortion.*

Definition 1.1.1 (Distortion)**:** An embedding $f : \mathcal{X} \to \mathcal{Y}$ has distortion $Q > 1$ if there exists a constant $c > 0$ such that $\forall u, v \in \mathcal{X}$

$$c \cdot d_{\mathcal{X}}(u, v) \leq d_{\mathcal{Y}}(f(u), f(v)) \leq cQ \cdot d_{\mathcal{X}}(u, v) \tag{1.1}$$

Compactly, we have $\mathrm{dist}(f) = \frac{\sup_{u,v\in\mathcal{X}} \mathrm{dist}_f(u,v)}{\inf_{u,v\in\mathcal{X}} \mathrm{dist}_f(u,v)}$, where $\mathrm{dist}_f(u, v) := \frac{d_{\mathcal{Y}}(f(u),f(v))}{d_{\mathcal{X}}(u,v)}$.

Having formalized, the notion of distortion we are ready to formalize the problem of *dimensionality reduction.*

Given a metric space (V, d) find a mapping $\phi : V \to \ell_p^d$ with **"small" distortion** and **"low" dimension**

.

There is a great body of work that deals with the exact trade-offs between distortion and dimension, depending also on the structure of the metric space with applications in optimization (approximation algorithms), networking (distance oracles), machine learning (feature extraction, kernel approximation) and indexing (approximate nearest neighbor search).

Euclidean Space - The JL lemma The simplest metric space is arguable the Euclidean space $\mathbb{R}^d$ with the $\|\cdot\|_2$ metric. Its simplicity arises from the vector space and inner product structures along with resulting rotational invariance. For this metric space we now have a complete picture of what is possible and the exact (up to) multiplicative constant tradeoff between distortion and dimension.

Theorem 1.1.2. *For any n-point euclidean space $(V, \|\cdot\|_2)$ and any $\epsilon > 0$:*

- *there exists a map $f : \mathbb{R}^d \to \mathbb{R}^D$ with $D = O(\epsilon^{-2} \log n)$ such that $\forall x, y \in V$*

$$(1-\epsilon)^2 \|x\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1+\epsilon)^2 \|x\|_2^2 \tag{1.2}$$

- *For any $d, n \geq 2$ and $1/\min\{n, d\}^{0.4999} < \epsilon < 1$, there exists a subset V_{LN} of $\mathbb{R}^d$ such that for any map $f : V_{LN} \to \mathbb{R}_2^D$ satisfying (1.2), it must be true that*

$$D = \Omega(\epsilon^{-2} \log n)$$

Distributional JL-Lemma Interestingly, there exist many ways to achieve the guarantees given by the above theorem. However, the simplest one is to generate a random gaussian matrix of appropriate dimensions and define f to be the resulting linear map. One can easily prove the first part of the theorem if one has a so called *Distributional JL-Lemma.*

Lemma 1.1.3. *For any integer $n \geq 2$ and $\epsilon, \delta \in (0, 1)$ there exists a distribution $\mathcal{D}_{\epsilon,\delta}$ on linear maps $f : \mathbb{R}^d \to \mathbb{R}^D$ with $D = O(\epsilon^{-2} \log(1/\delta))$ such that for any two $x, y \in \mathbb{R}^d$,*

$$\mathbb{P}_{\mathcal{D}_{\epsilon,\delta}} \left((1-\epsilon)^2 \|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1+\epsilon)^2 \|x - y\|_2^2 \right) \geq 1 - \delta \tag{1.3}$$

In fact most attempts of dimensionality reduction in Euclidean space have focused exclusively on linear maps and seeking characterizations of what kind of distributions work is an object of recent intensive theoretical investigation.

General metric spaces Similar investigations have been carried out for general metric spaces and have culminated in a very precise quantitative understanding.

Theorem 1.1.4 (Abraham, Bartal, Neiman'11). *For any $1 \leq p \leq \infty$, and any $\theta > 0$ every n-point metric space:*

- *embeds in ℓ_p with distortion $O(\log^{1+\theta} n)$ in dimension $O(\frac{\log n}{\theta \log \log n})$.*

- *for the metric of a 3-regular expander any map that has distortion at most $O(\log^{1+\theta} n)$ needs $\Omega(\frac{\theta \log n}{\log \log n})$ dimensions.*

Similar results exist that present a distortion dimension-tradeoff, where the dimension of the embedding depends on a notion of "intrinsic" dimension of the metric space (either through different notions of "decomposability" or through the doubling dimension). As we will see shortly these embedding results hinge on two facts:

(i) the existence of a **distributional lemma**, i.e., a randomized map that "succeeds" with good probability.

(ii) a **tensorization scheme** that composes many such maps to achieve guarantees with high probability.

Outlook These results show that in a very general sense we have a pretty good understanding of the quantitative trade-offs involved in obtaining low dimensional representations of metric spaces. Still extensions of the above ideas are constantly looked for.

1. **Relaxed guarantees:** last quarter we saw the notion of *scaling distortion* that is actually used to prove Theorem 1.1.4. This notion can be also used to provide results about average, or ℓ_q-notions of distortion and not worst-case. Scaling distortion essentially implies (perhaps counter-intuitively) that the map has *smaller distortion for points that are "relatively far apart"*, in the sense that there are many other points that are included in any ball around x or y of diameter equal to the distance $d(x, y)$.

2. **Prioritized Embeddings:** another line of work assigns an ordering/priority to vertices and seeks to obtain distortion guarantees for a pair (x, y) depending only on the highest priority of the pair.

3. **Local embeddings:** finally we may seek embeddings that approximate distances between close pairs well and we may not care at all for farther points. *This will be the topic of this talk.*

Definitions and Results

We first formalize the notion of locality we are going to use. A natural way to do so, is to consider for each point x the distance that includes its k-nearest neighbors.

Definition 1.2.1: Given $x \in V$ the k-nearest neighbor set and radius are define as

$$N_k(x) := \{k \text{ nearest neighbors of } x\} \tag{1.4}$$

$$r_k(x) := \min\{r > 0 | B(x, r) \subseteq N_k(x)\} \tag{1.5}$$

Definition 1.2.2: An embedding has *k-local distortion α* if for all $x, y \in V$ such that $y \in N_k(x)$,

$$\frac{1}{\alpha} d(x, y) \leq \|f(x) - f(y)\|_p \leq d(x, y) \tag{1.6}$$

Theorem 1.2.3. *For any n-point metric space (X, d) a parameter $k \leq n$ and an integer p satisfying $p \leq \ln k/2$ there exist an embedding into ℓ_p with k-local distortion $O(\log k/p)$ and dimension $O(e^p \log^2 k)$.*

Local Metric Embedding for Euclidean space In the case of euclidean metrics, we will be able to de better at least in a core neighborhood of $N_k(x)$. Let $r^*_k(x) = \frac{c_1 \epsilon r_k(x)}{\log k}$ and $r^*_k(x, y) = \max\{r^*_k(x), r^*_k(y)\}$.

Theorem 1.2.4. *Let $k \in \mathbb{N}$, given a discrete subset V of $\mathbb{R}^d$, for any $\epsilon > 0$ there exists an embedding $\Phi : V \to \mathbb{R}^D$, where $D = O(\epsilon^{-2} \log k)$ such that:*

$$\|\Phi(x) - \Phi(y)\|_2 \leq (1 + \epsilon)\|x - y\|, \forall x, y \in V \tag{1.7}$$

$$\|\Phi(x) - \Phi(y)\|_2 \geq \frac{1}{1 + \epsilon}\|x - y\|, \text{ if } \|x - y\| \leq \sqrt{\epsilon} \cdot r^*_k(x, y) \tag{1.8}$$

Definition 1.2.5: Let $x \sim y$ denote that $d(x, y) \leq t$. An embedding has *t-proximity* distortion α if for any $x, y \in V$ such that $x \sim y$,

$$\frac{1}{\alpha} d(x, y) \leq \|f(x) - f(y)\|_p \leq d(x, y) \tag{1.9}$$

General Approach

Fix any pair of vertices, $x, y \in V$ we are looking for a sequence of possibly randomized maps $f_1, \ldots, f_T : V \to \mathbb{R}$ such that:

- **Expansion:** each map$|f_t(x) - f_t(y)|^p \leq A_t^p d(x,y)^p$ doesn't "expand" distances by much. Typically $A_t = O(1)$ or even $A_t = 1$ (non-expanding embedding). As we will see this step is the easiest to achieve, because of *triangle inequality.* The main fact used is that for any $r > 0$ and $A \subseteq V$:

$$|\min\{d(x,A), r\} - \min\{d(y,A), r\}| \leq \min\{d(x,y), r\} \tag{1.10}$$

- **Contraction:** each map $|f_t(x) - f_t(y)|^p \geq \ell_t(x,y) \cdot d(x,y)$ "witnesses" the distance between x, y within some accuracy $\rho_t(x,y)$.

How do we produce a sequence of maps that witness all "local" distances?

1. **Distributional lemma:** typically we construct a probability distribution over maps such that we have a lower bound on the contraction with *constant probability.*The way we usually go about constructing these maps, is:
    - **Random Sampling:** construct a map for which we can lower bound its contribution *for all pairs of points* by some easy to use quantity.
    - **Probabilistic Partitions:** produce maps for which we get non-trivial guarantees for a specific set of distances. That entails, grouping distances together and producing a map that has constant probability of witnessing any distance from the group.
2. **Tensorization Scheme:** then we add many such maps together to boost the probability of success. Typically, the number of maps we add contribute to increasing the expansion. So, we want to keep the number of such maps (dimensions) as low as possible. Here there are two approaches:
    - use **Chernoff bounds** and *Union Bound* to argue that for each pair many such maps "succeed" with high probability.
    - use **Lovasz Local Lemma** to show that there is a choice of randomness such that by adding a few maps all distances are "witnessed". Typically, at this step we use some additional structure (growth condition) to control dependencies.

Frechet Embeddings

Given a sequence of numbers $\{w_t\}_{1\leq t\leq T}$ and a sequence of sets $\{W_t\}_{1\leq t\leq T}$, we define Frechet embeddings as the map:

$$F(x) := (f_1(x), \ldots, f_T(x)) = (w_1 d(x, W_1), \ldots, w_{|T|} d(x, W_T)) \tag{1.11}$$

Expansion For any such Frechét embedding we can bound the contraction for any pair $x, y \in V$ as $\|F(x) - F(y)\|_p^p \leq \|w\|_p^p d(x,y)^p$ as for each coordinate:

$$|f_t(x) - f_t(y)| = |w_t||d(x, W_t) - d(y, W_t)| \leq |w_t| d(x,y)$$

Bounding the Contraction

We follow the general strategy and come up with a distributional lemma and a tensorization scheme.

Lemma 1.4.1 (Distributional Lemma). *Let $s = 2^p$ and $T = \{1, \ldots, \log_s n\}$. Fix any $x, y \in V$ and set $\rho_0 = 0$ and $\rho_t := \max\{r_{s^t}(x), r_{s^t}(y)\}$. For all $t \in T$ there exists a distribution $\mathcal{D}_{t,s}$ over sets $W_t \subset V$ such that:*

$$\mathbb{P}(|d(x, W_t) - d(y, W_t)| > \rho_t - \rho_{t-1}) \geq \frac{1}{12s} \tag{1.12}$$

When the above event happens for a set $W_t \sim \mathcal{D}_{t,s}$ we say succeds.

Lemma 1.4.2 (Tensorization Lemma). *Given $R = c \cdot s \log n$ for $c > 0$ a large constant, with probability at least $1 - 1/n$ for all pairs $x, y \in V$ and scales $t \in T$ we have that at least $\frac{R}{24s}$ sets succeed.*

The embedding Let $\{f_t^{(i)}\}_{t,i\leq R}$ be functions as defined above where the superscript denotes independent samples of the sets $W_t \sim \mathcal{D}_{t,s}$. Our embedding will be:

$$F(x) := \bigoplus_{t=1}^{T} \bigoplus_{i=1}^{R} w_t f_t^{(i)}(x) \tag{1.13}$$

Here, the dimension of the embedding is $R \cdot \|w\|_0$, as we may zero out any scales t that we do not care about.

Contraction For a pair $x, y \in V$, let $m(L) \in T$ be the minimal index such that $\rho_m + \rho_{m-1} \geq L/4$. Assuming a non-increasing sequence of weights $\{w_t\}_{t \in T}$ we have:

$$\begin{aligned}\|F(x) - F(y)\|_p^p &= \sum_{t=1}^{T} \sum_{i=1}^{R} w_t^p |f_t^{(i)}(x) - f_t^{(i)}(y)|^p && (1.14)\\ &\geq \sum_{t=1}^{T} w_t^p \frac{R}{24s} |\rho_t - \rho_{t-1}|^p && (1.15)\\ &\geq \frac{R}{24s} \sum_{t=1}^{m} w_t^p |\rho_t - \rho_{t-1}|^p && (1.16)\\ &\geq \frac{R w_m^p}{24s} \frac{m}{m} \sum_{t=1}^{m} |\rho_t - \rho_{t-1}|^p && (1.17)\\ &\geq \frac{R w_m^p}{24 s m^{p-1}} \rho_m^p && (1.18)\end{aligned}$$

Global and Local Dimensionality Reduction

Application 1: Bourgain's Theorem Take $L = d(x, y)/2$, $m(L) = \log_s n$ and $w_t = 1$. We get that the distortion is:

$$\left(\frac{\text{“expansion”}}{\text{“contraction”}}\right)^{1/p} = \left(\frac{R \log_s n \cdot 24s8^p}{R} \log_s^{p-1}(n)\right)^{1/p} = O(c \log_s(n))$$

Application 2: Local Embedding with dimension $\Omega(\log n)$ Define $L = \min\{d(u, v), r_k(u)\}$, $w_t = \vartheta^{-1/p}(t)$, we know that $m \leq 1 + \log_s k$ where $\vartheta(t)$ positive non-decreasing and $\sum_{t=1}^{\infty} \frac{1}{\vartheta(t)} = 1$. Then, we get that local distortion:

$$\left(\frac{\text{“expansion”}}{\text{“contraction”}}\right)^{1/p} = \left(\frac{R \cdot 1 \cdot 26s8^p \log_s^{p-1}(k)}{R} \vartheta(\log k)\right)^{1/p} = O\left(\left(\frac{\log k}{p}\right)^{1-1/p} \left(\vartheta\left(\frac{\log k}{p}\right)\right)^{1/p}\right)$$

Fix any $\epsilon > 0$ an explicit function that achieves this is $\hat{\theta}(k) = \hat{c} k \log^{1+\epsilon} k$, for some constant $\hat{c}$ selected so that the infinite summation is 1.

Dimension The dimension here is $O(2^p \log n \log n)$ for all values of k simultaneously. The same proof technique can give distortion $O(\log k/p)$ and dimension $O(2^p \log n \log K)$ if we care for k up to a fixed value K. We observe that the main bottleneck in the dimension is the *tensorization* scheme that requires $O(\log(n))$ dimensions for each value $t \in [\log_s k]$ of interest.

How can we control dependencies so that to avoid the $O(\log n)$ bottleneck?

A different tensorization scheme: Lovasz Local Lemma

Looking back, we observe that if we insist on independent copies of the basic maps f_t given by the distributional lemma, then a dimension of $\Omega(\log n)$ is unavoidable. Thus, we need to depart from the assumption of independence and focus on being more careful with our construction. Fortunately, there is a light handed way to achieve this using a tool call the *Lovasz Local Lemma* (LLL).

Theorem 1.4.3. *Let $\mathcal{A}_1, \ldots, A_n$ be events in some probability space. Let $G(V, E)$ be a graph on n vertices with degree at most d, with each vertex i corresponding to an event $\mathcal{A}_i$. Assume that for any $i = 1, \ldots, n$:*

$$\mathbb{P}[\mathcal{A}_i | \bigwedge_{j \in Q} \neg \mathcal{A}_j] \leq p \tag{1.19}$$

for all $Q \subseteq \{j : (A_i, A_j) \notin E(G)\}$. Then,

$$p \leq \frac{1}{e(d+1)} \Rightarrow \mathbb{P}[\bigwedge_{i=1}^{n} \neg \mathcal{A}_i] > 0 \tag{1.20}$$

The lemma shows only the existence of random choices such that the probability of the joint event is greater than zero, but doesn't give a guide of how to find such choices. Fortunately, there are algorithmic versions of the Lovasz Local Lemma that can achieve this and they are actually an object of recent intensive theoretical investigation. To carry this approach forward we need:

1. **Impose structure:** so that we can bound the *degree of the dependency graph* of the events of interest.

2. **Tensorization:** make sure that the probability is large enough so that the *condition of the LLL are satisfied.*

Controlling Dependences First, let us define the event of interests. Let $\mathcal{A}_t(u, v)$ be the event that the Distribution lemma succeeds succeeds at scale t for the pair $u, v \in V$, i.e, there are at least $\Theta(R)$ coordinates $i \leq R$ that the lower bound on distance of (u, v) is true, and $\mathcal{A}(u, v) = \bigwedge_{t=1}^{m} \mathcal{A}_t(u, v)$ be the event that it succeeds for all $t \in [m]$. The event of interest is that $\mathcal{A} = \bigwedge_{u,v \in V} \mathcal{A}(u, v)$.

- *Dependency radius:* A natural idea is to use the fact that we are looking only to preserve distances "around" k-neighborhoods. In particular, observe that the event $\mathcal{A}(u, v)$ depend on the inclusion of points in $B(u, d(u, v)) \cup B(v, d(u, v)) \subseteq B(u, 2d(u, v))$. Thus if two events $\mathcal{A}(u, v)$, $\mathcal{A}(x, y)$ are independent if $B(u, 2d(u, v)) \cap B(x, 2d(x, y)) = \emptyset$.

- *Structure:* in order to make these balls disjoint between most pairs $(u,v),(x,y)$ and the resulting events independent, we will only be concern with pairs that are far away from the boundary of the k-neigborhoods, i.e.,

$$\forall u,v \in V, d(u,v) \le r_k(u)/8 \tag{1.21}$$

Lemma 1.4.4. *Let G be the graph with nodes $U = \{(u,v)|d(u,v) \le r_k(u)/8\}$ and edges $E = \{\{(u,v),(x,y)\}|u \in \bar{N}_k(x)\}$ (both belong in the k-neigborhood). The degree of this graph is bounded by k^2 and all events not connected by an edge are independent.*

Proof. Since there are at most k points $u \in \bar{N}_k(x)$ and each u has at most k points v such that $d(u,v) \le r_k(u)$, any event $\mathcal{A}(u,v)$ is connected to at most k^2 other events. To show that the events are independent it is sufficient to show that the balls are disjoint.

- Assume that $r_k(x) \le 2d(x,u)$ and$r_k(u) \le 2d(u,x)$.

$$2(d(u,v)+d(x,y)) \le \frac{1}{4}(r_k(u)+r_k(x)) \le \frac{1}{2}(d(u,x)+d(u,x)) = d(u,x)$$

- The assumption is true (Proof by contradiction).

□

Bounding the probability of an individual event Using Chernoff bounds we can prove that using $c/12 \log k$ independent maps, each event $A_t(u,v)$ is not satisified with probability at most k^{-4}. Taking a union bound among the $m \le k$ values we get that $p \le k^{-3}$ and hence we have satisfied the conditions of the LLL.

How can we handle larger distances $r_k/8 < d(u,v) \le r_k$?

Non-Frechet Embedding

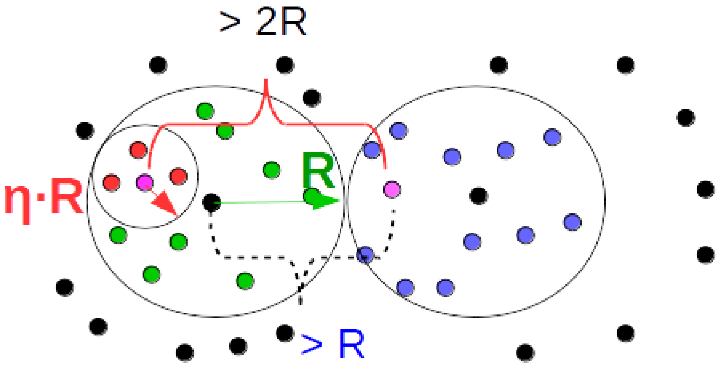
The main impediment in handling all distance of interest is that beyond that we do not control the dependencies that our distributional lemma imposes to the events of interest. Thus, we need to search for a *distributional lemma* that:

- Needs to handle only pairs with $\frac{r_k(u)}{8} < d(u,v) \le r_k(u)$.
- The resulting dependency graph has degree $O(k^2)$, so that we still get dimension $\log^2(k)$

The fact that we have the lower bound on the distance allows us to focus on producing partitions that are informed by this distance.

Distributional lemma based on Probabilistic Partitions

- We first obtain a *coarsening* of the metric space at scale k by partitioning the space into clusters $\mathcal{C} = \{C_1, \dots, C_s\}$ with centers $u_1, \dots, u_s$ compactly given by a function $C : \mathcal{X} \to \mathcal{C}$ such that
  - **Boundedness:** the diameter of clusters is bounded by Δ_{u_i}.
  - **Padding property:** with constant probability $B(x, \eta_i \Delta_{u_i}) \subsetneq C(x)$ for a padding parameter $\eta : \mathcal{X} \to (0, 1)$.



- **Coloring:** the final set W_k is constructed by including each cluster $C_1, \dots, C_s$ with probability 1/2.

Since, the points belong in different clusters then with constant probability exactly one of them will not belong to W_k in which case we have that:

$$|d(x, W_k) - d(y, W_k)| \geq d(x, \mathcal{X} \setminus C(x)) \geq \eta \cdot R \approx \eta \cdot d(x, y)$$

Definition 1.5.1 (Uniformly padded Local PP)**:** Given $\Delta > 0$ and $0 < \delta \leq 1$, let $\hat{\mathcal{P}}$ be a Δ-bounded probabilistic partition of $(\mathcal{X}, d)$. Given a collection of functions $\eta = \{\eta_C : \mathcal{X} | C \in \hat{\mathcal{P}}\}$, we say that $\hat{\mathcal{P}}$ is locally (η, δ)-*locally padded* if the event $B(x, \eta(x)\Delta) \subseteq C(x)$ occurs with probability at least δ regardless of the structure of the partition outside $B(x, 2\Delta)$. We say that $\hat{\mathcal{P}}$ is strongly $(\eta, \hat{\delta})$-locally padded if for any $\hat{\delta} \leq \delta \leq 1$, $\hat{\mathcal{P}}$ is $(\eta \ln(1/\delta), \delta)$-padded. We say that $\hat{P}$ is (η, δ)-*uniformly* locally padded if η is uniform with respect to $\hat{\mathcal{P}}$.

We apply a modification of the uniformly local padded decomposition on a specific sequence of centers carefully selected. We get a sequence of cluster $C_1, \dots, C_s$. Using those clusters we define the embedding as a summation of the functions.

$$\boxed{g^t(x) = \bar{D}^{-1/p} d(u, V \setminus C^{(t)}(x)) \cdot \sigma(C^{(t)(x)}) \qquad (1.22)}$$

Dependency Graph One of the main novelties of [ABN'15] is to bypass the random coloring procedure (and the concomittant dependencies that it introduces) by using an equivalent deterministic coloring procedure.

Lemma 1.5.2 (Coloring lemma)**.** *For any integer $\bar{D} > 1$ and $\delta \in (\Omega(1/\bar{D}), 1/2]$ there exists a set $S \subseteq \{-1, +1\}^{\bar{D}}$, $|S| \geq 2^{\bar{D}(1-H(\delta))/2}$ (entropy), such that for any $u \neq v \in S$ the hamming distance is at least $\delta\bar{D}$.*

Beyond Bourgains approach: Randomized Nash Device

For any $x, \omega \in \mathbb{R}^d$ and $\sigma > 0$ let $\phi(x, \sigma, \omega) := \frac{1}{\sigma}\begin{bmatrix}\cos(\sigma\omega^\top x)\\ \sin(\sigma\omega^\top x)\end{bmatrix}$ be a two dimensional map, that has the following two properties:

$$|\phi(x, \sigma, \omega) - \phi(y, \sigma, \omega)|^2 = 2\sigma^{-2}(1 - \cos(\sigma\omega^\top(x - y))) \tag{1.23}$$

$$\mathbb{E}[|\phi(x, \sigma, \omega) - \phi(y, \sigma, \omega)|^2] = 2\sigma^{-2}\left(1 - \exp\left(-\frac{1}{2}\sigma^2\|x - y\|^2\right)\right) \tag{1.24}$$

Randomized Nash Device Let $\sigma_1, \ldots, \sigma_D \in (0, \sigma_m)$ and $\omega_1, \ldots, \omega_D$ be samples from a d-dimensional gaussian $N(0, I_d)$. Define the map:

$$\Theta(x) = \frac{1}{\sqrt{D}} \bigoplus_{t=1}^{D} \phi(x, \sigma_t, \omega_t) \tag{1.25}$$

Lemma 1.6.1. *Let $\epsilon \in (0, 1/2)$ and $x, y \in \mathbb{R}^d$:*

- $\|\Theta(x) - \Theta(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2$ *with probability greater than* $1 - \exp(-\frac{D}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}))$
- *if* $\|x - y\| \leq \frac{\sqrt{\epsilon}}{\sigma_m}$, $\|\Theta(x) - \Theta(y)\|^2 \geq (1 - \epsilon)\|x - y\|^2$ *with probability greater than* $1 - \exp(-\frac{3D\epsilon^2}{128})$.
- *if* $\|x - y\| \geq \frac{1}{\sqrt{2}\sigma_m}$, $\|\Theta(x) - \Theta(y)\|^2 \geq \frac{1}{4\sigma_m^2}$ *with probability greater than* $1 - \exp(-\frac{D}{128})$.