# ON THE ESTIMATION OF GEODESIC PATHS ON SAMPLED MANIFOLDS UNDER RANDOM PROJECTIONS

*Mona Mahmoudi* [†], *Pierre Vandergheynst* [‡], *Matteo Sorci* [‡]

[†]Electrical and Computer Engineering, University of Minnesota
Minneapolis, MN, 55455, USA

[‡]Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

## ABSTRACT

In this paper, we focus on the use of random projections as a dimensionality reduction tool for sampled manifolds in high-dimensional Euclidean spaces. We show that geodesic paths approximations from nearest neighbors Euclidean distances are well-preserved by Gaussian projections and we characterize the distribution of geodesic lengths in the reduced dimensional point cloud. A stylized application to a real-world data set of human faces is presented to validate our theoretical findings.

***Index Terms***— Smooth manifolds, Geodesic distance, Random projection

## 1. INTRODUCTION

With the advent of the digital age and the widespread access to both efficient communication and storage facilities, we are now over-flooded with data. A recent report, [1], estimates the total amount of digital information created in 2006 at a blazing 161 exabytes, that is $161 \cdot 10^{18}$ bytes! Clearly, we are facing a data landslide and appropriate technology must be developed to overcome this challenge.

Finding information in large, high-dimensional data sets is the modern-day example of the "needle in the haystack". Very often, though, data sets exhibit some sort of organization that can be exploited. In this paper, we target, in particular, large sets of signals that can be seen as samples of high-dimensional smooth manifolds embedded in $\mathbb{R}^d$. In these cases, using the intrinsic differential geometry of the manifold greatly enhances tasks such as classification, recognition, finding nearest neighbors, etc. Some promising algorithms have been introduced recently to exploit the geometry of point clouds [2, 3, 4, 5, 6]. One crucial step in most of them is to correctly approximate geodesic distances between distant elements on the basis of nearest neighbor Euclidean distances, usually by means of shortest paths algorithms. However, when the embedding space is very high-dimensional, evaluating shortest paths on large point clouds can become numerically daunting. It is thus absolutely vital to reduce the dimensionality of the point cloud, while ensuring that geodesic distances are preserved.

In this paper, we study the behavior of shortest path approximations to geodesic distances under dimensionality reduction by random projections. Recently, there has been an increasing interest in understanding how a random embedding would affect the geometrical characteristics of a smooth manifold, for example, in [7] the authors characterize closed manifold properties. In this paper, however, we focus on the properties of point clouds sampled from smooth (not necessarily closed) manifolds and, in particular, we study outcome of shortest path computations. We show that, under mild assumptions on the underlying manifold, the distribution of path lengths after random projection is sharply distributed around the true length. Finally, we illustrate our findings by exploring a set of human faces displaying several levels of emotions, showing that approximate geodesic paths between any two pictures of the same person preserve the correct identity, i.e the shortest paths very rarely ventures through a wrong person.

## 2. THEORY

In this part, we investigate some properties of geodesic distance and path lengths in the point cloud data after random projection. In Section 2.1, we prove that if the Euclidean distance between each pair of projected points is preserved in the range of $(1 \pm \epsilon)$ of its original value, the geodesic distance is also preserved within the same ratios of its original values under certain conditions on the selection of neighbors. Furthermore, in Section 2.2, we find bounds for the expected value and variance of the path lengths after projection, assuming that the projection matrix $A$ has independent $\mathcal{N}(0, \frac{1}{n})$ entries.

### 2.1. Preserved Geodesic Distance Under Projection

**Theorem 1** *Under an embedding that satisfies the following inequality for Euclidean distance*

$$(1-\epsilon) \leq \frac{d^A(Ax, Ay)}{d(x,y)} \leq (1+\epsilon), \qquad (1)$$

*the geodesic distances on a modified neighborhood graph also satisfy*

$$(1-\epsilon) \leq \frac{d_G^A(Ax, Ay)}{d_G(x,y)} \leq (1+\epsilon), \tag{2}$$

*for all points $x$ and $y$ on the manifold, where $d(x,y)$ and $d^A(Ax, Ay)$ are Euclidean distances between $x$ and $y$ in the original and projected space, respectively, and $d_G$ similarly represents the geodesic distance.*

*Proof:* In the process of computing geodesic distances, changes in the neighborhood graph may result in different paths and geodesic distance values in the projected space compared to the original ones. Thus, we propose a constraint on the neighborhoods, which is including all the points in the original K-neighborhood of each point in the selected neighborhood after projection, to preserve geodesic distances. If path $P_1$ is the shortest path between points $i$ and $j$ in the original space and path $P_2'$ is the shortest path between them in the projected space, we can prove that its length is in the range of $(1 \pm \epsilon)|P_1|$ ($|P_1|$ is the length of $P_1$) when the proposed constraint is satisfied. If $P_1'$ is the projected $P_1$ and $P_2$ is the original $P_2'$ then

$$(1-\epsilon)|P_1| \leq |P_1'| \leq (1+\epsilon)|P_1|$$

$$(1-\epsilon)|P_2| \leq |P_2'| \leq (1+\epsilon)|P_2|.$$

Since $|P_2'|$ is the shortest path and it is smaller than $|P_1'|$, unless it is within $(1\pm\epsilon)|P_1|$, it should be less than $(1-\epsilon)|P_1|$, so

$$(1-\epsilon)|P_2| \leq |P_2'| \leq (1-\epsilon)|P_1|$$

which concludes that $|P_2| < |P_1|$[1] which is a contradiction to $P_1$ being the shortest path in the original space. Thus, we have proved that $(1-\epsilon)|P_1| \leq |P_2'| \leq (1+\epsilon)|P_1|$.

Now, we need to find the condition so that all the points in the original K-neighborhood are also in the new one. Assume $d_1$ and $d_2$ are the distances between point $i$ with points $j_1$ and $j_2$ (i.e. $d_1 = d(i, j_1)$ and $d_2 = d(i, j_2)$) and their projected values are $d_1'$ and $d_2'$, also assume that $j_1$ is in the original K-neighborhood of $i$, but $j_2$ is not, and the new neighborhood includes $j_2$ but not $j_1$ which concludes that

$$\begin{gathered} d_1'(1-\delta_1) = d_1 \leq d_2 = d_2'(1+\delta_2) \\ d_1' \geq d_2' \end{gathered} \tag{3}$$

where $\delta_1$ and $\delta_2$ are positive. Moreover, we know that

$$\begin{aligned} & d_1(1+\epsilon) = d_1'(1-\delta_1)(1+\epsilon) \geq d_1' \\ \Rightarrow\ & (1-\delta_1) \geq \tfrac{1}{1+\epsilon} \\ \Rightarrow\ & \delta_1 \leq \tfrac{\epsilon}{1+\epsilon} \end{aligned}$$

and

$$\begin{aligned} & d_2(1-\epsilon) = d_2'(1+\delta_2)(1-\epsilon) \leq d_2' \\ \Rightarrow\ & (1+\delta_2) \leq \tfrac{1}{1-\epsilon} \\ \Rightarrow\ & \delta_2 \leq \tfrac{\epsilon}{1-\epsilon} \end{aligned}$$

considering these inequalities and Equation (3) we have

$$\frac{d_1'}{d_2'} \leq \frac{1+\delta_2}{1-\delta_1} \leq \frac{1+\epsilon}{1-\epsilon}. \tag{4}$$

To define the neighborhood in the projected space, we first compute the K-neighborhood for each point and define $d'_{max}$ as its maximum distance to the points in its K-neighborhood. Based on Equation (4), if we also include all the points which are in the ball of radius $\frac{1+\epsilon}{1-\epsilon}d'_{max}$ in the neighborhood, all the points in the original K-neighborhood are also in the new neighborhood.

Moreover, to avoid shortcuts in the manifold after the projection, we need to define a constraint for the condition number. If $\tau$ is the inverse of condition number [8], since we need the maximum radius of the neighborhood to be less than $\frac{\tau}{2}$, it should satisfy the following condition

$$\max_i(\max_{j \in N_K(i)}(d_{ij}))\frac{1+\epsilon}{1-\epsilon} < \frac{\tau}{2} \tag{5}$$

where $N_K(i)$ is the K-neighborhood of point $i$ in the original space. Finally, we proved that if Equations (5) and (1) are satisfied and we define the neighborhood as explained above, the geodesic distances satisfy Equation (2).

The Johnson-Lindenstrauss Lemma guarantees that one can embed $p$ points from $\mathbb{R}^d$ into $\mathbb{R}^q$, $q \geq O(\epsilon^{-2}\log p)$, while satisfying (1). It is also well-known that random embedding by Gaussian projections satisfy Equation (1) with high probability, see [9] for more details. Under random projections, though, our reduced dimensional points, the actual paths, and their lengths become random variables, and this motivates us to study their distribution in Section 2.2.

### 2.2. Path Lengths Variation

In the previous section, we found the maximum perturbation for estimated geodesic distances which represents the worse cases. In this section, we investigate some properties of path lengths distributions by finding the upper and lower bounds of their expected values and the upper bound of their variances.

#### *2.2.1. Expected Value Of The Path Lengths*

Assume $A$ is the $n \times d$ projection matrix, and each $v_i$ represents a vector of difference between two points. We have the following inequality as the upper bound for the expected value of a path length with edge weights $\|v_i\|, i = 1..P$

$$\begin{gathered} E\left[\sum_{i=1}^P \|Av_i\|\right] = \sum_{i=1}^P E\left[\sqrt{\|Av_i\|^2}\right] \leq \\ \sum_{i=1}^P \sqrt{E\left[\|Av_i\|^2\right]} = \sum_{i=1}^P \|v_i\| \end{gathered} \tag{6}$$

[1] *There is a possibility that $P_2$ does not exist in the original graph. In this case, since $|P_2| < |P_1|$, the original K-neighborhood selection is not good enough to find the real shortest path.*

which is concluded from Jensen's inequality for expected value and concavity of square root and the fact that $E[\|A^T A\|] = \mathbf{1}$.

For the lower bound of this expected value, considering Markov's inequality and for arbitrary $t$, we have

$$\begin{aligned} E\left[\|Av_i\|\right] &> \sqrt{(1-t)\|v_i\|^2} P\left[\|Av_i\|^2 > (1-t)\|v_i\|^2\right] \\ &= \|v_i\|\sqrt{1-t} P\left[1 - \frac{\|Av_i\|^2}{\|v_i\|^2} < t\right] \end{aligned}$$

where by normalizing $v_i$, $x = \frac{v_i}{\|v_i\|}$, we can use the bound on probability introduced in [10] (Equation (3.1)) as follows

$$\begin{aligned} E\left[\|Av_i\|\right] &> \|v_i\|\sqrt{1-t} P\left[\|x\|^2 - \|Ax\|^2 < t\right] \\ &= \|v_i\|\sqrt{1-t}\left(1 - P\left[\|x\|^2 - \|Ax\|^2 \geq t\right]\right) \\ &\geq \|v_i\|\sqrt{1-t}\left(1 - e^{-n\frac{t^2}{C_1+C_2 t}}\right) \\ &= f\left(t, n, \|v_i\|\right), \end{aligned} \tag{7}$$

where $n$ is the projected dimension. We would like to show that, when $n$ grows, $E\left[\|Av_i\|\right] > \|v_i\|$ for some $t$. Thus, let us take $t = n^{-\alpha}$ with $\alpha > 0$. This gives :

$$\begin{aligned} f\left(t, n, \|v_i\|\right) &= \sqrt{1-n^{-\alpha}}\left(1 - e^{-\frac{n^{1-2\alpha}}{C_1+C_2 n^{-\alpha}}}\right)\|v_i\| \\ &\approx \sqrt{1-n^{-\alpha}}\left(1 - e^{-C_3 n^{1-2\alpha}}\right)\|v_i\| \end{aligned}$$

that indeed converges to $\|v_i\|$ when $n$ grows if we further impose $\alpha < 1/2$. Together with (6), this shows that estimating the geodesic distances with the projected points becomes unbiased as $n$ grows.

### 2.2.2. *Variance Of The Path Lengths*

For the variance of the path length, similar to previous section, we have

$$\begin{aligned} &\text{var}\left(\sum_{i=1}^P \|Av_i\|\right) \\ &= E\left[\left(\sum_{i=1}^P \|Av_i\|\right)^2\right] - E\left[\sum_{i=1}^P \|Av_i\|\right]^2 \\ &\leq \sum_{i=1}^P \|v_i\|^2 + \sum_{i\neq j}\sqrt{E\left[\|Av_i\|^2 . \|Av_j\|^2\right]} \\ &- \left(\sum_{i=1}^P \max_t f\left(t, n, \|v_i\|\right)\right)^2 \end{aligned} \tag{8}$$

where, after expanding, the second sentence is concluded from the concavity of square root and Jensen's inequality and the third sentence is obtained from Equation (7). To find the exact values in the second sentence, if we assume $y = v_i$ and $z = v_j$, $g_{lk}$ for $l = 1..n, k = 1..d$ are independent with Gaussian distribution, and $a_{lk} = \frac{1}{\sqrt{n}} g_{lk}$ then

$$\begin{aligned} &E\left[\|Av_i\|^2 . \|Av_j\|^2\right] \\ &= \frac{1}{n^2} E\Big[\left(\sum_{l=1}^n \sum_{k=1}^d \sum_{m=1}^d g_{lk} g_{lm} y_k y_m\right) \\ &\left(\sum_{l'=1}^n \sum_{k'=1}^d \sum_{m'=1}^d g_{l'k'} g_{l'm'} z_{k'} z_{m'}\right)\Big] \end{aligned}$$

using the linearity, independence of $g_{sr}$'s, and the fact that $E\left[g_{sr}^2\right] = 1$ and $E\left[g_{sr}^4\right] = 3$, the following equation is concluded

$$\begin{aligned} &E\left[\|Av_i\|^2 . \|Av_j\|^2\right] \\ &= \frac{2}{n}\sum_{k\neq m} y_k y_m z_k z_m + \sum_{k=1}^d \sum_{k'=1}^d y_k^2 z_{k'}^2 \\ &- \frac{1}{n}\sum_{k=1}^d y_k^2 z_k^2 + \frac{3}{n}\sum_{k=1}^d y_k^2 z_k^2 \\ &= \frac{2}{n} < y, z >^2 + \|y\|^2 . \|z\|^2 \end{aligned}$$

where, combined with Equation (8), we get

$$\begin{aligned} &\text{var}\left[\sum_{i=1}^P \|Av_i\|\right] \leq \sum_{i=1}^P \|v_i\|^2 \\ &+ \sum_{i\neq j}\sqrt{\frac{2}{n} < v_i, v_j >^2 + \|v_i\|^2 . \|v_j\|^2} \\ &- \left(\sum_{i=1}^P \max_t f\left(t, n, \|v_i\|\right)\right)^2 \end{aligned} \tag{9}$$

From the previous section, we know that, as $n$ grows, $f\left(t, n, \|v_i\|\right)$ goes to $\|v_i\|$ for some $t$. We also have that :

$$f\left(t, n, \|v_i\|\right) \leq \left(1 - e^{-n\frac{t^2}{C_1+C_2 t}}\right)\|v_i\| \leq \|v_i\| .$$

Hence, $\max_t f\left(t, n, \|v_i\|\right)$ goes to $\|v_i\|$ and the right hand side of inequality (9) goes to zero as $n$ increases since the first two terms cancel the third one. Also, we can analyze the behavior of the expression in Equation (9) in another way

$$\begin{aligned} &\text{var}\left[\sum_{i=1}^P \|Av_i\|\right] \leq \sqrt{\frac{2}{n}+1}\left(\sum_{i=1}^P \|v_i\|\right)^2 \\ &- \left(\sum_{i=1}^P \max_t f\left(t, n, \|v_i\|\right)\right)^2 \end{aligned} \tag{10}$$

which is concluded from Cauchy-Schwartz inequality, the fact that $\sqrt{\frac{2}{n}+1} > 1$ and that $\sum_{i=1}^P \|v_i\|^2$ is positive. This upper bound for variance of the path lengths also goes to zero when $n$ increases.

## 3. EXPERIMENTAL RESULTS

In this section, we present some experiments on the manifold of 1271 face images from 11 people with different expressions obtained from the database in [11]. The results show that most of the shortest paths between these images are preserved after random projection. For each image, we have a feature vector of length 348 representing the appearance parameters of an active appearance model (AAM) [12] combined with some descriptive measures computed on the face mask [13]. We project these vectors to 30 dimensions by a matrix which has independent $\mathcal{N}(0, \frac{1}{30})$ elements. Then, for each person $i$ we compute a path histogram. In this histogram, the value of bin $j$ is the number of times any point corresponding to person $j$ lies on a shortest path (computed using 10 nearest neighbors) between two points corresponding to

person $i$. When the shortest path connecting two end points, which belong to person $i$, is empty, one of the end points is counted in the histogram to emphasize on these empty paths. Since the feature vectors of each person's expressions are expected to be closer to each other than to those of the other persons, we expect that the peak of the histogram of each person belong to the same person. This type of histograms can be useful for recognition applications. In Fig. 1, we have presented these histograms for both the original and projected datasets where the projected histograms are averages of histograms obtained from 10 different projections.

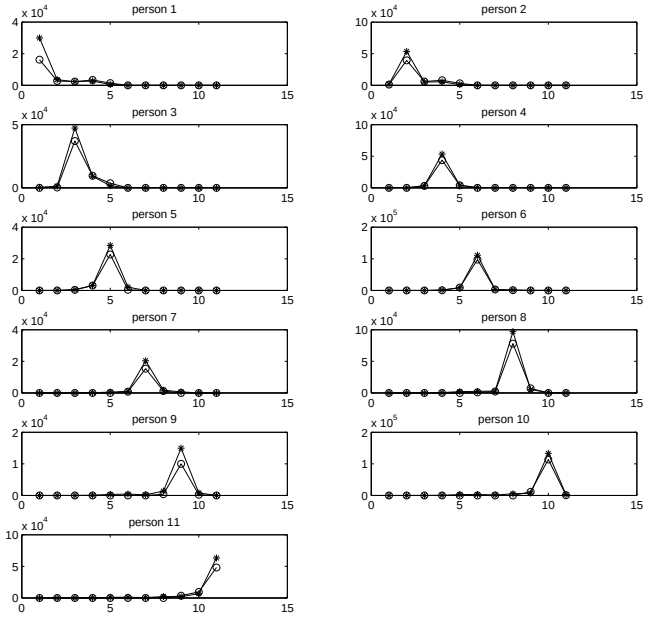


**Fig. 1**. The path histograms of 11 people are presented. The $o$'s are for original feature vectors and the *'s for the projected ones.

## 4. CONCLUSIONS

In this paper, we proved that under projections that preserve Euclidean distances, estimated geodesic distances and geodesic path lengths are sharply distributed around their original values. In addition, we characterized the first two moments of the distribution of these geodesic path lengths under Gaussian random projections. We numerically studied some properties of geodesic paths on a dataset of human face images which validated our theoretical results and which are motivating some future work on face recognition.

## 5. REFERENCES


[1] J. F. Gantz et al., "The expanding digital universe: A forecast of worldwide information growth through 2010," *IDC White Paper sponsored by EMC*, March 2007.

[2] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[3] J. Langford J. Tenenbaum M. Bernstein, V. de Silva, "Graph approximations to geodesics on embedded manifolds," *Tech. Rep., Stanford University*, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Tech. Rep. TR 2002-01, Univ. Chicago, Dept. Comp. Sci. and Statistics*, January 2002.

[5] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.

[6] R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F.J. Warner, and S.W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data. part i: Diffusion maps," *Proc. Nat. Acad. Sci.*, vol. 102, pp. 7426–7431, May 2005.

[7] C Hegde and R. Baraniuk, "Random projections for manifold learning," *Neural Information Processing Systems (NIPS)*, 2007.

[8] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, pp. 941–944, 2006.

[9] D. Achlioptas, "Database-friendly random projections," *Proc. 20th ACM Sympos. Principles Database Syst.*, pp. 274–281, 2001.

[10] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *submitted to IEEE Trans. Information Theory*, 2007.

[11] M. Sorci, G. Antonini, J.-Ph. Thiran, and M. Bierlaire, "Facial Expressions Evaluation Survey," Tech. Rep., 2007, ITS.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681–685, June 2001.

[13] B. Cerretani, M. Sorci, and J.-Ph. Thiran, "Modelling human perception of static expressions by discrete choice models," ITS, 2007.