

Landmark Alternating Diffusion*

Sing-Yuan Yeh[†], Hau-Tieng Wu[‡], Ronen Talmon[§], and Mao-Pei Tsui[¶]

Abstract. Alternating Diffusion (AD) is a commonly used diffusion-based sensor fusion algorithm. While it has been successfully applied to various problems, its computational burden remains a limitation. Inspired by the use of landmarks considered in the Robust and Scalable Embedding via Landmark Diffusion (ROSELAND), we propose a variation of AD, called Landmark AD (LAD), which captures the essence of AD while offering superior computational efficiency. We provide a series of theoretical analyses of LAD under the manifold setup and apply it to the automatic sleep stage annotation problem with two electroencephalogram channels to demonstrate its application.

Key words. diffusion maps, manifold learning, multimodal data, common manifold, sensor fusion

MSC codes. 53Z50, 65D18, 62M15, 57R40, 57M50

DOI. 10.1137/24M1656682

1. Introduction. With the advancement of sensing technology, it is becoming increasingly common to collect data from multiple sensors simultaneously. Sensor fusion is a technique that involves combining data from multiple sensors to obtain an accurate, reliable, and comprehensive understanding of the underlying phenomenon or environment being monitored [20, 12]. This process aims to overcome limitations inherent in the usage of individual sensors, such as noise, bias, or limited coverage, by leveraging complementary information from multiple sources. Various sensor fusion techniques and algorithms exist, including statistical methods, machine learning algorithms, Bayesian inference, and mathematical modeling. Specific examples include linear algorithms such as the canonical correlation analysis (CCA) and its variants [16, 15, 13, 17], nonparametric canonical correlation analysis (NCCA) [29], Alternating Diffusion (AD) [22, 36] and its generalizations [18, 34, 35], time-coupled diffusion maps [28], and multiview diffusion maps [25], to mention just a few. See [34, 40, 35] and the references therein for a more comprehensive literature review. Some of these algorithms aim

*Received by the editors April 25, 2024; accepted for publication (in revised form) January 6, 2025; published electronically May 7, 2025.

<https://doi.org/10.1137/24M1656682>

Funding: The first and fourth authors are supported in part by the National Science and Technology Council grants 112-2115-M-002-015.

[†]Corresponding author. Data Science Degree Program, National Taiwan University and Academia Sinica, Taipei, 106, Taiwan (d10948003@ntu.edu.tw).

[‡]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA (hauwu@cims.nyu.edu).

[§]Viterbi Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology, Haifa, 3200003 Israel (ronen@ef.technion.ac.il).

[¶]Department of Mathematics, National Taiwan University, Taipei, 106, Taiwan, and National Center for Theoretical Sciences, Mathematics Division, Taipei, 106, Taiwan (maopei@math.ntu.edu.tw).

at extracting common information shared by the datasets acquired by the different sensors while others attempt to identify the different and unique information among them.

Our focus in this paper is centered on the diffusion-based approach, particularly the AD algorithm. AD has shown remarkable efficacy in solving a diverse array of problems, producing significant results across various fields. These applications include, but are not limited to, audio-visual voice activity detection [10], sequential audio-visual correspondence [9], automatic sleep stage identification using multiple sensors [23], and advanced interpretation of electroencephalogram (EEG) signals [26, 27]. Furthermore, AD has been instrumental in seismic event modeling [24], predicting Intelligence Quotient (IQ) using two functional Magnetic Resonance Imaging (fMRI) paradigms [38], multimicrophone speaker localization [21], fetal electrocardiogram analysis [34], and the development of multiview spectral clustering methods [31]. This diverse range of applications highlights AD's versatility and its potential for driving significant advancements in various scientific and technological domains. Additionally, various theoretical results have been established to support the efficacy of AD. For example, in [22], it is shown that AD extracts the common information under the general metric space model. Under the manifold setup, it is demonstrated in [36] that AD asymptotically converges to the Laplace–Beltrami operator of the common manifold, which is insusceptible to the sensor-specific nuisance information modeled by a product manifold structure. The Laplace–Beltrami operator might be deformed if the common information shared by the two sensors is not identical but diffeomorphically related. The behavior of AD under the null hypothesis has been reported in [7] using random matrix theory and free probability framework, showing that asymptotically the spectrum converges to a deterministic distribution. Under the same framework as in [7], various noisy setups under the nonnull hypothesis have been studied in [8], including scenarios that model situations when one sensor malfunctions.

Although AD has been extensively researched and applied in various fields, a notable limitation arises from its computational demands. This is primarily due to the eigendecomposition, which is a critical yet resource-intensive step in the algorithm. In this paper, we present a new variation of AD that is theoretically sound and retains the core principles of the original method. This variation is designed to significantly improve computational efficiency. Our approach makes use of the landmark diffusion concept originally introduced in the algorithm *Robust and Scalable Embedding via Landmark Diffusion* (ROSELAND) [32, 33] to expedite the diffusion maps algorithm, which is a central manifold learning method [6]. The concept of landmark diffusion involves constraining diffusion from one point to another using a predesigned landmark set. By leveraging this landmark set, one can alleviate the computational demand of the eigendecomposition of a large kernel matrix of the entire dataset using the eigendecomposition of the much smaller kernel matrix of the landmark set. Specifically, by forcing the diffusion on each sensor to go through a predetermined landmark set, the AD algorithm can be sped up. We refer to the proposed algorithm as *Landmark Alternating Diffusion* (LAD). Besides demonstrating its computational superiority, we provide a series of theoretical analyses under the manifold framework. Furthermore, we showcase the application of LAD to the automatic sleep stage annotation problem, illustrating its practical utility.

The remainder of this paper is organized as follows. In section 2, we review AD. In section 3.1, we present the proposed LAD algorithm. Justification of LAD on the linear algebra level is given in section 3.2. In section 4, we analyze the asymptotic behavior of LAD under

a manifold setting, which equates LAD and AD on the asymptotical region. This analysis also provides us with a method to choose the hyperparameters of LAD. In section 5, we apply Landmark AD to several simulation datasets to validate the theory proposed in section 4. In section 6, LAD is applied to the sleep dataset by fusing signals from two channels to analyze sleep stages. The supplementary material (supplement.pdf [local/web 988KB]) provides the proof of the main theorems.

2. Background on alternating diffusion (AD). We start with a quick review of the AD algorithm. Suppose we have n pairs of measurements recorded from two sensors simultaneously, denoted as $\mathcal{X} := \{(s_i, r_i)\}_{i=1}^n \subset (\mathfrak{M}_1, d_1) \times (\mathfrak{M}_2, d_2)$, where $(\mathfrak{M}_1, d_1)$ and $(\mathfrak{M}_2, d_2)$ are metric spaces; that is, the dataset $\mathcal{X}_1 := \{s_i\}_{i=1}^n \subset \mathfrak{M}_1$ is from the first sensor, and $\mathcal{X}_2 := \{r_i\}_{i=1}^n \subset \mathfrak{M}_2$ is from the second sensor. Take two nonnegative smooth kernel functions $\tilde{K}^{(1)}, \tilde{K}^{(2)} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_+$ that decay sufficiently fast. To simplify the discussion, we assume $\tilde{K}^{(1)}$ and $\tilde{K}^{(2)}$ are Gaussian functions $\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$, but other kernels can be considered. Define the affinity matrices $\mathbf{W}^{(1)}, \mathbf{W}^{(2)} \in \mathbb{R}^{n \times n}$ for the first and second sensors as

$$\mathbf{W}_{ij}^{(1)} = \tilde{K}^{(1)}\left(\frac{d^{(1)}(s_i, s_j)^2}{\epsilon_1}\right) \quad \text{and} \quad \mathbf{W}_{ij}^{(2)} = \tilde{K}^{(2)}\left(\frac{d^{(2)}(r_i, r_j)^2}{\epsilon_2}\right),$$

where $\epsilon_1, \epsilon_2 > 0$ are kernel bandwidths that can be chosen by the user. Define the associated diagonal degree matrices by $\mathbf{D}_{ii}^{(\ell)} = \sum_{j=1}^n \mathbf{W}_{ij}^{(\ell)}$, where $\ell = 1, 2$, and the diffusion matrices

$$\mathbf{M}^{(\ell)} = \mathbf{D}^{(\ell)-1} \mathbf{W}^{(\ell)}.$$

The *alternating diffusion matrix* introduced in [22, 36] is given by

$$(2.1) \quad \mathbf{M} = \mathbf{M}^{(1)} \mathbf{M}^{(2)} \in \mathbb{R}^{n \times n}.$$

$\mathbf{M}$ is also a diffusion matrix in the sense that the sums of rows equal 1 and the spectral norm is bounded by 1, but, in general, it might not be diagonalizable.

Suppose the eigenvalue decomposition (EVD) of $\mathbf{M} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1}$ exists, where $\mathbf{U} \in GL(n)$ and $\mathbf{\Lambda}$ is a diagonal matrix. Denote the eigenvalues satisfying $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and the associated eigenvectors $u_1, u_2, \dots, u_n$. Choose $1 \leq q \leq n$, and denote $\mathbf{U}_q = [u_2, \dots, u_{q+1}] \in \mathbb{R}^{n \times q}$ and corresponding eigenvalues $\mathbf{\Lambda}_q = \text{diag}(\lambda_2, \lambda_3, \dots, \lambda_{q+1})$. We then embed the dataset into a q -dim Euclidean space via Ψ_t by

$$\Psi_t : (s_i, r_i) \mapsto e_i^\top \mathbf{U}_q \mathbf{\Lambda}_q^t \in \mathbb{R}^q,$$

where e_j is a unit column vector with 1 in the j th entry, $\top$ means the transpose, and $t > 0$ is the diffusion time chosen by the user. Ψ_t is the AD *with the initial diffusion on the second sensor*, or in short the AD. While the AD has been applied to a broad range of problems, e.g., [23, 18, 26, 21], it is limited by its computational time, which is governed by the EVD. Note that even if $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ are both sparse, which might speed up the EVD, the product in (2.1) hinders the sparsity structure.

3. Proposed landmark alternating diffusion algorithm. In this section, we present the proposed LAD algorithm with initial theoretical support from a linear algebra viewpoint.

3.1. Landmark alternating diffusion. To alleviate the computational burden and speed up AD, inspired by a recent landmark diffusion algorithm ROSELAND [32, 33], we propose a new algorithm, called *landmark AD* (LAD).

Take a set $\tilde{\mathcal{Z}} := \{(a_k, b_k)\}_{k=1}^m \subset (\mathfrak{M}_1, d_1) \times (\mathfrak{M}_2, d_2)$, where $m < n$. We call $\tilde{\mathcal{Z}}$ the *landmark set*, $\tilde{\mathcal{Z}}_1 := \{a_k\}_{k=1}^m \subset (\mathfrak{M}_1, d_1)$ the landmark set of the first sensor, and $\tilde{\mathcal{Z}}_2 := \{b_k\}_{k=1}^m \subset (\mathfrak{M}_2, d_2)$ the landmark set of the second sensor. In practice, $\tilde{\mathcal{Z}}$ could be a subset of $\mathcal{X}$. The choice of landmark points is user-defined, and different selections result in different embeddings, as demonstrated in the theoretical analysis in section 4 and the experimental results in section 5. See these sections for guidance on selecting an appropriate landmark set. With the selected landmark points, construct two *landmark affinity matrices* $\bar{\mathbf{W}}^{(1)}, \bar{\mathbf{W}}^{(2)} \in \mathbb{R}^{n \times m}$, which are defined by

$$(3.1) \quad \bar{\mathbf{W}}_{ik}^{(1)} = \tilde{K}^{(1)} \left(\frac{d^{(1)}(s_i, a_k)^2}{\epsilon_1} \right) \quad \text{and} \quad \bar{\mathbf{W}}_{ik}^{(2)} = \tilde{K}^{(2)} \left(\frac{d^{(2)}(r_i, b_k)^2}{\epsilon_2} \right).$$

Construct a diagonal matrix $\bar{\mathbf{D}}^{(2)}$ as follows:

$$(3.2) \quad \bar{\mathbf{D}}^{(2)} = \text{diag}(\bar{\mathbf{W}}^{(2)\top} \bar{\mathbf{W}}^{(2)} \mathbf{1}_m),$$

where $\mathbf{1}_m \in \mathbb{R}^{m \times 1}$ is a column vector with all entries 1. Then, fix a *normalization factor* $\alpha \in [0, 1]$ and define the *landmark diffusion matrix* of the second sensor

$$(3.3) \quad \bar{\mathbf{M}}_\alpha^{(2)} := \bar{\mathbf{W}}^{(2)} (\bar{\mathbf{D}}^{(2)})^{-\alpha} \in \mathbb{R}^{n \times m}.$$

Next, construct a diagonal matrix $\bar{\mathbf{D}}_\alpha^{(1)}$ as

$$(3.4) \quad \bar{\mathbf{D}}_\alpha^{(1)} = \text{diag}(\bar{\mathbf{W}}^{(1)} \bar{\mathbf{M}}_\alpha^{(2)\top} \mathbf{1}_n) \in \mathbb{R}^{n \times n}.$$

Then, the *landmark diffusion matrix* of the first sensor are defined as

$$(3.5) \quad \bar{\mathbf{M}}_\alpha^{(1)} = (\bar{\mathbf{D}}_\alpha^{(1)})^{-1} \bar{\mathbf{W}}^{(1)} \in \mathbb{R}^{n \times m}.$$

The *landmark AD matrix* is defined as

$$\bar{\mathbf{M}}_\alpha = \bar{\mathbf{M}}_\alpha^{(1)} \bar{\mathbf{M}}_\alpha^{(2)\top} \in \mathbb{R}^{n \times n}.$$

To speed up the computation, consider the matrix $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)} \in \mathbb{R}^{m \times m}$, compute the eigenstructure of $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)}$ on m landmarks and extrapolate to the whole database in the following way. Denote the EVD of $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)}$ by

$$(3.6) \quad \bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)} = \bar{\mathbf{V}} \bar{\mathbf{\Lambda}} \bar{\mathbf{V}}^{-1} \in \mathbb{R}^{m \times m},$$

where the diagonal entries of $\bar{\mathbf{\Lambda}}$ are $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \dots \geq \bar{\lambda}_{m'} > \bar{\lambda}_{m'+1} = \dots = \bar{\lambda}_m = 0$. Set $\bar{\mathbf{U}} = \bar{\mathbf{M}}_\alpha^{(1)} \bar{\mathbf{V}} \in \mathbb{R}^{n \times m}$. Choose $1 \leq q \leq m'$. Let $\bar{\mathbf{U}}_q \in \mathbb{R}^{n \times q}$ contain the first q columns of $\bar{\mathbf{U}}$ and $\bar{\mathbf{\Lambda}}_q$ a diagonal matrix containing the associated eigenvalues. The α -normalized LAD with the initial diffusion on the second sensor, or α -LAD in short, is defined as

$$\bar{\Psi}_t : (s_i, r_i) \mapsto e_i^\top \bar{\mathbf{U}}_q (\bar{\mathbf{\Lambda}}_q)^t \in \mathbb{R}^q,$$

Algorithm 3.1. LAD.

Input: Two datasets, $\{s_i\}_{i=1}^n \subset (\mathfrak{M}_1, d_1)$ and $\{r_i\}_{i=1}^n \subset (\mathfrak{M}_2, d_2)$, are sampled simultaneously from two sensors. Choose $\epsilon_1, \epsilon_2 > 0$, embedding dimension $0 < q < m$, diffusion time $t > 0$, $\alpha \in [0, 1]$, and the landmark set $\{(a_k, b_k)\}_{k=1}^m$.

- 1: Build the $n \times m$ landmark affinity matrix $\bar{\mathbf{W}}^{(1)}$ by $\bar{\mathbf{W}}_{ij}^{(1)} = \tilde{K}_{\epsilon_1}^{(1)}(s_i, a_k)$;
- 2: Build the $n \times m$ landmark affinity matrix $\bar{\mathbf{W}}^{(2)}$ by $\bar{\mathbf{W}}_{ij}^{(2)} = \tilde{K}_{\epsilon_2}^{(2)}(r_i, b_k)$;
- 3: Construct $\bar{\mathbf{D}}^{(2)} = \text{diag}(\bar{\mathbf{W}}^{(2)\top} \bar{\mathbf{W}}^{(2)} \mathbf{1}_m)$;
- 4: Construct the landmark diffusion matrix $\bar{\mathbf{M}}_\alpha^{(2)} = \bar{\mathbf{W}}^{(2)} (\bar{\mathbf{D}}^{(2)})^{-\alpha}$;
- 5: Construct the $n \times n$ diagonal landmark degree matrices $\bar{\mathbf{D}}_\alpha^{(1)}$ by $\bar{\mathbf{D}}_\alpha^{(1)} = \text{diag}(\bar{\mathbf{W}}^{(1)} \bar{\mathbf{M}}_\alpha^{(2)\top} \mathbf{1}_n)$;
- 6: Construct the landmark diffusion matrix $\bar{\mathbf{M}}_\alpha^{(1)} = (\bar{\mathbf{D}}_\alpha^{(1)})^{-1} \bar{\mathbf{W}}^{(1)}$;
- 7: Apply EVD on the matrix $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)} = \mathbf{V} \bar{\Lambda} \mathbf{V}^{-1}$ where eigenvalues are $\bar{\lambda}_1 \geq \dots \geq \bar{\lambda}_m$;
- 8: Set $\bar{\mathbf{U}} = \bar{\mathbf{M}}_\alpha^{(1)} \mathbf{V} = [\bar{u}_1 \dots \bar{u}_m]$; {Approximate $\mathbf{U}$ }
- 9: Set $\bar{\mathbf{U}}_q = [\bar{u}_2 \dots \bar{u}_{q+1}]$ and $\bar{\Lambda}_q = [\bar{\lambda}_2 \dots \bar{\lambda}_{q+1}]$;

Output: $e_i^\top \bar{\mathbf{U}}_q \bar{\Lambda}_q^t$ as the embedded point of (s_i, r_i) , where $i = 1, \dots, n$.

where $t > 0$ is the diffusion time chosen by the user. The output is the diffusion result on $\{r_i\}_{i=1}^n$. The pseudocode of α -LAD is summarized in Algorithm 3.1.

The computational complexity of evaluating the landmark affinity matrices and landmark diffusion matrices, getting $\bar{\mathbf{D}}_\alpha^{(1)}$, $\bar{\mathbf{D}}^{(2)}$, obtaining $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)}$ and the EVD of $\bar{\mathbf{M}}_\alpha^{(2)\top} \bar{\mathbf{M}}_\alpha^{(1)}$ are $\Theta(nm)$, $O(nm)$, $O(nm^2)$, and $O(m^{2.81})$, respectively. Therefore, the overall computational complexity of α -LAD is $\Theta(nm) + O(nm^2)$ since we assume $m < n$. Eventually, if $m < n^{1/2}$, α -LAD has a preferable computational complexity compared with AD.

3.2. Justification of LAD from a linear algebra perspective. Suppose we have a vector $v \in \mathbb{R}_{\geq 0}^n$ of mass distribution on the measurements from the second sensor such that $v^\top \mathbf{1}_n = 1$. Geometrically, we could understand AD as two diffusion steps of v . The first step is a diffusion on the second dataset, i.e., $\mathbf{M}^{(2)}v$. Then, the result is bijectively transferred to the first dataset by considering the range of $\mathbf{M}^{(2)}$ as the domain of $\mathbf{M}^{(1)}$. The second diffusion step is then applied on the first dataset, i.e., $\mathbf{M}^{(1)}\mathbf{M}^{(2)}v$. It is noteworthy that when $\mathcal{X}_1 = \mathcal{X}_2$, AD reduces to a single diffusion step with double diffusion time.

To speed up the computation time, α -LAD attempts to approximate the effective diffusion process above by incorporating intermittent diffusion steps on the landmarks. The initial diffusion is transferred from $\mathcal{X}_2$ to $\tilde{\mathcal{Z}}_2$ by considering $\bar{\mathbf{M}}_\alpha^{(2)}v$. Then, the result is bijectively transferred from $\tilde{\mathcal{Z}}_2$ to $\tilde{\mathcal{Z}}_1$. Finally, diffusion from $\tilde{\mathcal{Z}}_1$ to $\mathcal{X}_1$ is applied, i.e., $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)}v$. Overall, we obtain a diffusion from $\mathcal{X}_2$ to $\mathcal{X}_1$ as in AD, but the intermittent diffusion steps are performed on the landmarks.

First, we show that the landmark AD matrix $\bar{\mathbf{M}}_\alpha$ is a diffusion matrix. By construction, based on the kernels, all the entries of $\bar{\mathbf{M}}_\alpha$ are nonnegative. By (3.4) and (3.5), we have

$$\begin{aligned} \bar{\mathbf{M}}_\alpha^{(1)} \bar{\mathbf{M}}_\alpha^{(2)\top} \mathbf{1}_n &= \text{diag}(\bar{\mathbf{W}}^{(1)} \bar{\mathbf{M}}_\alpha^{(2)\top} \mathbf{1}_n)^{-1} \bar{\mathbf{W}}^{(1)} (\bar{\mathbf{D}}^{(2)})^{-\alpha} \bar{\mathbf{W}}^{(2)} \mathbf{1}_n \\ (3.7) \quad &= \text{diag}(\bar{\mathbf{W}}^{(1)} (\bar{\mathbf{D}}^{(2)})^{-\alpha} \bar{\mathbf{W}}^{(2)} \mathbf{1}_n)^{-1} \bar{\mathbf{W}}^{(1)} (\bar{\mathbf{D}}^{(2)})^{-\alpha} \bar{\mathbf{W}}^{(2)} \mathbf{1}_n = \mathbf{1}_n, \end{aligned}$$

which concludes that $\bar{\mathbf{M}}_\alpha$ is a diffusion matrix. Moreover, the above calculation shows that the L^∞ norm of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ is 1. Since the spectral norm is bounded by the L^∞ norm, all eigenvalues, which might be complex, are bounded by 1. When all the eigenvalues of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ are simple, the EVD exists. However, in general, the matrix might be defective. In section 4, we will show that under the manifold setup, asymptotically, $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ pointwisely converges to a self-adjoint operator. By this result and numerical confirmation, we hypothesize that the eigenvalues of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ are asymptotically real, and $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ is asymptotically nondefective.

Next, we discuss the relationship between the spectral structures of $\bar{\mathbf{M}}_\alpha^{(2)\top}\bar{\mathbf{M}}_\alpha^{(1)}$ and $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$. Suppose $\bar{\mathbf{M}}_\alpha^{(1)}$ is of full rank. Let $\mathbf{V} \in Gl(m)$ be the eigenvector matrix of $\bar{\mathbf{M}}_\alpha^{(2)\top}\bar{\mathbf{M}}_\alpha^{(1)}$. By the rank-nullity theorem, the null space of $\bar{\mathbf{M}}_\alpha^{(1)}$ is degenerate and hence $\bar{\mathbf{U}} := \bar{\mathbf{M}}_\alpha^{(1)}\mathbf{V}$ is of full rank as well. By a direct calculation, the columns of $\bar{\mathbf{U}}$ form the m right eigenvectors of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$. If the rank of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ is m , then we obtain all the nontrivial right eigenvectors. Note that, in general, even if the rank of $\bar{\mathbf{M}}_\alpha^{(1)}$ is less than m , we can still recover the nontrivial eigenvectors of $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ from the nontrivial eigenvectors of $\bar{\mathbf{M}}_\alpha^{(2)\top}\bar{\mathbf{M}}_\alpha^{(1)}$. Overall, we get that we can speed up the computation of the nontrivial eigenvectors of the $n \times n$ matrix $\bar{\mathbf{M}}_\alpha^{(1)}\bar{\mathbf{M}}_\alpha^{(2)\top}$ using α -LAD by considering the eigendecomposition of the smaller $m \times m$ matrix $\bar{\mathbf{M}}_\alpha^{(2)\top}\bar{\mathbf{M}}_\alpha^{(1)}$.

We remark that the normalization factor $\alpha \in [0, 1]$ plays a similar role as in diffusion maps [6]. We will systematically study its effect in section 4 under the manifold setup. Broadly, we will show that under the manifold setup, when $\alpha = 1$, the impact of the landmark distribution is minimized. Conversely, when $\alpha = 0$, 0-LAD is impacted by the landmark distribution. In the special case that $\alpha = 0$ and $\mathcal{X}_1 = \mathcal{X}_2$, 0-LAD is reduced to ROSELAND [32, 33].

4. Asymptotic analysis of LAD under the manifold setup. To further understand LAD and connect it with the original AD, thereby showing that it recovers the common manifold, we continue to study the asymptotic behavior of LAD under the manifold setup.

4.1. Manifold setting. Suppose there is a common structure that two sensors aim to capture, represented by a d -dimensional compact and smooth manifold $\mathcal{M}$ without boundary. This manifold is referred to as the common manifold. In practice, different sensors may be of different modalities and capture different “views” or information about the common manifold. These differences are captured by different embeddings of $\mathcal{M}$. We mention that it is possible to consider a more complicated model that includes sensor-specific variables as in [22]. Let $\mathcal{M}$ be smoothly embedded into $\mathbb{R}^{p_1}$ via $\iota^{(1)}$ and into $\mathbb{R}^{p_2}$ via $\iota^{(2)}$, where p_1 and p_2 may not be equal. Here, $\iota_{(\ell)}$ encodes the features of the ℓ th sensor, and $\mathbb{R}^{p_\ell}$ is the space that hosts the acquired information. The collected dataset is modeled by independently and identically distributed (i.i.d.) samples from two random vectors, $X^{(\ell)} := \iota^{(\ell)} \circ X$, where $X : (\Omega, \mathcal{F}, P) \rightarrow \mathcal{M}$ is a random vector with induced measure on $\mathcal{M}$, denoted by $\nu := X_*P$, defined on the Borel sigma algebra on $\mathcal{M}$. It is important to note that we cannot directly access $\mathcal{M}$, but rather observe it through a dataset collected by two sensors, $\{(s_i, r_i)\}_{i=1}^n$ where $s_i = \iota^{(1)}(x_i) \in \mathbb{R}^{p_1}$ and $r_i = \iota^{(2)}(x_i) \in \mathbb{R}^{p_2}$. In contrast to many sensor fusion techniques that merely aim to merge the multisensor information, in LAD, as in AD, the ultimate goal is to recover the common manifold by recovering $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathcal{M}$ from the dataset $\{(s_i, r_i)\}_{i=1}^n$.

Since $\iota^{(\ell)}$ is a smooth embedding, $\iota^{(1)}(\mathcal{M})$ and $\iota^{(2)}(\mathcal{M})$ diffeomorphic to each other and $\Phi := \iota^{(1)} \circ \iota^{(2)^{-1}}$ is a diffeomorphism. Assume the metric of $\iota^{(\ell)}(\mathcal{M})$, denoted as $g^{(\ell)}$, is induced from the canonical metric of $\mathbb{R}^p$. Let $dV^{(\ell)}$ be the measure associated with the Riemannian volume form induced from $g^{(\ell)}$. Assume $\nu_{\mathcal{M}}$ is absolutely continuous with respect to $dV^{(\ell)}$. By the Radon–Nikodym theorem, there exists a probability density function (p.d.f.) $p^{(\ell)}$ on $\mathcal{M}$ such that $d\nu = p^{(\ell)} dV^{(\ell)}$.

The landmark set $\mathcal{Z}$ is modeled in a similar way. Assume $\mathcal{Z} = \{z_k\}_{k=1}^m$ is sample from a $\mathcal{M}$ -valued random variable $Z : (\Omega, \mathcal{F}, P) \rightarrow \mathcal{M}$ with induced measure $\nu^{\mathcal{Z}} = Z_* P$. Assume Z is independent of X . By the Radon–Nikodym theorem, there exists a p.d.f. $p_{\mathcal{Z}}^{(\ell)}$ of $Z^{(\ell)}$ on $\mathcal{M}$ such that $d\nu^{\mathcal{Z}} = p_{\mathcal{Z}}^{(\ell)} dV^{(\ell)}$, where $\ell = 1, 2$.

With the sampled data $\{(s_i, r_i)\}_{i=1}^n \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and the landmark set $\{(a_i, b_i)\}_{i=1}^m \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, where $a_i = \iota^{(1)}(z_i)$ and $b_i = \iota^{(2)}(z_i)$, we can run LAD. Below we provide a series of analyses studying the behavior of LAD under the manifold setup. Note that while, in general, we can consider embeddings into more general metric spaces, in this analysis we focus on $(\mathfrak{M}_{\ell}, d_{\ell}) = (\mathbb{R}^{p_{\ell}}, \|\cdot\|)$ for $\ell = 1, 2$.

With the above model, to proceed with the analysis, we specify the following assumptions.

Assumption 4.1.

1. The d -dimensional manifold $\mathcal{M}$ is compact without boundary and embedded into $\mathbb{R}^{p_{\ell}}$ via $C^4(\mathcal{M})$ map $\iota^{(\ell)}$.
2. The kernel functions are C^3 , positive and decay exponentially fast; that is, exist $c_1, c_2 > 0$ such that $\tilde{K}^{(\ell)}(t) < c_1 e^{-c_2 t^2}$ and $|\partial_t \tilde{K}^{(\ell)}(t)| < c_1 e^{-c_2 t^2}$. Furthermore, define $\mu_{r,k}^{(\ell)} := \int_{\mathbb{R}^d} \|x\|^r \partial^{(k)} \tilde{K}^{(\ell)}(\|x\|) dx$. Assume $\mu_{0,0}^{(\ell)} = 1$ for $\ell = 1, 2$.
3. ν is absolutely continuous with respect to $dV^{(\ell)}$. Moreover, $p^{(\ell)} \in C^4(\mathcal{M})$ and is bounded below from zero.

Without loss of generality, we analyze LAD assuming it starts from the second sensor. A similar analysis holds when it starts from the first sensor. First, we define the degree function on the second manifold. To simplify the tedious notation, we systematically use $K_{\epsilon}^{(\ell)}(x_i, x_j) := \tilde{K}^{(\ell)}(\frac{\|\iota^{(\ell)}(x_i) - \iota^{(\ell)}(x_j)\|^2}{\epsilon})$ for $\ell = 1, 2$.

Definition 4.2. Define the landmark kernel function $K_{\text{lan},\epsilon}^{(2)} : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}_+$ as

$$K_{\text{lan},\epsilon}^{(2)}(z, z') := \int_{\mathcal{M}} K_{\epsilon}^{(2)}(z, x) K_{\epsilon}^{(2)}(x, z') d\nu(x),$$

and define the landmark degree function, denoted as $d_{\text{lan},\epsilon}^{(2)} : \mathcal{M} \rightarrow \mathbb{R}_+$, by

$$d_{\text{lan},\epsilon}^{(2)}(z) = \int_{\mathcal{M}} K_{\text{lan},\epsilon}^{(2)}(z, z') d\nu^{\mathcal{Z}}(z').$$

First, recall the following result from [33, Theorem 10], which states that the quantity $d_{\text{lan},\epsilon}^{(2)}$ approximates the quantity $p^{(2)} p_{\mathcal{Z}}^{(2)}$.

Theorem 4.3 (Theorem 10 in [33]). For $z \in \mathcal{M}$, we have

$$\begin{aligned} d_{\text{lan},\epsilon}^{(2)}(z) &= \int_{\mathcal{M}} K_{\text{ref},\epsilon}^{(2)}(z, z') p_{\mathcal{Z}}^{(2)}(z') dV^{(2)}(z') \\ &= \epsilon^d \left[p^{(2)}(z) p_{\mathcal{Z}}^{(2)}(z) + \epsilon \frac{\mu_{2,0}^{(2)}}{d} \left(p_{\mathcal{Z}}^{(2)}(z) \Delta p^{(2)}(z) + \frac{1}{2} p^{(2)}(z) \Delta p_{\mathcal{Z}}^{(2)}(z) \right. \right. \\ &\quad \left. \left. - w(z) p^{(2)}(z) p_{\mathcal{Z}}^{(2)}(z) + \nabla p^{(2)}(z) \cdot \nabla p_{\mathcal{Z}}^{(2)}(z) \right) \right] + \mathcal{O}(\epsilon^{d+2}) \\ &:= \epsilon^d p^{(2)}(z) p_{\mathcal{Z}}^{(2)}(z) (1 + \epsilon \bar{E}(z)) + \mathcal{O}(\epsilon^{d+2}), \end{aligned}$$

where

$$\bar{E}(x) := \frac{\Delta p^{(2)}(x)}{p^{(2)}(x)} + \frac{\Delta p_{\mathcal{Z}}^{(2)}(x)}{2p_{\mathcal{Z}}^{(2)}(x)} + \frac{\nabla p^{(2)}(x) \cdot \nabla p_{\mathcal{Z}}^{(2)}(x)}{p^{(2)}(x) p_{\mathcal{Z}}^{(2)}(x)} - w(x),$$

$w(x) = \frac{1}{3}s(x) - \frac{d}{12|S^{d-1}|} \int_{S^{d-1}} [\Pi_x^{(2)}(\theta, \theta)]^2 d\theta$, $s(x)$ is the scalar curvature at x , and $\Pi_x^{(2)}$ is the second fundamental form of the embedding at x , and $|S^{d-1}|$ is the volume of the canonical $(d-1)$ -dim sphere.

Next, recall that the landmark kernel induced by the landmark set depends on the normalization factor α . We have the following definition.

Definition 4.4 (α -LAD kernel). The α -landmark alternating diffusion kernel, $K_{\text{lan},\epsilon,\alpha} : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}_+$, is defined by

$$K_{\text{lan},\epsilon,\alpha}(x, y) := \int_{\mathcal{M}} \frac{K_{\epsilon}^{(1)}(x, z') K_{\epsilon}^{(2)}(z', y)}{[d_{\text{lan},\epsilon}^{(2)}(z')]^{\alpha}} d\nu^{\mathcal{Z}}(z').$$

The α -landmark alternating degree function, $d_{\text{lan},\epsilon,\alpha} : \mathcal{M} \rightarrow \mathbb{R}_+$, is defined by

$$d_{\text{lan},\epsilon,\alpha}(x) = \int_{\mathcal{M}} K_{\text{lan},\epsilon,\alpha}(x, y) d\nu(y).$$

With the α -landmark alternating diffusion kernel and α -landmark alternating degree function, we can define the LAD operator.

Definition 4.5 (LAD operator). The LAD operator, $T_{\text{lan},\epsilon} : C(\mathcal{M}) \rightarrow C(\mathcal{M})$, is defined as

$$T_{\text{lan},\epsilon,\alpha} f(x) = \int_{\mathcal{M}} \frac{K_{\text{lan},\epsilon,\alpha}(x, y)}{d_{\text{lan},\epsilon,\alpha}(x)} f(y) d\nu(y).$$

It is clear that $T_{\text{lan},\epsilon,\alpha} 1 = 1$, where 1 is a constant function with value 1.

It is immediate to see that the above quantities are continuous companions of LAD matrices used in the LAD algorithm. Denote $\nu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ to be the empirical measure associated with the samples on the common manifold and $\nu_m^{\mathcal{Z}} = \frac{1}{m} \sum_{i=1}^m \delta_{z_i}$ to be the empirical measure associated with the landmark set. Note that $d_{\text{lan},\epsilon}^{(2)}(z_i) = \int_{\mathcal{M}} K_{\text{lan},\epsilon}^{(2)}(z_i, z') d\nu^{\mathcal{Z}}(z')$

is the continuous version of $d_{\text{lan},\epsilon,m}^{(2)}(z_i) := \int_{\mathcal{M}} K_{\text{lan},\epsilon,n}^{(2)}(z_i, z') d\nu_m^{\mathcal{Z}}(z')$, where $K_{\text{lan},\epsilon,n}^{(2)}(x_i, z') := \int_{\mathcal{M}} K_{\epsilon}^{(2)}(z, x) K_{\epsilon}^{(2)}(x, z') d\nu_n(x)$. By a direct expansion, we have the diagonal entry $\bar{\mathbf{D}}_{ii}^{(2)} = \int_{\mathcal{M}} K_{\text{lan},\epsilon,n}^{(2)}(x_i, z') d\nu_m^{\mathcal{Z}}(z')$. Similarly, $d_{\text{lan},\epsilon,\alpha}(x_i) = \int_{\mathcal{M}} K_{\text{lan},\epsilon,\alpha}(x_i, y) d\nu(y)$ is the continuous version of $\int_{\mathcal{M}} K_{\text{lan},\epsilon,\alpha,m}(x_i, y) d\nu_n(y)$, where $K_{\text{lan},\epsilon,\alpha,m}(x, y) := \int_{\mathcal{M}} \frac{K_{\epsilon}^{(1)}(x, z') K_{\epsilon}^{(2)}(z', y)}{[d_{\text{lan},\epsilon,m}^{(2)}(z')]^{\alpha}} d\nu_m^{\mathcal{Z}}(z')$. By a direct expansion, we have $\bar{\mathbf{D}}_{\alpha,ii}^{(1)} = \int_{\mathcal{M}} K_{\text{lan},\epsilon,\alpha,n}(x_i, y) d\nu_n(y)$. Along similar lines, we see that $K_{\text{lan},\epsilon,\alpha}(x_i, x_j)$ is the continuous version of $(\bar{\mathbf{W}}^{(1)} \bar{\mathbf{M}}_{\alpha}^{(2)\top})_{ij}$ and $T_{\text{lan},\epsilon,\alpha}$ is the continuous version of the landmark AD matrix $\bar{\mathbf{M}}_{\alpha} = \bar{\mathbf{M}}_{\alpha}^{(1)} \bar{\mathbf{M}}_{\alpha}^{(2)\top}$.

The main theoretical result is about the asymptotic behavior of $\bar{\mathbf{M}}_{\alpha}$, which is composed of the variance analysis and bias analysis. In short, in the variance analysis, asymptotically when $n \rightarrow \infty$, we quantify the stochastic fluctuation of the convergence of $\bar{\mathbf{M}}_{\alpha}$ to the integral operator $T_{\text{lan},\epsilon}$ caused by finite sampling points. In the bias analysis, asymptotically when $\epsilon \rightarrow 0$, we show that $T_{\text{lan},\epsilon}$ converges to an elliptic operator defined on $(\mathcal{M}, g^{(2)})$ and quantify the convergence rate. The proof of Theorem 4.6 is postponed to the supplement.

Theorem 4.6. Fix $x \in \mathcal{M}$. Fix normal coordinates around x associated with $g^{(\ell)}$ and the associated exponential maps $\exp^{(\ell)}$, where $\ell = 1, 2$. Let $\mathcal{B} := \{E_i\}_{i=1}^d \subset T_x \mathcal{M}$ be an orthonormal basis associated with $g^{(2)}$. Set

$$R_x = \left[d \exp^{(1)}|_0 \right]^{-1} [d\iota^{(1)}]^{-1} \nabla \Phi[d\iota^{(2)}] \left[d \exp^{(2)}|_0 \right] \in \mathbb{R}^{d \times d},$$

with all quantities represented with the basis $\mathcal{B}$. By the singular value decomposition (SVD), we have $R_x = U_x \Lambda_x V_x^T$, where $\Lambda_x = \text{diag}[\lambda_1, \dots, \lambda_d]$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, and $U_x, V_x \in O(d)$. Then, for a $f \in C^3(\mathcal{M})$, when $\epsilon > 0$ is sufficiently small, we have

$$\begin{aligned} T_{\text{lan},\epsilon,\alpha} f(x) &= f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \Delta^{(2)} f(x) + \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) \\ &\quad + \frac{\epsilon \mu_{2,0}^{(1)}}{d} \sum_{i=1}^d \lambda_i \left(\frac{\nabla_{E_i}^{(2)} p^{(2)}(x)}{p^{(2)}} + \frac{\nabla_{E_i}^{(2)} q_{\alpha}(x)}{q_{\alpha}(x)} \right) \nabla_{E_i}^{(2)} f(x) \\ &\quad + \frac{\epsilon \mu_{2,0}^{(2)}}{d} \frac{\nabla^{(2)} p^{(2)}(x) \cdot \nabla^{(2)} f(x)}{p^{(2)}(x)} + \mathcal{O}(\epsilon^{3/2}), \end{aligned}$$

where $\nabla^{(\ell)}$ is the covariant derivative associated with $\iota^{(\ell)}$, $q_{\alpha}(x) := \frac{p_{\mathcal{Z}}^{(2)}(x)^{1-\alpha}}{p^{(2)}(x)^{\alpha}}$ and the implied constant depends on the scalar curvature at x , second fundamental form at x , C^3 norm of f , and C^2 norms of $p^{(2)}$ and $p_{\mathcal{Z}}^{(2)}$.

By a direct calculation, asymptotically when $\epsilon \rightarrow 0$, $\frac{2d}{\mu_{2,0}^{(2)}} \frac{T_{\text{lan},\epsilon,\alpha} - I}{\epsilon}$ converges to

$$\mathcal{L}_{\alpha} := \Delta^{(2)} + 2 \frac{\nabla^{(2)} p^{(2)}}{p^{(2)}} \cdot \nabla^{(2)} + \frac{\mu_{2,0}^{(1)}}{\mu_{2,0}^{(2)}} \left\{ \sum_{i=1}^d \lambda_i \left[\nabla_{E_i E_i}^{(2)^2} + 2 \left(\frac{\nabla_{E_i}^{(2)} p^{(2)}}{p^{(2)}} + \frac{\nabla_{E_i}^{(2)} q_{\alpha}}{q_{\alpha}} \right) \cdot \nabla_{E_i}^{(2)} \right] \right\},$$

which is self-adjoint in nature, and involves not only the intrinsic operator, the Laplace–Beltrami operator $\Delta^{(2)}$, but also first-order terms that depend on the sampling densities

and landmark densities. The theorem immediately leads to the following corollaries, which demonstrate the connection between LAD and other related algorithms.

Corollary 4.7. Consider $p^{(2)} = p_{\mathcal{Z}}^{(2)}$. If we take $\alpha = 1/2$, then $q_\alpha = 1$ and Theorem 4.6 is reduced to original AD; that is, for $f \in C^3(\mathcal{M})$, we have

$$\begin{aligned} T_{\text{lan}, \epsilon, \alpha} f(x) &= f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \Delta^{(2)} f(x) + \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) \\ &\quad + \frac{\epsilon \mu_{2,0}^{(1)}}{d} \left(\sum_{i=1}^d \lambda_i \frac{\nabla_{E_i}^{(2)} p^{(2)}(x) \nabla_{E_i}^{(2)} f(x)}{p^{(2)}(x)} \right) + \frac{\epsilon \mu_{2,0}^{(2)}}{d} \frac{\nabla^{(2)} p^{(2)}(x) \cdot \nabla^{(2)} f(x)}{p^{(2)}(x)} + \mathcal{O}(\epsilon^{3/2}). \end{aligned}$$

If we further assume $p^{(2)}$ is constant; that is, the sampling is uniform, then we obtain

$$T_{\text{lan}, \epsilon, \alpha} f(x) = f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \Delta^{(2)} f(x) + \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) + \mathcal{O}(\epsilon^{3/2}).$$

This corollary implies that if the landmark distribution is the same as the data sample distribution; that is, $p^{(2)} = p_{\mathcal{Z}}^{(2)}$, then if we take $\alpha = 1/2$, the asymptotic operator is the same as that of AD.

Corollary 4.8. Consider $\alpha = 1$. Then $q_\alpha(x) = \frac{1}{p^{(2)}(x)}$, which is independent on $p_{\mathcal{Z}}$. For $f \in C^3(\mathcal{M})$, we have

$$\begin{aligned} T_{\text{lan}, \epsilon, \alpha} f(x) &= f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \Delta^{(2)} f(x) + \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) \\ &\quad + \frac{\epsilon \mu_{2,0}^{(2)}}{d} \frac{\nabla^{(2)} p^{(2)}(x) \cdot \nabla^{(2)} f(x)}{p^{(2)}(x)} + \mathcal{O}(\epsilon^{3/2}). \end{aligned}$$

If we further assume $p^{(2)}$ is constant; that is, the sampling is uniform, then we obtain

$$T_{\text{lan}, \epsilon, \alpha} f(x) = f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \Delta^{(2)} f(x) + \frac{\epsilon \mu_{2,0}^{(1)}}{2d} \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) + \mathcal{O}(\epsilon^{3/2}).$$

This corollary implies that the normalization factor α plays the same role as the α -normalization in the original diffusion map algorithm [6] in the sense that the landmark distribution does not impact the asymptotic operator. Clearly, while the asymptotic operator does not depend on the landmark set, the operator is, in general, different from that of AD, unless $p^{(2)}$ is constant.

Corollary 4.9. Consider $p_{\mathcal{Z}}^{(2)} \propto \frac{1}{p^{(2)^3}}$ and take $\alpha = 1/2$, $K^{(1)} = K^{(2)}$. We have $q_\alpha(x) = \frac{1}{p^{(2)}(x)^2}$. For $f \in C^3(\mathcal{M})$, we have

$$\begin{aligned} T_{\text{lan}, \epsilon, \alpha} f(x) &= f(x) + \frac{\epsilon \mu_{2,0}^{(2)}}{d} \left(\sum_{i=1}^d (1 - \lambda_i) \frac{\nabla_{E_i}^{(2)} p^{(2)}(x) \nabla_{E_i}^{(2)} f(x)}{p^{(2)}(x)} \right) \\ &\quad + \frac{\epsilon \mu_{2,0}^{(2)}}{2d} \left[\Delta^{(2)} f(x) + \sum_{i=1}^d \lambda_i \nabla_{E_i E_i}^{(2)^2} f(x) \right] + \mathcal{O}(\epsilon^{3/2}). \end{aligned}$$

Furthermore, if $R_x = I_d$, then $\lambda_i = 1$ and the result is independent on probability density. If $\iota^{(1)} = \iota^{(2)}$, then we obtain a landmark design result to recover the Laplace–Beltrami operator that is similar to Remark 1 in [32].

Next, we present the variance analysis of LAD. The proof is postponed to the supplement. We show that as $n \rightarrow \infty$, the matrix $I_n - (\mathbf{D}_\alpha^{(1)})^{-1} \mathbf{W}_\alpha^{(1)} \mathbf{M}_\alpha^{(2)}$ behaves asymptotically like the integral operator $1 - T_{\text{lan}, \epsilon, \alpha}$. The analysis primarily focuses on the stochastic fluctuation of the convergence from finite sampling points. Since the quantity of interest is of the order ϵ , the stochastic fluctuation induced by the finite sampling points should be much smaller than ϵ so that we can obtain the quantify of interest.

Theorem 4.10. Take $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{Z} = \{z_k\}_{k=1}^m$, where $m = \lceil n^\beta \rceil$ for some $0 < \beta \leq 1$ and $\lceil x \rceil$ is the smallest integer greater than or equal to $x \in \mathbb{R}$. Take $f \in C^3(\mathcal{M})$ and denote $\mathbf{f} \in \mathbb{R}^n$ such that $\mathbf{f}_i = f(x_i)$. Let $\epsilon = \epsilon(n)$ so that $\frac{\sqrt{\log n}}{n^{\beta/2} \epsilon^{d/4+1}} \rightarrow 0$ and $\epsilon \rightarrow 0$ when $n \rightarrow \infty$. Then with probability higher than $1 - \mathcal{O}(1/n^2)$, we have

$$\frac{1}{\epsilon} \left[\left(I_n - (\mathbf{D}_\alpha^{(1)})^{-1} \mathbf{W}_\alpha^{(1)} \mathbf{M}_\alpha^{(2)} \right) \mathbf{f} \right] (i) = \frac{f(x_i) - T_{\text{lan}, \epsilon, \alpha} f(x_i)}{\epsilon} + \mathcal{O} \left(\frac{\sqrt{\log n}}{n^{\beta/2} \epsilon^{d/4+1}} \right),$$

for all $i = 1, 2, \dots, n$.

We shall mention that the convergence rate is not optimal. In short, according to a numerical exploration reported in the end of the supplement, the convergence rate should depend on $\epsilon^{d/4+1/2}$ but not $\epsilon^{d/4+1}$. This discrepancy comes from the fact that we handle the convergence of landmark and dataset separately since handling the dependence caused by the diffusion among dataset and landmark is challenging. We leave this problem to our future work.

5. Simulation results. In this section, we use simulated data to validate the theoretical behavior of LAD, and compare it with the original AD. All experiments, except those in section 5.1, were conducted on an 8-core Apple M1 CPU laptop using MATLAB 9.10 (R2021a, 64-bit). Section 5.1 experiments were run on a 10-core Intel Xeon E5-2650 v3 CPU with MATLAB 9.7 (R2019b, 64-bit).

To compare the results of α -LAD and AD, set $\epsilon_1 = \epsilon_2 = \epsilon$, where $\sqrt{\epsilon}$ is the kernel bandwidth parameter, for both algorithms. We choose ϵ such that the $\sqrt{\epsilon}$ -neighborhood of each data point contains a sufficient number of points (approximately over 30 points). Besides controlling bandwidth $\sqrt{\epsilon}$, we also normalize the eigenvectors of both LAD and AD by $\sqrt{n}$.

In addition to comparing the computational time, we also evaluate how accurately α -LAD approximates AD. We consider three quantities. Denote (λ_l, u_l) as the l th eigenpair of α -LAD, and (μ_l, v_l) as the l th eigenpair of AD. The l th eigenvalue difference is quantified by $|\lambda_l - \mu_l|/\mu_l$, while the eigenvector difference is quantified by the inner product of $\langle u_l, v_l \rangle / \|u_l\| \|v_l\|$. Third, we quantify the pairwise embedding difference [2, 14]. Assume we have n points and two embeddings E_1 and E_2 , where the embedding dimension is q . Let $E_\ell(i) \in \mathbb{R}^q$ denote the embedded i th data point recorded from the ℓ th sensor, where $\ell = 1, 2$ and $i = 1, \dots, n$. However, it is well known that eigenvector embedding has a rotational degree of freedom [39]. Thus, the third quantity is

$$(5.1) \quad d(E_1, E_2) := \min_{O \in O(q)} \left(\frac{1}{n} \sum_{i=1}^n \|E_1(i) - OE_2(i)\|_2 \right);$$

that is, the embedded points are rotationally aligned before evaluating the mean pairwise distance. Note that in (5.1), we focus on the pairwise distance between the i th point of E_1 and the i th point of E_2 instead of the global similarity of two embedding manifolds E_1 and E_2 . In this section, unless otherwise mentioned, the pairwise embedding difference is evaluated with $q = 6$.

5.1. The impact of landmark size on α -LAD. We study the computational time and how well LAD approximates AD with various landmark sizes. Consider a torus $\mathbb{T}^2$ and a deformed torus S , defined as

$$\mathbb{T}^2 = \left\{ (x, y, z) \left| \begin{array}{l} x(u, v) = (R + r \cos v) \cos u \\ y(u, v) = (R + r \cos v) \sin u \\ z(u, v) = r \sin v \end{array} \right. \right\},$$

and

$$S = \left\{ (x, y, z) \left| \begin{array}{l} x(u, v) = [R + (1 + 0.5 \cos 4u)r \cos v] \cos u \\ y(u, v) = [R + (1 + 0.5 \cos 4u)r \cos v] \sin u \\ z(u, v) = (1 + 0.5 \cos 4u)r \sin v \end{array} \right. \right\},$$

where $R = 2.5$, $r = 1$, and $u, v \in [0, 2\pi)$. The factor $1 + 0.5 \cos 4u$ deformed the tube radius of the torus. We treat S and $\mathbb{T}^2$ as $\iota^{(1)}(\mathcal{M})$ and $\iota^{(2)}(\mathcal{M})$, respectively. There exist a diffeomorphism map, $\Phi_2 : \mathbb{T}^2 \rightarrow S$. We uniformly sample 20,000 points on $\mathbb{T}^2$ and map them to S using Φ_2 , and collect these points as $\{(r_i, s_i)\}$. We consider different landmark sizes as $m = \lfloor 2^{i/2} \rfloor$ for $i = 11, 12, 13, \dots, 24$, where $\lfloor \cdot \rfloor$ is the floor function. The above process, which includes generating sample points and landmark points, is repeated independently for 30 times. We focus on $\alpha = 1/2$ since 1/2-LAD gives us the best approximation of AD according to the established theorem. The results are shown in Figure 1. Clearly, when the landmark size increases, the computational time increases, but the top eigenvalues and eigenvectors of 1/2-LAD better approximate those of AD.

Furthermore, if we increase n to 500,000 (1,000,000 respectively), a typical laptop would no longer be able to compute AD. However, when using LAD with $m = 1,000$, the computation time is 4.5 (12.3 respectively) minutes. The main limitation going beyond one million points is out of the memory of the laptop.

5.2. The dependence on the initial diffusion. Consider the canonical $\mathbb{S}^1$ and Trefoil knot C , defined as

$$\begin{aligned} \mathbb{S}^1 &= \{(\cos \theta, \sin \theta)\} \subset \mathbb{R}^2, \\ C &= \{(\sin \theta + 2 \sin 2\theta, \cos \theta - 2 \cos 2\theta, -\sin 3\theta)\} \subset \mathbb{R}^3, \end{aligned}$$

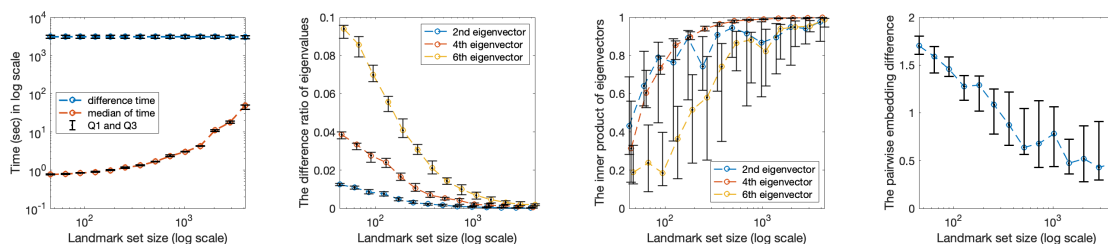


Figure 1. From left to right: a comparison of computational time between AD and 1/2-LAD with different landmark sizes, the difference ratio of the 1st, 3rd, 5th nontrivial eigenvalues between AD and 1/2-LAD, the inner product between the 1st, 3rd, 5th nontrivial eigenvectors of AD and 1/2-LAD and the pairwise embedding difference between AD and 1/2-LAD.

where $\theta \in [-\pi, \pi)$. We treat C and S^2 as $\iota^{(1)}(\mathcal{M})$ and $\iota^{(2)}(\mathcal{M})$, respectively. Note that S^1 and C are diffeomorphic via the map $\Phi_0 : S^1 \rightarrow C$:

$$\Phi_0(x, y) = \begin{cases} x = \sin\left(\arctan \frac{y}{x}\right) + 2 \sin 2\left(\arctan \frac{y}{x}\right), \\ y = \cos\left(\arctan \frac{y}{x}\right) - 2 \cos 2\left(\arctan \frac{y}{x}\right), \\ z = -\sin 3\left(\arctan \frac{y}{x}\right). \end{cases}$$

Uniformly sampling 1000 points from S^1 and mapping them to C using Φ_0 , we collect these points as the dataset $\tilde{\mathcal{X}}$. We then uniformly choose 100 landmark points from $\tilde{\mathcal{X}}$ and set $\alpha = 1/2$. In Figure 2, we observe the impact of the initial sensor chosen for diffusion in both AD and 1/2-LAD. Notably, the results of both 1/2-LAD and AD are contingent upon the selected initial sensor. Additionally, the resulting embedding closely resembles the diffusion maps of the ending sensor, indicating that the embedding is predominantly influenced by diffusion in the ending sensor, which encodes the geometric information for the embedding. While C is diffeomorphic to S^1 , it possesses a distinctive exterior topology different from S^1 . However, since AD and LAD are both diffusion-based algorithms capturing local geometric information, they ultimately cannot capture this exterior topology.

5.3. The impact of landmark distribution and α -normalization. To study the impact of landmark distribution, consider the canonical S^1 and ellipse E , defined as

$$S^1 = \{(\cos \theta, \sin \theta)\} \subset \mathbb{R}^2, \quad E = \{(2 \cos \theta, \sin \theta)\} \subset \mathbb{R}^2,$$

where $\theta \in [-\pi, \pi)$. We treat E and S^1 as $\iota^{(1)}(\mathcal{M})$ and $\iota^{(2)}(\mathcal{M})$, respectively, and define $\Phi_1 : S^1 \rightarrow E$ by

$$\Phi_1(x, y) = \left(2 \cos\left(\arctan \frac{y}{x}\right), \sin\left(\arctan \frac{y}{x}\right)\right).$$

First, the dataset is generated by uniformly sampling n points on S^1 and mapping them to E by Φ_1 , i.e., $p^{(2)}$ is constant. We consider the following two scenarios. In the first scenario, $p_{\mathcal{Z}}^{(2)}$ is uniform, and in the second scenario, we set the nonuniform $p_{\mathcal{Z}}^{(2)}(\theta) = \frac{58}{50}[0.48 \cos \theta + 0.52]$.

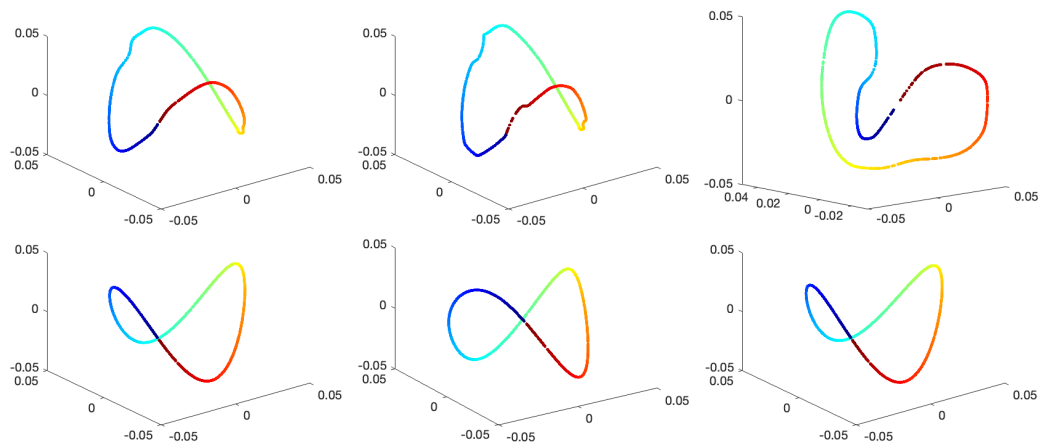


Figure 2. Upper left: AD starting from $\mathbb{S}^1$. Upper middle: 1/2-LAD starting from $\mathbb{S}^1$. Upper Right: Diffusion maps (DM) of dataset on C . Lower left: AD starting from Trefoil knot C . Lower middle: 1/2-LAD starting from Trefoil knot C . Lower right: DM of dataset on $\mathbb{S}^1$.

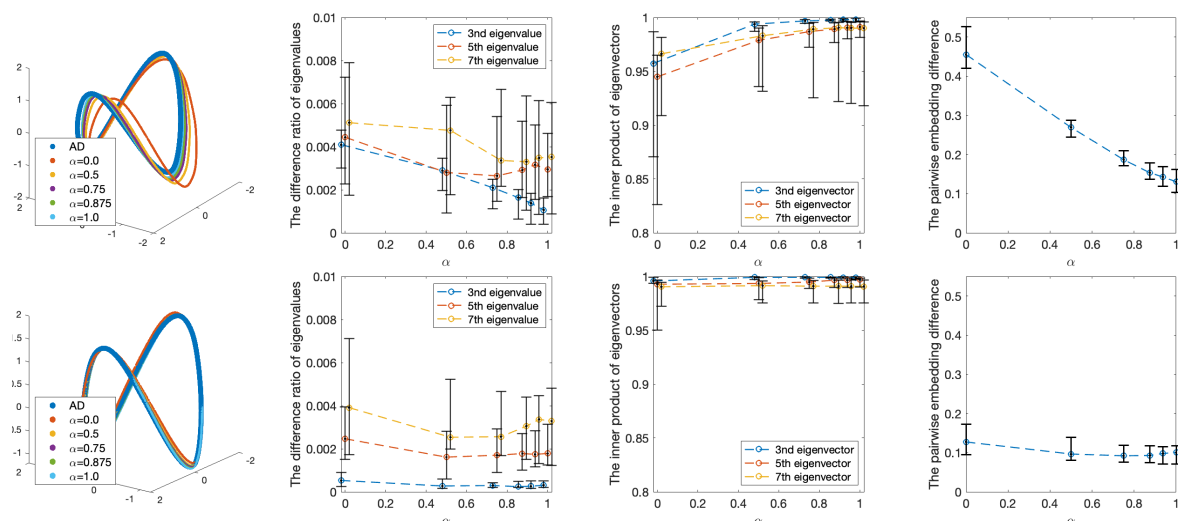


Figure 3. Top row, from left to right: α -LAD with various α , the difference ratio of the 2nd, 4th, 6th nontrivial eigenvalues between AD and α -LAD, the inner product between the 2nd, 4th, 6th nontrivial eigenvectors of AD and α -LAD, and the pairwise embedding difference between AD and α -LAD, where the landmark distribution is nonuniform. Bottom row shows the same thing but when the landmark distribution is uniform.

In both cases, we randomly sample 2500 points $\{(s_i, r_i) : s_i \in E, r_i \in \mathbb{S}^1\}$ and choose 1000 landmark points according to $p_{\mathcal{Z}}^{(2)}$.

α -LAD with various α under different $p_{\mathcal{Z}}^{(2)}$ setups are shown in the left column of Figure 3. We observe that when landmarks are uniformly selected, the choice of α seems to have less impact on the embeddings. However, in the case of nonuniform landmark selection, the

embeddings vary and depend on the value of α . When $\alpha = 1$, the embedding (colored in cyan) becomes similar to LAD with uniform sampling. However, we will see below that when the landmark sampling is nonuniform, LAD fails to recover AD when $\alpha = 1$. The quantification of the deviation of LAD from AD is shown in Figure 3. As α approaches 1, the impact of nonuniform landmark distribution on LAD is mitigated, corroborating the theoretical analysis.

Next, we know from the theoretical analysis shown in Corollary 4.8 that when $\alpha = 1$, although the impact of nonuniform landmark sample on LAD is eliminated, it do not recover AD. To further explore this result, consider the same $\mathbb{S}^1$ setup. But this time, the dataset is sampled nonuniformly following the distribution $p^{(2)}$ as shown in the top left subfigure of Figure 4. Additionally, we arbitrarily select four different landmark distributions. See the top row of Figure 4 for an illustration. We sample 3500 points $\{(s_i, r_i) : s_i \in E, r_i \in \mathbb{S}^1\}$ following $p^{(2)}$ and choose 1000 landmark points following different landmark distributions. We label these four different $p_Z^{(2)}$ cases as “Case 1” to “Case 4.”

The embedding manifolds under these different landmark distributions as α approaches 1 are shown in the middle row of Figure 4. We see that as α approaches 1, the embeddings by LAD are more similar, but these embeddings are not necessarily similar to the embedding

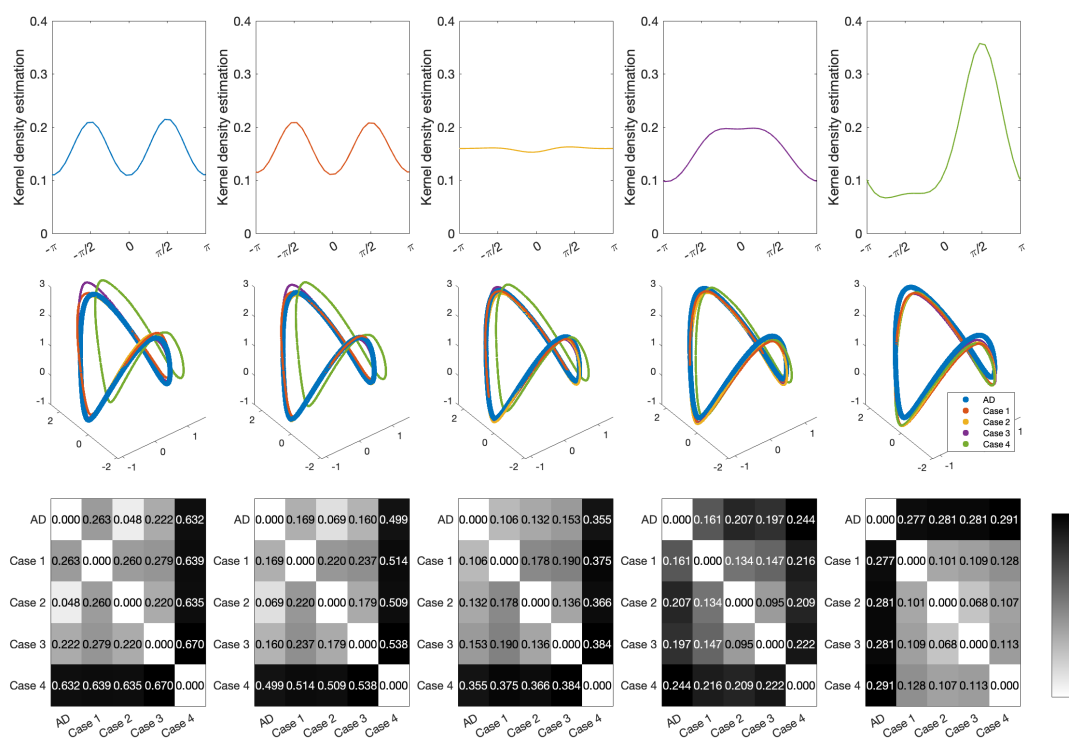


Figure 4. Illustration of various embeddings by α -LAD under different landmark distributions. Top row, from left to right: the sampling density function $p^{(2)}$ (blue) and four landmark distributions $p_Z^{(2)}$, labeled “Case 1” to “Case 4.” Middle row, from left to right: the outputs of α -LAD with $\alpha = 0, 1/4, 1/2, 3/4$, and 1, respectively. The blue curve is the result of AD, which is the same from left to right. The curves in other colors correspond to different landmark distributions. Lower row, from left to right: the pairwise embedding difference between different embeddings with $\alpha = 0, 1/4, 1/2, 3/4$, and 1, respectively.

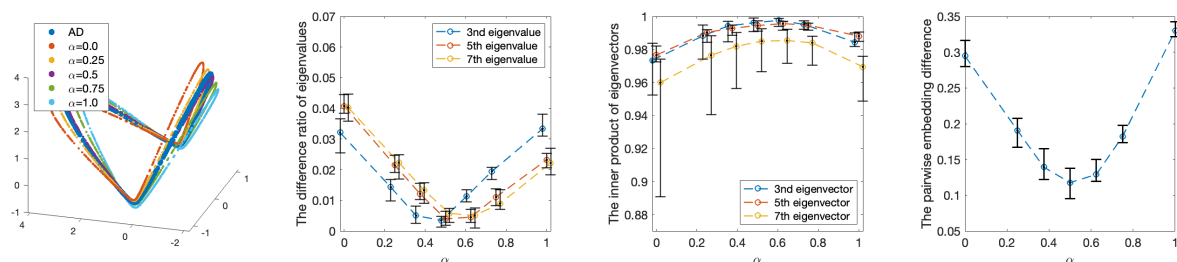


Figure 5. Left: The α -LAD with $p_{\mathcal{Z}}^{(2)} = p^{(2)}$ and $\alpha = 0, 1/4, 1/2, 3/4, 1$. Middle left: The difference ratio of the 2nd, 4th, 6th nontrivial eigenvalues of AD and α -LAD with $\alpha = 0, 1/4, 3/8, 1/2, 5/8, 3/4, 1$. Middle right: The inner product of the 2nd, 4th, 6th nontrivial eigenvectors of AD and α -LAD with $\alpha = 0, 1/4, 3/8, 1/2, 5/8, 3/4, 1$. Right: The embedding difference between AD and α -LAD.

by AD. The quantification of pairwise embedding differences is shown at the bottom row of Figure 4. While a numerical difference between AD and LAD results is evident in the bottom row of Figure 4, the middle row shows no significant visual differences in the embedding manifolds. The supplementary material (supplement.pdf [local/web 988KB]) includes an example with a more noticeable visual difference.

Finally, we ask whether we can design landmark distribution and choose α appropriately so that α -LAD can recover the original AD. Consider the same $\mathbb{S}^1$ model with $\Phi_1 : \mathbb{S}^1 \rightarrow E$, but with the data sampling distribution $p^{(2)}$ nonuniform. According to the established theorem, this can be achieved if $p_{\mathcal{Z}}^{(2)} = p^{(2)}$ and $\alpha = 1/2$. The sample point follows the probability density function $p^{(2)}(\theta) = \frac{1}{\pi} [\tan^{-1}(\frac{1}{2} \tan \theta)]$. We randomly select 2500 points $\{(s_i, r_i) : s_i \in E, r_i \in \mathbb{S}^1\}$ and uniformly choose 1000 landmark points from them, i.e., $p_{\mathcal{Z}}^{(2)} = p^{(2)}$. The quantification results with various α are illustrated in Figure 5. We observe that 1/2-LAD approximates the embedding by AD. This result is consistent with our theoretical expectations.

5.4. Comparison with Nyström extension. Last but not least, consider a related algorithm, the Nyström extension [37, 11], which applies eigen-decomposition to a small dataset subset and extrapolates the eigenvectors to the full set. Below we consider a comparison with combining the modified Nyström extension [10, 19, 32] with AD.

Since our data lies on two different manifolds, we regard the term $\mathbf{W}^{(1)}(\mathbf{D}^{(2)})^{-1}\mathbf{W}^{(2)}$ in (2.1), $(\mathbf{D}^{(1)})^{-1}\mathbf{W}^{(1)}(\mathbf{D}^{(2)})^{-1}\mathbf{W}^{(2)}$, as an effective affinity matrix. Please refer to Definition 3.2 in [36] or (SM1.2) in the supplementary material (supplement.pdf [local/web 988KB]) for more details. Then, we define the *effective kernel* as $\tilde{K}_\epsilon^{(e)}(x_i, x_j) = [\mathbf{W}^{(1)}(\mathbf{D}^{(2)})^{-1}\mathbf{W}^{(2)}]_{ij}$. Then, extract the subrows and the subcolumns of the effective affinity matrix $\mathbf{W}^{(1)}(\mathbf{D}^{(2)})^{-1}\mathbf{W}^{(2)}$ corresponding to the landmark points, resulting in a effective affinity matrix on landmark set with size $m \times m$. Similarly, we obtain the subrows of the effective affinity matrix corresponding to additional points, which is an extrapolation matrix with size $(n - m) \times m$. Then, apply modified Nyström extension. We call the AD accelerated by the modified Nyström extension Nyström-AD (NAD). Consider the same manifold setup as in section 5.1 and adopt the metric used in Figure 1 to compare performance. For clarity, we only compare the first and third nontrivial eigenvectors with different landmark sizes as $m = \lfloor 2^{i/2} \rfloor$ for $i = 11, 12, 13, \dots, 20$ and focus on $\alpha = 1/2$. The results are shown in Figure 6. We can observe that the results of LAD

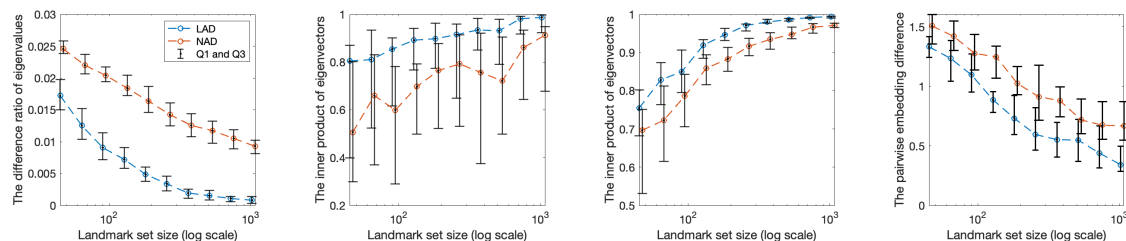


Figure 6. From left to right: a comparison of difference ratio of the 1st nontrivial eigenvalues between AD and 1/2-LAD as well as between AD and NAD with different landmark sizes, the inner product between the 1st, 3rd nontrivial eigenvectors of AD and 1/2-LAD as well as between AD and NAD, and the pairwise embedding difference between AD and 1/2-LAD as well as between AD and NAD.

are closer to those of AD compared to NAD. Note that the theoretical analysis of Nyström extension is limited [4], and it is unclear how the geometric information is lost. Since it is outside the scope of this work, we leave this topic to our future work.

6. Application to automatic sleep stage annotation. According to the American Academy of Sleep Medicine (AASM), sleep dynamics can be classified into two broad stages: rapid eye movement (REM) and nonrapid eye movement (NREM). The NREM stage is further divided into N1, N2, and N3. To assess the sleep stage, experts read electroencephalogram (EEG), electrooculogram, and electromyogram following the Rechtschaffen and Kales (R&K) criteria [3, 30].

AD has been applied to establish an automatic sleep stage annotation system from two EEG channels [26]. The key idea is that the common variable among various channels reveals a more reliable physiological process associated with sleep stages by eliminating sensor-specific disturbances. However, the computational burden limits the application of AD to short signals. We now demonstrate that LAD is a computationally efficient alternative to AD for extracting common sources of brain dynamics recorded by the O1-A2 and O2-A1 electrodes for an automatic sleep stage annotation system for five sleep stages, including awake, REM, N1, N2, and N3. We consider polysomnogram data from forty subjects without sleep apnea from the Taiwan Integrated Database for Intelligent Sleep (TIDIS). All subjects' recordings exceed six hours. All experiments were conducted on a computer equipped with two 12-core CPUs, Intel Xeon CPU E5-2697 v2, utilizing MATLAB software version 9.12 (R2022a) in its 64-bit configuration for experimental design and analysis.

6.1. Feature extraction. By the R&K criteria [3, 30], we take a 30-second epoch for sleep stage evaluation. Each EEG channel recording undergoes the following preprocessing. Suppose the i th epoch starts at time t_i and ends at $t_i + 30$. We extend the i th epoch to a 90-second EEG segment starting at $t_i - 60$ and ending at $t_i + 30$. Since the sampling rate is 200 Hz, each data vector is of size 18,000. Following [26], we apply the scattering transform [1] to extract spectral features from each 90-second EEG segment. For subject x , denote $u_{x,j}^{(\ell)}$ as the extracted feature of the j th epoch in the ℓ th channel, where $\ell = 1, 2$ represents O1-A2 and O2-A1, respectively, and $x = 1, \dots, 40$ represents 40 subjects. Let $\mathcal{U}^{(\ell)} = \bigcup_{x=1}^{40} \{u_{x,j}^{(\ell)}\}_{j=1}^{J_x}$

Table 1

Upper: The cumulative sum of all confusion matrices of AD. Middle: The cumulative sum of all confusion matrices of 0.5-LAD without class-balanced landmark set. Lower: The cumulative sum of all confusion matrices of 0.5-LAD* with class-balanced landmark set.

AD		Predict				
		Awake	REM	N1	N2	N3
True	Awake (16 %)	3477 (82%)	143 (3%)	370 (9%)	236 (6%)	4 (0%)
	REM (13 %)	63 (2%)	2517 (72%)	514 (15%)	414 (12%)	1 (0%)
	N1 (9 %)	432 (18%)	581 (24%)	868 (36%)	548 (23%)	0 (0%)
	N2 (52 %)	287 (2%)	478 (3%)	464 (3%)	12441 (88%)	523 (4%)
	N3 (10 %)	121 (4%)	2 (0%)	2 (0%)	844 (31%)	1760 (64%)
0.5-LAD		Predict				
		Awake	REM	N1	N2	N3
True	Awake (16 %)	3472 (82%)	173 (4%)	356 (8%)	226 (5%)	3 (0%)
	REM (13 %)	90 (3%)	2475 (71%)	461 (13%)	482 (14%)	1 (0%)
	N1 (9 %)	470 (19%)	620 (26%)	742 (31%)	597 (25%)	0 (0%)
	N2 (52 %)	276 (2%)	485 (3%)	447 (3%)	12422 (88%)	563 (4%)
	N3 (10 %)	30 (1%)	0 (0%)	2 (0%)	978 (36%)	1719 (63%)
0.5-LAD*		Predict				
		Awake	REM	N1	N2	N3
True	Awake (16 %)	3555 (84%)	163 (4%)	291 (7%)	216 (5%)	5 (0%)
	REM (13 %)	94 (3%)	2499 (71%)	439 (13%)	476 (14%)	1 (0%)
	N1 (9 %)	508 (21%)	592 (24%)	774 (32%)	554 (23%)	1 (0%)
	N2 (52 %)	280 (2%)	506 (3%)	470 (3%)	12431 (88%)	506 (4%)
	N3 (10 %)	24 (1%)	3 (0%)	14 (1%)	855 (31%)	1833 (67%)

be the scattering EEG spectral features of the ℓ th channel, where J_x is the number of epochs of subject x . In this dataset, the size of $\mathcal{U}^{(\ell)}$ is 27,090 for $\ell = 1, 2$. Next, we apply AD or LAD to fuse information from the two channels. Let $m = 850 \approx 5\sqrt{n}$ is the size of the landmark set. Furthermore, following the established theorem and the above simulation results, we set $\alpha = 0.5$ and uniformly sample landmark set from $\{(s_i, r_i)\}$, i.e., $p^{(2)} = p_{\mathcal{Z}}^{(2)}$ and the landmark set is a subset of the original database.

It is well known that N2 stages dominate other sleep stages, and usually, N1 and REM are much less frequent (see the leftmost column in Table 1, which show 52% epochs are labeled N2 and only 9% are labeled N1). While handling the imbalanced data is not the focus of this study, for the sake of completeness of the analysis, we consider a simple balancing scheme—we uniformly sample 170 epochs from each stage, resulting in a total of 850 epochs, as the landmark set. We conducted separate experiments with and without the class balancing. In the subsequent experiments, 0.5-LAD without a class balancing is denoted by 0.5-LAD*.

6.2. Classification. We chose the standard and widely used kernel support vector machine (SVM) [5] for the learning step. Kernel SVM finds a nonlinear hyperplane to separate the data set into two disjoint subsets. Specifically, we opted for the radial basis function (RBF) kernel. Since our sleep dynamics classification problem involves five classes, we needed to extend the kernel SVM to a multiclass SVM. To achieve this, we applied the one-versus-all (OVA) classification scheme. We balance the dataset using random down-sampling, which involves randomly selecting epochs from the majority class (N2) and deleting them from the training dataset before kernel SVM training. In our experiment, we randomly delete 50% of N2 epochs before training.

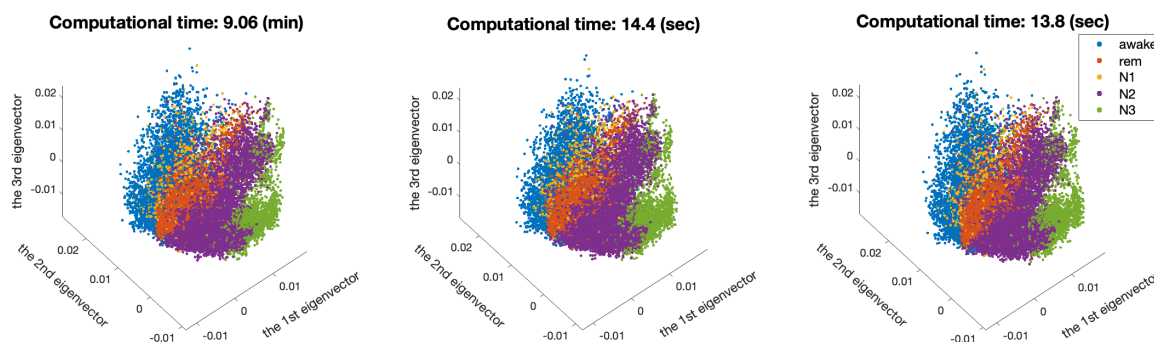


Figure 7. Left: embedding by AD. Middle: embedding by 0.5-LAD. Right: embedding by 0.5-LAD*.

6.3. Results. After applying 0.5-LAD, 0.5-LAD*, and AD to 29,070 epochs, the top three nontrivial eigenvectors are shown in Figure 7. It can be observed that the results of 0.5-LAD and 0.5-LAD* are *nearly* identical to those of AD. The inner product of the top three nontrivial eigenvectors between 0.5-LAD and 0.5-LAD* are 0.9940, 0.9937, and 0.9936, respectively, and the pairwise embedding difference is 0.006.

We conducted leave-one-subject-out cross-validation (LOSOCV) to assess the SVM prediction performance. The accuracy of AD, 0.5-LAD and 0.5-LAD* are 0.793 ± 0.052 , 0.780 ± 0.035 , and 0.785 ± 0.039 over LOSOCV, respectively. The macro F1 of AD, 0.5-LAD, and 0.5-LAD* are 0.655 ± 0.068 , 0.650 ± 0.050 , and 0.644 ± 0.058 over LOSOCV, respectively. By considering a p -value less than 0.05 as indicating statistical significance, both methods showed no significant difference in accuracy or macro F1 compared to AD using the Wilcoxon sign-rank test with Bonferroni correction (for accuracy, the p -values are 0.459 and 0.863 when comparing 0.5-LAD and 0.5-LAD* with AD respectively, and for macro F1, the p -values are 0.338 and 0.718 when comparing 0.5-LAD and 0.5-LAD* with AD, respectively). Notably, the computational times for 0.5-LAD and 0.5-LAD* are 14.4 and 13.8 seconds, respectively, which is shorter by over an order of magnitude than AD's average computation time, which is nine minutes. In summary, despite slightly lower performance without statistical significance, adopting 0.5-LAD substantially reduces the computational time, enhancing practical utility.

The distributions of accuracy and macro-F1 by LOSOCV obtained from different methods are depicted on the left-hand side and right-hand side of Figure 8, respectively. Through LOSOCV, we acquired 40 confusion matrices for each validation subject. It is evident that the outcomes of 0.5-LAD closely resemble those of AD, while significantly reducing computational time. The cumulative sum of all confusion matrices derived from AD and 0.5-LAD are depicted in Table 1. Notably, the recall of 0.5-LAD* outperforms that of 0.5-LAD in the classes with fewer data, namely REM and N1, which emphasizes the importance of balancing the data.

7. Conclusion. We introduce a novel algorithm called Landmark Alternating Diffusion (LAD) designed to enhance the performance of the commonly employed kernel sensor fusion method, Alternating Diffusion (AD). Furthermore, we integrate an α -normalization parameter to mitigate the influence of landmark sampling schemes. Apart from significantly improved

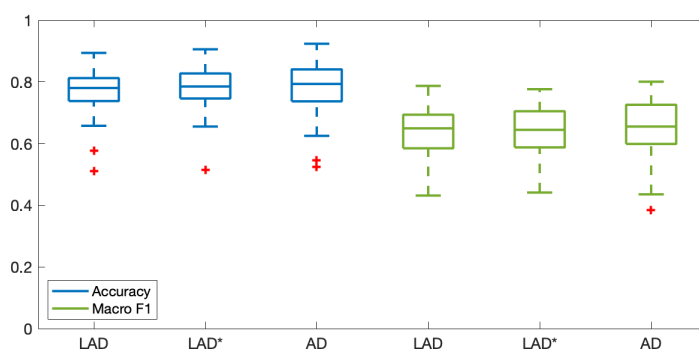


Figure 8. The box plot of 40-fold validation accuracies by different methods. LAD stands for Landmark AD without class-balanced landmark set fusion 2 channels; LAD* stands for Landmark AD with class-balanced landmark set fusion 2 channels; AD stands for AD fusion 2 channels.

computational efficiency, LAD exhibits asymptotic behavior akin to that of AD within a manifold framework.

Acknowledgment. We would like to thank Professor Chun-Hao Yang for valuable comments. We also thank the anonymous reviewers for their constructive feedback.

REFERENCES

- [1] J. ANDÉN AND S. MALLAT, *Deep scattering spectrum*, IEEE Trans. Signal Process., 62 (2014), pp. 4114–4128, <https://doi.org/10.1109/TSP.2014.2326991>.
- [2] K. S. ARUN, T. S. HUANG, AND S. D. BLOSTEIN, *Least-squares fitting of two 3-d point sets*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-9 (1987), pp. 698–700, <https://doi.org/10.1109/TPAMI.1987.4767965>.
- [3] R. B. BERRY, R. BROOKS, C. E. GAMALDO, S. M. HARDING, R. M. LLOYD, C. L. MARCUS, AND B. V. VAUGHN, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications : Version 2.3*, American Academy of Sleep Medicine, 2015.
- [4] L.-B. CHANG, Z. BAI, S.-Y. HUANG, AND C.-R. HWANG, *Asymptotic error bounds for kernel-based Nyström low-rank approximation matrices*, J. Multivariate Anal., 120 (2013), pp. 102–119, <https://doi.org/10.1016/j.jmva.2013.05.006>.
- [5] A. CHRISTMANN AND I. STEINWART, *Support Vector Machines*, Springer Science & Business Media, 2008, <https://doi.org/10.1007/978-0-387-77242-4>.
- [6] R. R. COIFMAN AND S. LAFON, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30, <https://doi.org/10.1016/j.acha.2006.04.006>.
- [7] X. DING AND H.-T. WU, *On the spectral property of kernel-based sensor fusion algorithms of high dimensional data*, IEEE Trans. Inform. Theory, 67 (2021), pp. 640–670, <https://doi.org/10.1109/TIT.2020.3026255>.
- [8] X. DING AND H.-T. WU, *How do kernel-based sensor fusion algorithms behave under high-dimensional noise?*, Inf. Inference, 13 (2024), iaad051, <https://doi.org/10.1093/imaiai/iaad051>.
- [9] D. DOV, R. TALMON, AND I. COHEN, *Kernel-based sensor fusion with application to audio-visual voice activity detection*, IEEE Trans. Signal Process., 64 (2016), pp. 6406–6416, <https://doi.org/10.1109/TSP.2016.2605068>.
- [10] D. DOV, R. TALMON, AND I. COHEN, *Sequential audio-visual correspondence with alternating diffusion kernels*, IEEE Trans. Signal Process., 66 (2018), pp. 3100–3111, <https://doi.org/10.1109/TSP.2018.2802445>.
- [11] C. FOWLKES, S. BELONGIE, F. CHUNG, AND J. MALIK, *Spectral grouping using the Nyström method*, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004), pp. 214–225, <https://doi.org/10.1109/TPAMI.2004.1262185>.

- [12] F. GUSTAFSSON, *Statistical Sensor Fusion*, Professional Publishing House, 2012.
- [13] D. R. HARDOON, S. SZEDMAK, AND J. SHAW-TAYLOR, *Canonical correlation analysis: An overview with application to learning methods*, Neural Comput., 16 (2004), pp. 2639–2664, <https://doi.org/10.1162/0899766042321814>.
- [14] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1160–1174, <https://doi.org/10.1137/0907079>.
- [15] P. HORST, *Relations among m sets of measures*, Psychometrika, 26 (1961), pp. 129–149, <https://doi.org/10.1007/BF02289710>.
- [16] H. HOTELLING, *Relations between two sets of variates*, Biometrika, 28 (1936), pp. 321–377, <https://doi.org/10.1093/biomet/28.3-4.321>.
- [17] H. HWANG, K. JUNG, Y. TAKANE, AND T. S. WOODWARD, *A unified approach to multiple-set canonical correlation analysis and principal components analysis*, Br. J. Math. Stat. Psychol., 66 (2013), pp. 308–321, <https://doi.org/10.1111/j.2044-8317.2012.02052.x>.
- [18] O. KATZ, R. TALMON, Y.-L. LO, AND H.-T. WU, *Alternating diffusion maps for multimodal data fusion*, Inform. Fusion, 45 (2019), pp. 346–360, <https://doi.org/10.1016/j.inffus.2018.01.007>.
- [19] S. LAFON, Y. KELLER, AND R. R. COIFMAN, *Data fusion and multicue data matching by diffusion maps*, IEEE Trans. Pattern Anal. Mach. Intell., 28 (2006), pp. 1784–1797, <https://doi.org/10.1109/TPAMI.2006.223>.
- [20] D. LAHAT, T. ADALI, AND C. JUTTEN, *Multimodal data fusion: An overview of methods, challenges, and prospects*, Proc. IEEE, 103 (2015), pp. 1449–1477, <https://doi.org/10.1109/JPROC.2015.2460697>.
- [21] B. LAUFER-GOLDSHTEIN, R. TALMON, AND S. GANNOT, *Data-driven multi-microphone speaker localization on manifolds*, Found. Trends Signal Process., 14 (2020), pp. 1–161, <https://doi.org/10.1561/20000000098>.
- [22] R. R. LEDERMAN AND R. TALMON, *Learning the geometry of common latent variables using alternating-diffusion*, Appl. Comput. Harmon. Anal., 44 (2018), pp. 509–536, <https://doi.org/10.1016/j.acha.2015.09.002>.
- [23] R. R. LEDERMAN, R. TALMON, H.-T. WU, Y.-L. LO, AND R. R. COIFMAN, *Alternating diffusion for common manifold learning with application to sleep stage assessment*, in Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, South Brisbane, Queensland, Australia, 2015, pp. 5758–5762, <https://doi.org/10.1109/ICASSP.2015.7179075>.
- [24] O. LINDENBAUM, Y. BREGMAN, N. RABIN, AND A. AVERBUCH, *Multiview kernels for low-dimensional modeling of seismic events*, IEEE Trans. Geosci. Remote Sens., 56 (2018), pp. 3300–3310, <https://doi.org/10.1109/TGRS.2018.2797537>.
- [25] O. LINDENBAUM, A. YEREDOR, M. SALHOV, AND A. AVERBUCH, *Multi-view diffusion maps*, Inform. Fusion, 55 (2020), pp. 127–149, <https://doi.org/10.1016/j.inffus.2019.08.005>.
- [26] G.-R. LIU, Y.-L. LO, J. MALIK, Y.-C. SHEU, AND H.-T. WU, *Diffuse to fuse EEG spectra—Intrinsic geometry of sleep dynamics for classification*, Biomed. Signal Process. Control, 55 (2020), 101576, <https://doi.org/10.1016/j.bspc.2019.101576>.
- [27] G.-R. LIU, Y.-L. LO, Y.-C. SHEU, AND H.-T. WU, *Explore intrinsic geometry of sleep dynamics and predict sleep stage by unsupervised learning techniques*, in Harmonic Analysis and Applications, Springer, 2020, pp. 279–324, https://doi.org/10.1007/978-3-030-61887-2_11.
- [28] N. F. MARSHALL AND M. J. HIRN, *Time coupled diffusion maps*, Appl. Comput. Harmon. Anal., 45 (2018), pp. 709–728, <https://doi.org/10.1016/j.acha.2017.11.003>.
- [29] T. MICHAELI, W. WANG, AND K. LIVESCU, *Nonparametric canonical correlation analysis*, in Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML’16, 2016, 48, pp. 1967–1976.
- [30] A. RECHTSCHAFFEN AND A. KALES, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*, Neurological Information Network, 1968, <https://books.google.com.tw/books?id=wzdRnQEACAAJ>.
- [31] A. ROMÁN-MESSINA, C. M. CASTRO-ÁRVIZU, A. CASTILLO-TAPIA, E. R. MURILLO-AGUIRRE, AND O. RODRÍGUEZ-VILLALÓN, *Multiview spectral clustering of high-dimensional observational data*, IEEE Access, 11 (2023), pp. 115884–115893, <https://doi.org/10.1109/ACCESS.2023.3323604>.

- [32] C. SHEN, Y.-T. LIN, AND H.-T. WU, *Robust and scalable manifold learning via landmark diffusion for long-term medical signal processing*, J. Mach. Learn. Res., 23 (2022), pp. 1–30, <https://www.jmlr.org/papers/v23/20-786.html>.
- [33] C. SHEN AND H.-T. WU, *Scalability and robustness of spectral embedding: Landmark diffusion is all you need*, Inf. Inference, 11 (2022), pp. 1527–1595, <https://doi.org/10.1093/imaiai/iaac013>.
- [34] T. SHNITZER, M. BEN-CHEN, L. GUIBAS, R. TALMON, AND H.-T. WU, *Recovering hidden components in multimodal data with composite diffusion operators*, SIAM J. Math. Data Sci., 1 (2019), pp. 588–616, <https://doi.org/10.1137/18M1218157>.
- [35] T. SHNITZER, H.-T. WU, AND R. TALMON, *Spatiotemporal analysis using Riemannian composition of diffusion operators*, Appl. Comput. Harmon. Anal., 68 (2024), 101583, <https://doi.org/10.1016/j.acha.2023.101583>.
- [36] R. TALMON AND H.-T. WU, *Latent common manifold learning with alternating diffusion: Analysis and applications*, Appl. Comput. Harmon. Anal., 47 (2019), pp. 848–892, <https://doi.org/10.1016/j.acha.2017.12.006>.
- [37] C. WILLIAMS AND M. SEEGER, *Using the Nyström method to speed up kernel machines*, Adv. Neural Inf. Process. Syst., 13 (2000), pp. 661–667, https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.
- [38] L. XIAO, J. M. STEPHEN, T. W. WILSON, V. D. CALHOUN, AND Y.-P. WANG, *A manifold regularized multi-task learning model for IQ prediction from two fMRI paradigms*, IEEE Trans. Bio-Med. Eng., 67 (2019), pp. 796–806, <https://doi.org/10.1109/TBME.2019.2921207>.
- [39] L. ZELNIK-MANOR AND P. PERONA, *Self-tuning spectral clustering*, Adv. Neural Inf. Process. Syst., 17 (2004), pp. 1601–1608.
- [40] X. ZHUANG, Z. YANG, AND D. CORDES, *A technical review of canonical correlation analysis for neuroscience applications*, Hum. Brain Mapp., 41 (2020), pp. 3807–3833, <https://doi.org/10.1002/hbm.25090>.