Out-of-sample extension of spectral embeddings: An optimization perspective

Chester Holtz[*]
University of California, San Diego
chholtz@ucsd.edu

Taanish Reja[*]
University of California, San Diego
treja@ucsd.edu

Gal Mishne[*]
University of California, San Diego
gmishne@ucsd.edu

Abstract

Graph-based manifold learning constructs / reveals low dimensional embeddings of high-dimensional data, however requires out-of-sample-extension methods to embed new data points. We propose a new framework, *ROSE* (Riemannian Out-of-Sample Extension), for out-of-sample extensions for spectral graph-based embedding algorithms. *ROSE* is motivated from an optimization perspective of the underlying eigenvector problem associated with classic manifold learning problems. Similar to graph-based semi-supervised learning, our approach exploits the geometry of new points in addition to the sampled points, by treating the in-sample embedddings as *labeled* data. *ROSE* Despite its nonconvexity, *ROSE* is solvable by first-order methods, which converge to global minimizers under certain assumptions.

1 Introduction

Given a graph $G = (V, E)$ on n vertices with adjacency matrix $W \in \mathbb{R}^{n \times n}$, the goal of graph embedding is to map the vertices of G to some d-dimensional vector space in such a way that geometry of the embedding preserves the geometry of G. For example, we may ask that vertices with high connectivity in G be assigned to nearby vectors in the embedding space.

Various approaches have been proposed to compute effective graph embeddings. A classic method that is particularly relevant is the manifold learning approach Laplacian Eigenmaps [1, 2], which utilizes the graph Laplacian matrix to produce low-dimensional representations of graph vertices.

One challenge of these approaches is that they do not provide an explicit mapping between the vertices and the low-dimensional embedding. Therefore, when new data is introduced or when large-scale datasets prevent an embedding of the full dataset due to computational limitations, out-of-sample extension (OOSE) methods are used to extend the embeddings from the training samples to the rest of the data or to new unseen data [3]. For instance, in the case of Laplacian Eigenmaps, out-of-sample extension can be performed using a Nyström approximation. The Nyström method [4] allows for the efficient approximation of the eigenfunctions of the graph Laplacian on the new data points by expressing them as a linear combination of the eigenfunctions computed from the original dataset [4–7]. Beyond these classical methods, more recent approaches the family of spectral and diffusion networks [3, 8, 9] have been proposed to handle out-of-sample extension in graph embeddings.

In this paper, we propose a new approach to graph-based out-of-sample extensions that builds upon these traditional methods but introduces a novel optimization perspective. Our method focuses on directly learning the mapping between the data and the embedding space through an optimization

[*]Equal contribution.

C. Holtz et al., Out-of-sample extension of spectral embeddings: An optimization perspective (Extended Abstract). Presented at the Second Learning on Graphs Conference (LoG 2023), Virtual Event, November 27–30, 2023.

framework. This approach allows for a more principled and systematic extension of embeddings to new data points while preserving the topological and geometric properties of the original graph embedding. We provide one interpretation of out-of-sample extensions of spectral algorithms as a quadratic optimization problem over a smooth manifold with special quadratic constraints, for which one can implement gradient or conjugate gradient methods and Newton methods over geodesic paths on the manifold [10, 11].

Let $St(n, r)$ denote Stiefel manifold defined as

$$St(n, r) = \{X \in \mathbb{R}^{n\times r} : X^\top X = I_r\}$$

where I_r denotes the $r \times r$ identity matrix. In short, $St(n, r)$ is the set of matrices in $\mathbb{R}^{n\times r}$ whose columns are orthonormal in $\mathbb{R}^n$ with respect to the inner product $\langle x, y\rangle = tr(x^\top y)$.

In this paper, we propose an efficient algorithm to solve out-of-sample extension for manifold learning problems. Our algorithm is based on a reduction to a quadratic problem of the form

$$\min_{X\in St(n,r)} \{F(X) = \langle X, AXC\rangle - \langle B, X\rangle\}, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $C \in \mathbb{R}^{r\times r}$ is positive definite and the the inner product $\langle R, T\rangle$ is the trace of the matrix $R^\top T$ for R, T of the same size.

2 Quadratic Minimization for Out-of-Sample-Extension

Given a dataset $D = \{x_1, \ldots, x_{m'}\}$ with $x_i \in \mathbb{R}^d$, D gives rise to a symmetric matrix $W \in \mathbb{R}^{m'\times m'}$ derived from a kernel $w_{ij} = K(x_i, x_j) \geq 0$, that measures the similarity between pairs of datapoints x_i and x_j such that $K(x_i, x_j)$ is large if x_i and x_j are similar and small otherwise. This matrix defines a graph where nodes are data points, and edges represent pairwise similarities, with its combinatorial Laplacian $L = \text{diag}(1^\top W) - W$ capturing the local geometric structure of the data. The eigenvectors of this matrix are used to embed the data in a lower-dimensional space, preserving local neighborhood relationships.

Introduce the graph $G = (V, E, W)$ induced from W and D and with m' vertices corresponding to the m' data points, where $V = \{v_1, v_2, \ldots, v_{m'}\}$ is the vertex set, E is the edge set and W is the weight matrix whose entries $w_{ij} \geq 0$ are the edge weights between v_i and v_j. Assume that the graph is symmetric, i.e., $w_{ij} = w_{ji}$. An embedding of the vertices into $\mathbb{R}^r$ is given by the eigenvectors X corresponding to the smallest r nontrivial eigenvalues,

$$\min_{X\in\mathbb{R}^{m'\times r}} \langle X, LX\rangle,\ 1_n^\top X = 0,\ X^\top X = I_r. \tag{2}$$

In the context of an out-of-sample extension, suppose we are given the embedding associated with a subset of the m' samples, and where new points need to be embedded without recomputing the eigendecomposition.

More concretely, we consider a set of pre-specified "in-sample" data, the "training set" to be available. I.e., let the in-sample data corresponds to first m vertices $V_l := \{v_1, v_2, \ldots, v_m\}$ with observations $\{y_1, y_2, \ldots, y_m\}$, where $0 < m \ll m'$. Let n denote the number of out-of-sample vertices, $n = m' - m$. Our task corresponds to smoothly propagating the observed values over the out-of-sample vertices $V_u := \{v_{m+1}, v_{m+2}, \ldots, v_{m'}\}$. Let

$$L = \begin{bmatrix} L_{l,l} & L_{l,u} \\ L_{u,l} & L_{u,u} \end{bmatrix},\ Y = \begin{bmatrix} Y_l \\ Y_u \end{bmatrix}, Y_l = [y_1, \ldots, y_m]^\top,\ X = \begin{bmatrix} X_l \\ X_u \end{bmatrix}. \tag{3}$$

where subscripts l and u correspond to in-sample and out-of-sample indices, respectively. Let X_l represent the **in-sample vertices**. Introduce the associated constraint set

$$\mathcal{X} := \{X \in \mathbb{R}^{m'\times r} : 1_n^\top X = 0, X^\top X = X_l^\top X_l + X_u^\top X_u = pI, X_l = Y_l\}, \tag{4}$$

for some scalar $p = m'/r$. The following proposition indicates that the unknown matrix X_u can be computed from one quadratic minimization problem over a Stiefel manifold, i.e., (8).

Proposition 2.1. *Let p be a positive scalar. Given X, L in (3). Given the observed in-sample matrix $X_l \in \mathbb{R}^{m\times r}$ and $c \in \mathbb{R}^r$, Consider the minimization*

$$\min_{X_u\in\mathbb{R}^{n\times r}} \{\langle X, LX\rangle : X \in \mathcal{X}\}. \tag{5}$$

Let c_l be the column sum of X_l, i.e. $= X_l^\top 1_m$ and $P = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and

$$A = PL_{u,u}P,\ \ B = -P(n^{-1}L_{u,u}1_n c_l^\top + L_{u,l}X_l),\ \ C = pI - X_l^\top X_l - \frac{1}{n}c_l c_l^\top. \tag{6}$$

Then, $X_u = XC^{1/2} + \frac{1}{n}1_n c_l^\top$, where X is the minimizer of

$$\min_{X\in\mathbb{R}^{n\times r}} \left\{\langle X, AXC\rangle - 2\langle X, BC^{1/2}\rangle : X \in St(n,r)\right\} \tag{7}$$

For brevity, we will recast $B = BC^{1/2}$. To reiterate, given a solution to (7), X^*, one recovers a solution to (5) via the transformation

$$X^*C^{1/2} + \frac{1}{n}\mathbf{1}r^\top \tag{8}$$

2.1 Optimality conditions

Optimization over the Stiefel manifold is a nonconvex problem. Generally, it is not possible to recover the global minimum. However, we show that in certain special cases recover of high-quality critical points is likely for first-order methods under an appropriate initialization. First, we define critical points to be those points that satisfy the following *first-order condition*

$$AXC = B + X\Lambda \tag{9}$$

for some $\Lambda \in \mathbb{R}^{n\times r}$. Points satisfying this condition can be local maximizers, minimizers, or saddle points. In general, there can be many critical points satisfying this condition. The following illustrates that a critical point X satisfying a certain second-order condition is a global minimizer of (1).

Proposition 2.2. *Let d_1 be the smallest eigenvalue of A and X' be a critical point of*

$$\min_X F(X) \quad \textit{s.t. } X^\top X = I \tag{10}$$

and let Λ' be the associated multipliers matrix. Suppose

$$d_1 C \succcurlyeq \Lambda'. \tag{11}$$

Then X' is a global minimizer. Suppose $d_1 C \succ \Lambda'$. Then, X' is the unique global minimizer.

Note that in general, the condition in (11) could be too strict to be fulfilled for any critical points. We introduce the following proposition to describe a less restrictive spectral condition that implies global optimality for certain special quadratics. The following non-degeneracy condition on B ensures that any critical point X satisfying a certain condition is a global minimizer. More concretely, the projection of B on V must sufficiently large compared to the spectral gap $d_r - d_1$ such that $\Lambda \succcurlyeq d_r C$.

Proposition 2.3. *Let $V = [v_1, v_2, \ldots, v_r] \in \mathbb{R}^{n\times r}$ be the eigenvectors of A corresponding to the smallest r nonzero eigenvalues $d_1 \le d_2 \le \ldots \le d_r$. Let (X, Λ) be a local solution satisfying*

$$AXC = B + X\Lambda$$

and $\lambda_1, \ldots, \lambda_r \le d_r$. Let s_1 be the smallest singular value of $V^\top BC^{-1}$. Suppose

$$d_r - \gamma_j \ge \sigma \textit{ for all } j = 1, \ldots, r, \textit{ and } \sigma > d_r - d_1 \tag{12}$$

Then, all eigenvalues $\gamma_1, \ldots, \gamma_r$ of the matrix ΛC^{-1} are less than d_1 and X is a global minimizer.

3 Riemannian Out-of-Sample Extension (ROSE)

In this section, we describe our algorithm, termed ROSE, for semi-supervised out-Of-Sample extension. A standard algorithm for minimizing a smooth objective over the Stiefel Manifold, e.g. (1), is given in "An introduction to optimization on smooth manifolds" by Nicolas Boumal. For each $U \in \mathbb{R}^{n\times r}$, define the projection $\mathcal{P}$ onto the tangent space at $X \in St(n,r)$,

$$\mathcal{P}_X(U) = Xskew(X^\top U) = X - Xsym(X^\top U) \tag{13}$$

Where $sym(Z) = \frac{1}{2}(Z + Z^\top)$. Let $F(X; A, B, C) = \langle X, AX\rangle/2 - \langle X, B\rangle$ be the objective of (1). Then, the Euclidean gradient is given by$\nabla F = AXC - B$ and the projected gradient is given by

$$gradF(X) = \mathcal{P}_X \nabla F(X) = \mathcal{P}(AXC - B) = (AXC - B) - Xsym(\Lambda). \tag{14}$$

Where $\Lambda = X^\top(AXC - B)$. The Riemannian Gradient method computes a sequence of iterates $X_0, X_1, \ldots, X_k$ where

$$X_{k+1} = \mathcal{R}_{X_k}(-\alpha_k gradF(X_k)) \tag{15}$$

Where $\mathcal{R}_{X_k}(g_k) = UV^\top$ for $U\Sigma V^\top = SVD(X_k - g_k)$.

3.1 Initialization of Riemannian Gradient

Convergence of first-order methods typically relies on initialization. We justify a computationally friendly initialization that relies on the approximation of an "ideal" initialization given below.

Proposition 3.1. *Let V be an isometric matrix, $V \in St(l,r), n > l \geq r$. Let $\mathcal{S}$ be an induced subspace, $X = \{V\widetilde{X} : \widetilde{X} \in St(l,r)\}$. Consider the subspace-restricted regularized problem,*

$$\begin{aligned} &\min_X \{F(X;A,B,C) : X = V\widetilde{X} \in \mathcal{S}, X \in St(n,r)\} \\ =&\min_{\widetilde{X}} \{F(\widetilde{X}; V^\top AV, V^\top B, C) : \widetilde{X} \in St(n,r)\} \end{aligned} \tag{16}$$

Consider $V = V_g$, i.e., the subspace-restricted problem

$$\min_{X \in St(n,r)} \{F(X; A, V_g V_g^\top B, C)\} \tag{17}$$

Then the associated multiplier $\Lambda = X^\top(AXC - V_g V_g^\top B)$ satisfies $\Lambda \preccurlyeq d_r C$

Proof. Let $X = V_g Q$, where Q is one orthogonal matrix, which maximizes $\langle Q, V_g^\top B\rangle$. Note that

$$\Lambda = X^\top(AXC - B) = d_r C - Q^\top V_g^\top B \tag{18}$$

Is symmetric. Hence, Λ can be expressed as the difference between two Hermitian matrices, and by Weyl's inequality, the proof is complete. □

Since computing V_g is expensive, we instead consider an estimate of $V_g Q$, $X_0 = PD_u^{-1}W_{ul}Y$.

4 Preliminary Experiments

The experiment in Figures 1 and 2 show the average-percentage of neighborhood overlap between the $k$nn graphs derived from data in the ambient space and the $k$nn graphs derived in the embedding spaces. Each line graph is generated from averages over 10 trials, where in each trial different in-sample sets were chosen uniformly. In Figure 2, we report a measure of global distortion. These experiments demonstrates ROSE's ability to preserve local neighborhoods in relative to a full eigenvector decomposition.

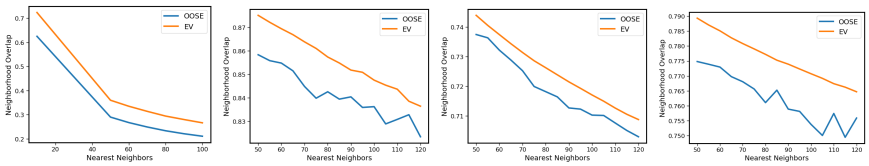

Figure 1: Neighborhood overlap at 10% in-sampled data for the noisy swiss roll dataset, MNIST, Fashion MNIST, and CIFAR-10.

# in-sample	Eigenvectors	ROSE[1]	ROSE[2]	Nyström
1%	2.24 ± 19.36	2.06 ± 3.46	3.69 ± 5.49	9.86 ± 19.13
5%	2.18 ± 16.16	2.47 ± 9.33	5.47 ± 12.66	8.37 ± 26.65
10%	2.28 ± 18.15	2.66 ± 13.64	5.01 ± 12.73	6.76 ± 25.29
25%	2.29 ± 15.68	2.90 ± 15.44	3.94 ± 9.10	4.45 ± 12.25

Table 1: Distortion of embedding methods on MNIST. Distortion$(\Phi_k, \mathcal{U}_k) = ||\Phi_k||_{Lip}\, ||\Phi_k^{-1}||_{Lip}$

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317. 1

[2] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006. URL http://jmlr.org/papers/v7/belkin06a.html. 1

[3] Gal Mishne, Uri Shaham, Alexander Cloninger, and Israel Cohen. Diffusion nets. *Applied and Computational Harmonic Analysis*, 47(2):259–285, 2019. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2017.08.007. URL https://www.sciencedirect.com/science/article/pii/S1063520317300957. 1

[4] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf. 1

[5] Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/cf05968255451bdefe3c5bc64d550517-Paper.pdf.

[6] Petros Drineas and Michael W. Mahoney. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(72):2153–2175, 2005. URL http://jmlr.org/papers/v6/drineas05a.html.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004. doi: 10.1109/TPAMI.2004.1262185. 1

[8] Uri Shaham, Kelly Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJ_aoCyRZ. 1

[9] Ziyu Chen, Yingzhou Li, and Xiuyuan Cheng. Specnet2: Orthogonalization-free spectral embedding by neural networks. In Bin Dong, Qianxiao Li, Lei Wang, and Zhi-Qin John Xu, editors, *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pages 33–48. PMLR, 15–17 Aug 2022. URL https://proceedings.mlr.press/v190/chen22a.html. 1

[10] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. 2

[11] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3. 2