Graph based Gaussian processes on restricted domains

David B. Dunson1 | Hau-Tieng Wu2 | Nan Wu3

1Department of Statistical Science, Duke University, Durham, North Carolina, USA
2Departments of Mathematics and Department of Statistical Science, Duke University, Durham, North Carolina, USA
3Department of Mathematics, Duke University, Durham, North Carolina, USA

Correspondence
Nan Wu, Department of Mathematics, 120 Science Drive, Physics Building, Duke University, Durham, NC 27708-0320, USA.
Email: nan.wu@duke.edu

Funding information
European Research Council (ERC), Grant/Award Number: 856506

Abstract
In nonparametric regression, it is common for the inputs to fall in a restricted subset of Euclidean space. Typical kernel-based methods that do not take into account the intrinsic geometry of the domain across which observations are collected may produce sub-optimal results. In this article, we focus on solving this problem in the context of Gaussian process (GP) models, proposing a new class of Graph Laplacian based GPs (GL-GPs), which learn a covariance that respects the geometry of the input domain. As the heat kernel is intractable computationally, we approximate the covariance using finitely-many eigenpairs of the Graph Laplacian (GL). The GL is constructed from a kernel which depends only on the Euclidean coordinates of the inputs. Hence, we can benefit from the full knowledge about the kernel to extend the covariance structure to newly arriving samples by a Nyström type extension. We provide substantial theoretical support for the GL-GP methodology, and illustrate performance gains in various applications.

KEYWORDS
Bayesian, graph Laplacian, heat kernel, manifold, nonparametric regression, restricted domain, semi-supervised

1 | INTRODUCTION

We are interested in problems in which data are collected on ‘inputs’ \( x = (x_1, \ldots, x_D) \in S \subset \mathbb{R}^D \) and ‘outputs’ \( y \in \mathbb{R} \), with \( S \) a subset of \( \mathbb{R}^D \). Labelled training data \( \mathcal{D} = \{x_i, y_i\}_{i=1}^m \) are available for
samples $i = 1, \ldots, m$, along with (possibly) unlabelled data $\mathcal{U} = \{x_i\}_{i=m+1}^{m+n}$. There are many settings in which this problem arises, including nonparametric regression focused on using features $x$ to predict outcome $y$ and computer model emulation in which $x$ corresponds to an input into a computer model that is expensive to run and $y$ corresponds to the output. Gaussian process (GP) models are routinely used in these settings, but without explicitly taking into account the geometry of $S$ or using the unlabelled data to improve estimation of the unknown input-output function. It is common for $S$ to be highly restricted and non-linear. For example, in computer model emulation, the outputs commonly satisfy some physical laws or differential equations constrained over the domain of the inputs with complicated geometry.

For concreteness, we focus on the following model throughout the paper, while noting that many elaborations are straightforward within the framework we will propose,

$$y_i = f(x_i) + \eta_i, \quad \eta_i \sim N(0, \sigma^2_{\text{noise}}),$$

where $f : S \to \mathbb{R}$ is an unknown regression function that is assigned a GP prior centred at zero with covariance function $C(x, x')$ and $\sigma^2_{\text{noise}}$ is the measurement error variance (this can be set to zero for deterministic computer models). The choice of $C(x, x')$ has a fundamental impact on the results; the most common choices of covariance function are the squared exponential and its generalization, the Matérn. Both choices depend critically on the distance between inputs $x$ and $x'$; by far the most common choice in practice is the Euclidean distance.

We use the following example to illustrate the problems with ignoring the geometry of $S$ and simply using a Euclidean distance-based kernel. Raynaud’s disease is a disorder of the blood vessels in the fingers and toes. When a person is cold or feels emotional stress, it causes the blood vessels to narrow so that the blood can not get to the skin. As a result, the affected parts on the fingers and toes turn white and then blue and there is a significant difference between the temperature over the affected part and the unaffected part. In Figure 1, we plot a simulation of the skin temperature (degree Celsius) of a patient with Raynaud’s disease from a 3D scan of a right hand, a surface in $\mathbb{R}^3$. We hold out the temperature values at all but a relatively small number of sensor locations; the top left panel shows the labelled data and the top right all the data. We fit a GP with the square exponential of the Euclidean distance in the covariance, and show the predicted values in the bottom left panel of Figure 1. There is a substantial change in temperature between the index finger and the middle finger and between the ring finger and the little finger. For the GP with squared exponential covariance, there is inappropriate smoothing between different fingers, leading to poor predictions over the regions indicated in the boxes.

There have been attempts to solve related problems in the literature. Cheng and Wu (2013) assume $S$ is an unknown submanifold in a Euclidean space and develop a locally linear regression method on the manifold; see also Zhu et al. (2003), Zhu (2005), Belkin et al. (2006), Nadler et al. (2009), Dunlop et al. (2020), and Wang and Lerman (2015) for semi-supervised approaches. An alternative is to choose a covariance in a GP prior that respects the geometry of $S$, but it is not clear how to specify such a covariance. When the subset $S$ is a submanifold in a Euclidean space, the heat kernel of $S$ that characterizes the diffusion of heat flowing out of a point in $S$ is a potentially appealing choice. Castillo et al. (2014) studies theoretical properties of the posterior distribution, such as rates of contraction, for related models. Unfortunately, the heat kernel is typically intractable to calculate. For known manifolds, Lin et al. (2019) propose an extrinsic GP that embeds the manifold in a higher-dimensional Euclidean space, while Niu et al. (2019) propose an intrinsic GP relying on a Monte Carlo (MC) approximation to the heat kernel. The intrinsic GP is only applicable to known manifolds and their MC approximation is very expensive.
FIGURE 1 Points are sampled from a 3D scan of a right hand. Each point corresponds to the skin temperature (degree Celsius). Top left panel shows a subset of labelled data and the top right panel shows all the data. We fit Gaussian process (GP) models to predict held-out temperature data. The bottom left panel shows the results for a GP with squared exponential covariance, while the bottom right panel shows the results for our Graph Laplacian based GP (GL-GP) [Colour figure can be viewed at wileyonlinelibrary.com]

computationally, relying on simulating Brownian motion many times and calculating proportions of paths from a starting point ending up close to a target point. There are also valid covariance structures defined on a submanifold in a Euclidean space other than the heat kernel. For example, the generalized Mååter covariance is discussed in Li et al. (2021) and Borovitskiy et al. (2020). When the manifold is known, such covariances can be approximated by numerically solving stochastic partial differential equations on the manifold.

In this article, we develop a novel Graph Laplacian-based GP (GL-GP) to solve the above problem with predictors on an unknown subset $S$ of $\mathbb{R}^D$, which is not necessarily a manifold. The key idea is to construct a covariance that incorporates the intrinsic geometry of $S$. This is accomplished through taking finitely many eigenpairs of the graph Laplacian (GL) using the labelled and unlabelled predictor values. The GL is widely studied in spectral graph theory (Chung, 1996), and corresponds to the infinitesimal generator of a random walk on the sampled data points. The covariance in the GL-GP approximates a diffusion process on $S$ with respect to intrinsic distances between data points. Tuning parameters in the covariance functions can be estimated by maximizing marginal likelihoods (Rasmussen, 2003). We propose a Nyström extension method to extend the covariance structure determined from an existing dataset to a newly arriving dataset. To provide a teaser illustrating practical advantages of the GL-GP, we fit our GL-GP to the skin temperature example. The predicted values are shown in the bottom right panel of Figure 1.
The proposed GL-GP can be used broadly in place of existing GP models. The method is designed to adapt to the support of the sample points, and one does not need to know a priori that the data have constrained support. We find in practice that the GL-GP often outperforms GPs with standard off the shelf covariance functions in general applications, particularly when the predictors have a nontrivial geometric or topological structure. In addition to the novel GL-GP methodology, a major contribution of this paper is the theory we have developed in support of the GL-GP. We first associate the GL with an integral operator for any subset \( S \) of the Euclidean space. We discuss theoretical properties of the integral operator and define the covariance function of the GL-GP by using finitely many eigenpairs of the integral operator. When \( S \) is an embedded submanifold of \( \mathbb{R}^D \), we provide the convergence rate of the GL-GP covariance function. We show the stability of the GL (hence the GL-GP algorithm) when there are measurement errors so that predictors do not fall exactly within \( S \). Finally, when \( S \) is an embedded submanifold of \( \mathbb{R}^D \), we provide theory on contraction of the posterior for \( f \) around the true function \( f_0 \) under some regularity assumptions.

2 | GRAPH BASED GAUSSIAN PROCESSES ON SUBSETS OF EUCLIDEAN SPACE

Focusing on Equation (1), we propose a new approach for choosing the covariance function \( C \) in the Gaussian process; our proposed approach differs from current standard methods in not treating \( C \) as a pre-specified function having a small number of unknown parameters (e.g. squared exponential or Mátern) but instead estimates the covariance in a manner that takes into account the structure of the support \( S \) as well as information in both the labelled and unlabelled data. Before introducing the proposed covariance, we provide a review of prediction based on GPs.

2.1 | Gaussian process review

Denote \( f \in \mathbb{R}^m \) to be the discretization of a continuous function \( f \) over \( x_1, x_2, \ldots, x_m \) so that \( f(i) = f(x_i) \). A GP prior for \( f \) implies \( p(f|x_1, x_2, \ldots, x_m) = \mathcal{N}(0, \Sigma) \), where \( \Sigma \in \mathbb{R}^{m \times m} \) is the covariance matrix induced from \( C \), with the \( (i, j) \) element of \( \Sigma \) corresponding to \( \text{cov}\{f(x_i), f(x_j)\} = C(x_i, x_j) \), for \( 1 \leq i, j \leq m \). The prior distribution \( \mathcal{N}(0, \Sigma) \) can be combined with the information in the likelihood function under model (1) to obtain the posterior distribution, which will be used as a basis for inference.

We want to predict \( f(x_i) \), where \( x_i \in \mathcal{X} \). Denote \( f_* \in \mathbb{R}^n \) with \( f_*(i) = f(x_{m+i}) \) for \( i = 1, \ldots, n \). Under a GP prior for \( f \), the joint distribution of \( f \) and \( f_* \) is \( p(f, f_*) = \mathcal{N}(0, \Sigma) \), where \( \Sigma \) is an \((m + n) \times (m + n)\) covariance matrix that can be expressed as \( \Sigma = \begin{bmatrix} \Sigma_{ff} & \Sigma_{ff_*} \\ \Sigma_{f_*f} & \Sigma_{f_*f_*} \end{bmatrix} \), where \( \Sigma_{ff} \in \mathbb{R}^{m \times m} \), \( \Sigma_{f_*f} \in \mathbb{R}^{n \times m} \), and \( \Sigma_{f_*f_*} \in \mathbb{R}^{n \times n} \) are induced from the covariance function \( C \) respectively. Denote \( y \in \mathbb{R}^m \) with \( y_i = y_i \) for \( i = 1, \ldots, m \) to be the observations over \( \{x_1, x_2, \ldots, x_m\} \). Under model (1) and a GP prior, we have \( p(y, f_*) = \mathcal{N}(0, \tilde{\Sigma}) \), where

\[
\tilde{\Sigma} = \Sigma + \begin{bmatrix} \sigma^2_{\text{noise}} I_{m \times m} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{ff} + \sigma^2_{\text{noise}} I_{m \times m} & \Sigma_{ff_*} \\ \Sigma_{f_*f} & \Sigma_{f_*f_*} \end{bmatrix}.
\]
By a direct calculation, the predictive distribution is
\[
p(f_*|y) = \mathcal{N}(\Sigma_{f}f(\Sigma_{ff} + \sigma^2_{\text{noise}}I_{m \times m})^{-1}y, \Sigma_{f}f - \Sigma_{f}f(\Sigma_{ff} + \sigma^2_{\text{noise}}I_{m \times m})^{-1}\Sigma_{ff}).
\]

We refer the readers to Rasmussen (2003) for more background on Gaussian processes.

### 2.2 Graph Laplacian and graph based Gaussian process

The GL is a fundamental tool in spectral graph theory (Chung, 1996). In this section, we first summarize the GL and then introduce the GL-GP, which uses the spectral structure of GL to define a covariance matrix \( \Sigma \) to be used as described in the previous subsection.

Given a dataset \( \mathcal{X} = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}\} \subset \mathbb{R}^D \), we first define a kernel
\[
k_{\varepsilon}(x, x') = \exp \left( -\frac{||x - x'||^2_{\mathbb{R}^D}}{4\varepsilon^2} \right),
\]
(2)

where \( \varepsilon > 0 \) is a bandwidth parameter. Although we can choose a more general kernel within our proposed methodology, we focus on this Gaussian-like choice for simplicity. This kernel is used to define an \((m+n) \times (m+n)\) affinity matrix \( W \) as
\[
W_{ij} := \frac{k_{\varepsilon}(x_i, x_j)}{q_{\varepsilon}(x_i)q_{\varepsilon}(x_j)},
\]
(3)

where \( i, j = 1, \ldots, m+n \) and \( q_{\varepsilon}(x) := \sum_{i=1}^{m+n} k_{\varepsilon}(x, x_i) \). We construct an \((m+n) \times (m+n)\) diagonal matrix \( D \) so that its \( i \)th diagonal entry is
\[
D_{ii} = \sum_{j=1}^{m+n} W_{ij},
\]
(4)

and define the row stochastic transition matrix as \( A = D^{-1}W \). Our main quantity of interest is the GL matrix, which is defined as
\[
L := \frac{A - I}{\varepsilon^2}.
\]
(5)

The affinity matrix \( W \) in Equation (3) is symmetric. Hence \( W \) can be considered as the affinity matrix of an undirected complete graph \( G = (V, E, W) \), where \( V = \{x_i\}^{m+n}_{i=1} \), \( E \) consists of edges connecting any pair of points in \( V \), and \( L \) is the GL over the affinity graph \( G \). Moreover, the GL can be viewed as an infinitesimal generator of a random walk on \( G \) (Chung, 1996).

**Remark 1** In the graph theory literature, \( D - W \) is typically called the unnormalized GL and the matrix \( I - A = I - D^{-1}W \) is called the normalized (or random walk) GL. We call the matrix \( L \) the kernel normalized GL since the affinity is normalized by \( q_{\varepsilon}(x_i)q_{\varepsilon}(x_j) \) in Equation (3).

**Remark 2** The affinity matrix \( W \) in Equation (3) is also considered in many kernel based-machine learning algorithms, for example, diffusion maps (Coifman & Lafon, 2006). In Coifman and
Lafon (2006), the term (3) is called the \textit{a-normalization} with \(a = 1\). The kernel \(k_{\epsilon}(x_{i},x_{j})\) is normalized by \(q_{\epsilon}\) to adjust for the non-uniform sampling density.

A basic spectral property of \(L\) is summarized in the following proposition.

**Proposition 1** Let \(\mu\) be an eigenvalue of \(-L\). Then, \(\mu\) is real and \(0 \leq \mu \leq \frac{1}{\epsilon^{2}}\). In particular, \(0\) is the smallest eigenvalue of \(-L\).

**Proof.** It is sufficient to show that the eigenvalues of \(A\) are real and between \(0\) and \(1\) with \(1\) the largest eigenvalue of \(A\). Let \(K_{ij} = k_{\epsilon}(x_{i},x_{j})\). Then, \(K\) is a symmetric matrix generated by the Gaussian. By Bochner’s theorem, \(K\) is positive semidefinite. If \(Q\) is the diagonal matrix with \(Q_{ii} = \sum_{j=1}^{m+n} K_{ij}(x_{i},x_{j})\), then \(W = Q^{-1}WQ^{-1}\) is positive semidefinite. Let \(\tilde{A} = D^{-1/2}WD^{-1/2}\), which is a symmetric matrix. Hence, the eigenvalues of \(\tilde{A}\) are real and \(\tilde{A}\) is positive semidefinite. Since \(A\) is similar to \(\tilde{A}\) through \(A = D^{-1/2}\tilde{A}D^{1/2}\), the eigenvalues of \(A\) and \(\tilde{A}\) are the same. Finally, since \(A\) is row stochastic, the largest eigenvalue of \(A\) is \(1\) by the fact that the spectral norm is bounded by the \(L^{\infty}\) norm of \(A\).

Suppose the dataset \(\mathcal{X} = \{x_{1}, \ldots, x_{m}, x_{m+1}, \ldots, x_{m+n}\}\) lies within a subset \(S\) of \(\mathbb{R}^{D}\) and we construct the GL matrix \(L\) based on \(\mathcal{X}\) as in Equation (5). Suppose the eigenpairs of the GL over the affinity graph \(G\) are \((-\mu_{i,m+n,\epsilon}, v_{i,m+n,\epsilon})\). Denote \(v_{i,m+n,\epsilon}\) to be the eigenvector associated with the eigenvalue \(\mu_{i,m+n,\epsilon}\) of \(-L\) normalized in the \(\ell^{2}\) norm, where \(i = 0, \ldots, m+n-1\). By Proposition 1, we order \(\mu_{i,m+n,\epsilon}\) so that \(0 = \mu_{0,m+n,\epsilon} < \mu_{1,m+n,\epsilon} \leq \ldots \leq \mu_{m+n-1,m+n,\epsilon}\), where \(\mu_{0,m+n,\epsilon} < \mu_{1,m+n,\epsilon}\) since the graph is connected. Then, we define

\[
H_{\epsilon,K,t} = (m+n) \sum_{i=0}^{K-1} e^{-\mu_{i,m+n,\epsilon}t}v_{i,m+n,\epsilon}v_{i,m+n,\epsilon}^{T} \in \mathbb{R}^{(m+n) \times (m+n)},
\]

(6)

to be the covariance matrix for GP regression over \(\mathcal{X}\).

Since \(H_{\epsilon,K,t}\) is constructed from the GL, we refer to a GP with the covariance (6) as the GL-GP. We will show later that the GL-GP covariance matrix in Equation (6) is an approximation to the GL-GP covariance function on \(S \times S\) and it is associated with the kernel of a compact self-adjoint integral operator over \(L^{2}(S)\). The total dimension of the eigenspaces corresponding to the non-zero eigenvalues of the integral operator is \(K\). We will also show that the eigenvalue \(e^{-\mu_{i,m+n,\epsilon}t}\) is an approximation to the \(i\)th largest eigenvalue of the integral operator and \(\sqrt{m+n}v_{i,m+n,\epsilon}\) is an approximation to the corresponding eigenfunctions normalized in the \(L^{2}\) norm. We will discuss more details of formulation (6) after we define the GL-GP covariance function on \(S \times S\).

For illustration, we consider a two balloons example in \(\mathbb{R}^{3}\). In this case, \(S\) is a set with singularities consisting of 2 dimensional spheres and 1 dimensional line segments. We sample 2530 points on \(S\) as the dataset \(\mathcal{X}\). Through this toy example, we discuss how the parameters \(\epsilon, K\) and \(t\) in the GL-GP covariance matrix \(H_{\epsilon,K,t}\) are related to the geometric structure of set \(S\). We also compare the GL-GP covariance and the GP with squared exponential covariance:

\[
C(x, x') = A \exp\left(-\frac{\|x - x'\|^{2}_{\mathbb{R}^{D}}}{\rho^{2}}\right).
\]

(7)

The parameter \(\epsilon\) in the GL should be large enough so that there are sufficient numbers of points in an \(\epsilon\)-sized Euclidean neighbourhood around \(x\) to obtain information about the local
geometry but not so large as to include points that are not close to \( x \) in intrinsic distance within \( S \). In the two balloons example, we focus on a point \( x_i \) located on the right edge of the left balloon; this point is close in Euclidean distance to points on the left edge of the right balloon. We plot the covariance \( H_{\varepsilon,K,t} \) between this point and the other points in Figure 2 letting \( K = 38 \) and \( t = 0.01 \). In the left panel, \( \varepsilon = \sqrt{0.007} \), which is an appropriate size to define an intrinsic neighbourhood on the set, while in the right panel \( \varepsilon = \sqrt{0.1} \), which bridges the gap and leads to inappropriate covariance across balloons. The parameter \( K \) controls fluctuations in the GL-GP covariance. When \( K \) is larger, higher frequency oscillations are considered in constructing the GL-GP covariance. However, due to the factor \( e^{-H_{\varepsilon,K,t}} \) in the covariance matrix, the amplitudes of those high frequency oscillations are relatively small. In Figure 3, we plot the covariance relative to the same \( x \) as in the previous figure but for different choices of \( K \). When \( K = 8 \) the covariance decreases monotonically with increasing intrinsic distance from \( x_i \), while for \( K = 38 \) there are oscillations with small amplitudes contributing to non-monotonicity. Finally, the diffusion time \( t \) controls the rate of decrease in the covariance as the intrinsic distance increases. Figure 4 shows the impact of varying \( t \) on the covariance relative to the impact of varying the bandwidth in the squared exponential covariance.

From the above discussion, the parameters \( t, \varepsilon \) and \( K \) in the GL-GP covariance have an important impact on the performance. Let \( H_{\varepsilon,K,t} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \), where \( A \) is an \( m \times m \) matrix. We propose to estimate these parameters by maximizing the log of the marginal likelihood,

\[
\log p(y|\theta, H_{\varepsilon,K,t}) = \sum_{i=1}^{n} \log p(y_i|\theta, H_{\varepsilon,K,t}) - \frac{1}{2} \log |H_{\varepsilon,K,t}| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \mathbf{y}^{T} H_{\varepsilon,K,t}^{-1} \mathbf{y}
\]

Figure 2 When \( K = 38, t = 0.01 \) in \( H_{\varepsilon,K,t} \), we plot the row of the Graph Laplacian-based Gaussian process covariance matrix corresponding to a point in the dataset over the 2530 sampled points on \( S \subset \mathbb{R}^3 \) for \( \varepsilon = \sqrt{0.007} \) and \( \varepsilon = \sqrt{0.1} \) respectively [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 3 When \( \varepsilon = \sqrt{0.007}, t = 0.01 \) in \( H_{\varepsilon,K,t} \), we plot the row of Graph Laplacian-based Gaussian process covariance matrix corresponding to a point in the dataset over the 2530 sampled points on \( S \subset \mathbb{R}^3 \) for \( K = 8 \) and \( K = 38 \) respectively [Colour figure can be viewed at wileyonlinelibrary.com]
When $\epsilon = \sqrt{0.007}, K = 38$ in $H_{\epsilon,K,t}$, we plot a row of Graph Laplacian-based Gaussian process covariance matrix corresponding to a point in the dataset over the 2530 sampled points on $S \subset \mathbb{R}^3$ for $t = 0.01$, $t = 0.3$ and $t = 1$ respectively in the three panels on the left. In comparison, we plot the same row of the square exponential covariance matrix over the 2530 sampled points on $S \subset \mathbb{R}^3$ induced by Equation (7) for $A = 12$ and $\rho = \sqrt{0.015}, \rho = \sqrt{0.2}$ and $\rho = \sqrt{0.5}$ respectively in the three panels on the right [Colour figure can be viewed at wileyonlinelibrary.com]

\[
\log p(y|\epsilon, K, t, \sigma_{\text{noise}}) = -y^T(A + \sigma_{\text{noise}}^2 I_{m \times m})^{-1}y - \log(\det(A + \sigma_{\text{noise}}^2 I_{m \times m})) - \frac{m}{2} \log(2\pi). \tag{8}
\]

An identical strategy is used routinely in the literature for estimating GP covariance parameters (Rasmussen, 2003). The parameters $\epsilon$ and $K$ are related to the GL over the dataset, while the parameters $t$ and $\sigma_{\text{noise}}$ are not. To maximize Equation (8), we propose to alternate between a grid search for $\epsilon$ and $K$ and gradient descent for $t$ and $\sigma_{\text{noise}}$.

### 2.3 Nyström extension of the GL-GP covariance matrix

Suppose the dataset $\mathcal{X} = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}\} \subset S \subset \mathbb{R}^D$, $L$ is the GL constructed from $\mathcal{X}$ following Equation (5), and $(\mu_{i,m+n,\epsilon}, \tilde{v}_{i,m+n,\epsilon})$ is the $i$th eigenpair of $-L$ with $\tilde{v}_{i,m+n,\epsilon}$
normalized in $\ell^2$. Then, construct the GL-GP covariance matrix $H_{\varepsilon,K,t}$ as in Equation (6). Now, suppose we have $\ell$ additional samples $\{x_{m+n+1}, \ldots, x_{m+n+\ell}\}$ on $S$ and set $\mathcal{X}^* = \{x_1, \ldots, x_m, x_{m+n+1}, \ldots, x_m+n+\ell\}$. In this subsection, we propose a computationally efficient Nyström extension algorithm to find an extension of $H_{\varepsilon,K,t}$ over $\mathcal{X}^*$ without needing to rerun the whole GL-GP algorithm.

The idea is using the full knowledge of the kernel to construct an extension of the eigenvector $\tilde{v}_{n,m+n,\varepsilon}$ to $\mathcal{X}^*$. Recall that $q_\varepsilon(x) := \sum_{i=1}^{m+n} k_\varepsilon(x, x_i)$. Define the extension matrix $E \in \mathbb{R}^{(m+n+\ell) \times (m+n)}$ by

$$E_{ij} := \frac{k_\varepsilon(x_i, x_j)}{q_\varepsilon(x_i) q_\varepsilon(x_j)}$$

where $x_i \in \mathcal{X}^*$ and $x_j \in \mathcal{X}$. We have the following immediate proposition.

**Proposition 2** For $x_j \in \mathcal{X}$, we have $\frac{1}{1-\varepsilon^2 \mu_{l,m+n,\varepsilon}} E \tilde{v}_{l,m+n,\varepsilon}(j) = \tilde{v}_{l,m+n,\varepsilon}(j)$.

**Proof.** Observe that $A_{ij} = E_{ij}$ for $1 \leq i, j \leq m+n$. Since $\tilde{v}_{l,m+n,\varepsilon}$ is the eigenvector of $A$ corresponding to the eigenvalue $1 - \varepsilon^2 \mu_{l,m+n,\varepsilon}$, for $1 \leq j \leq m+n$, we have $\frac{1}{1-\varepsilon^2 \mu_{l,m+n,\varepsilon}} E \tilde{v}_{l,m+n,\varepsilon}(j) = \frac{1}{1-\varepsilon^2 \mu_{l,m+n,\varepsilon}} A \tilde{v}_{l,m+n,\varepsilon}(j) = \tilde{v}_{l,m+n,\varepsilon}(j)$.

This proposition says that $\frac{1}{1-\varepsilon^2 \mu_{l,m+n,\varepsilon}} E \tilde{v}_{l,m+n,\varepsilon}$ is an extension of the eigenvector $\tilde{v}_{l,m+n,\varepsilon}$ from $\mathcal{X}$ to $\mathcal{X}^*$. With the parameters $\varepsilon, K$, and $t$, the Nyström extension gives

$$H_{\varepsilon,K,t}^* := E \left( \sum_{i=0}^{K-1} \frac{e^{-\mu_{l,m+n,\varepsilon} t}}{(1-\varepsilon^2 \mu_{l,m+n,\varepsilon})^2} \tilde{v}_{l,m+n,\varepsilon} \tilde{v}_{l,m+n,\varepsilon}^\top \right) E^\top \in \mathbb{R}^{(m+n+\ell) \times (m+n+\ell)},$$

which is an extension of the GL-GP covariance matrix $H_{\varepsilon,K,t}$ over $\mathcal{X}^*$. Based on the definition of the extension matrix $E$, for $1 \leq i, j \leq l$, $H_{\varepsilon,K,t}^*(m+n+i, m+n+j)$ only depends on $\mathcal{X}$, $x_{m+n+i}$ and $x_{m+n+j}$ and not on the remaining samples. Hence, it will not change when further samples are added.

We will justify the error in the Nyström extension of the GL-GP covariance structure and the error in the prediction after we introduce the GL-GP covariance function in Theorem 3. A simulation result of the Nyström extension is provided in Section H of the Supplementary Materials.

**Remark 3** The Nyström extension can be used to induce a covariance function for the GL-GP. For any $x, y \in S$, let $\mathcal{X} = \{x_1, \ldots, x_m\}$ and $\mathcal{X}^* = \{x_1, \ldots, x_{m+n}, x, y\}$. Construct the Nyström extension matrix $H_{\varepsilon,K,t}^*$ based on $\mathcal{X}^*$ and define $C_{\varepsilon,K}(x, y, t)$ to be the corresponding entry in the matrix. Such $C_{\varepsilon,K}(x, y, t)$ shares the property that the restriction over $\mathcal{X}$ is equal to the covariance matrix. However, this definition relies on the samples $\mathcal{X}$. Later, we will introduce a more natural way to define the covariance function that is independent of samples.
3 | APPLICATIONS

In this section, we apply our GL-GP approach in three different examples. In all cases, we compare the GL-GP with the GP with the covariance (7). The first case is the two balloons example. The second case is a spiral, which is a compact manifold with boundary. The third case is a complicated 2-dimensional compact manifold with boundary coming from the 3D scan of a human’s right hand.

3.1 | GL-GP on two balloons

We have two unit spheres $S_1$ and $S_2$ centred at $(1.2, 0, 2)$ and $(-1.2, 0, 2)$ respectively. We connect the south poles $(1.2, 0, 1)$ and $(-1.2, 0, 1)$ to the origin by two line segments $L_1$ and $L_2$; $S = S_1 \cup S_2 \cup L_1 \cup L_2$. Globally, $S$ does not have a manifold structure, and the Hausdorff dimension of $S$ is 2. We sample 30 points from a uniform density on one of the spheres and sample 3 points from a uniform on the attached line segment. We then find the 33 points symmetrical to the sampled points on the other sphere and line segment. Those 66 points are the labelled data points $\{x_1, \ldots, x_{66}\}$. 2200 unlabelled points $\{x_{67}, \ldots, x_{2266}\}$ are sampled in the same way with 2000 points on the spheres and 200 points on the line segment. The labels are sampled via (1) with $\sigma_{\text{noise}} = 1$ and $f(x) = 5 \times d_S(x, a)$, for $x \in S$, with $a = (-1.2, 0, 3)$ then the north pole of $S_2$ and $d_S$ denoting the distance metric for the space $S$. The distance on $S$ between two points on the same sphere is the length of the shorter part of the great circle connecting those points. The distance between a point $x$ on a sphere and a point $x'$ on the segment attached to the sphere is the sum of the distance on $S$ between $x$ to the attaching point and the Euclidean distance between the attaching point to $x'$. Hence, the formula for $d_S(x, a)$ is

$$d_S(x, a) = \begin{cases} 
\arccos(|x - (1.2, 0, 2)| \cdot (0, 0, 1)) & \text{if } x \in S_2, \\
\pi + ||x - (1.2, 0, 1)||_{\mathbb{R}^3} & \text{if } x \in L_2, \\
\pi + ||(1.2, 0, 1)||_{\mathbb{R}^3} + ||x||_{\mathbb{R}^3} & \text{if } x \in L_1, \\
\pi + 2||(1.2, 0, 1)||_{\mathbb{R}^3} + \arccos(|x - (1.2, 0, 2)| \cdot (0, 0, -1)) & \text{if } x \in S_1.
\end{cases}$$

We plot the labels over $x_i$ for $i = 1, \ldots, 66$ and the true values in the top two panels in Figure 5.

We use both the GL-GP and the GP with square exponential covariance to predict the response values for the unlabelled data. In both cases, covariance parameters are chosen by maximizing the marginal likelihood. We calculate the root mean square error (RMSE) relative to the true value of the regression function at the unlabelled data points. Maximizing (8), the parameters are $K = 30$, $\epsilon = \sqrt{0.012}$, $t = 0.33$ and $\sigma_{\text{noise}} = \sqrt{0.9}$, leading to an RMSE 0.721. For the GP, we obtain $A = 250$, $\rho = \sqrt{0.54}$, $\sigma_{\text{noise}} = \sqrt{0.9}$ and an RMSE 1.812. Figure 5 compares the prediction by the GL-GP and the GP over $x_i$, for $i = 67, \ldots, 2266$. For better visualization, in Figure 6, we compare the difference between the prediction and the ground truth over $x_i$, for $i = 67, \ldots, 2266$. The GL-GP performs better over the regions indicated in the boxes and their symmetric parts on the other sphere. Since the square exponential covariance in the GP tries to smooth the predictive values between region 1 and its symmetric part, the prediction in region 1 is lower than the true value.
Figure 5  The true regression function is 5 times the distance of a point to the north pole of the sphere on the left. Let \( \{x_1, \ldots, x_{66}\} = \) labelled points and \( \{x_{67}, \ldots, x_{2266}\} = \) unlabelled points. Top left: response variables over labelled points. Top right: ground truth over all points. Bottom left: prediction of Graph Laplacian based Gaussian process (GL-GP) over unlabelled points with \( RMSE = 0.721 \). Bottom right: prediction of (squared exponential Euclidean) GP over unlabelled points with \( RMSE = 1.812 \) [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 6  Left: difference between prediction of Graph Laplacian based Gaussian process (GL-GP) and ground truth over \( \{x_i\}, i = 67, \ldots, 2266 \). Right: difference between prediction of GP and ground truth over \( \{x_i\}, \) for \( i = 67, \ldots, 2266 \) [Colour figure can be viewed at wileyonlinelibrary.com]
3.2 Spiral case

We consider a spiral $M$ embedded in $\mathbb{R}^2$ parametrized by

$$\gamma(\theta) = ((\theta + 4)^{0.7} \cos(\theta), (\theta + 4)^{0.7} \sin(\theta)) \in \mathbb{R}^2, \quad \text{where } \theta \in [0, 8\pi).$$

We sample 60 labelled points $\{\theta_1, \ldots, \theta_{60}\}$ and 1500 unlabelled points $\{\theta_{61}, \ldots, \theta_{1560}\}$ on $[0, 8\pi)$ from the uniform density. Let $x_i = \gamma(\theta_i)$ for $i = 1, \ldots, 1560$. The labels are sampled under Equation (1) with $\sigma_{\text{noise}} = 1$ and $f(\gamma(\theta)) = 3 \sin \left( \frac{\theta}{10} \right) + 3 \cos \left( \frac{\theta}{2} \right) + 4 \sin \left( \frac{4\theta}{5} \right)$. We plot the labels over $x_i$ for $i = 1, \ldots, 60$, and the ground truth in the top two panels in Figure 7. Maximizing the marginal likelihoods, we obtain $K = 9, \varepsilon = \sqrt{0.1}, t = 0.02$ and $\sigma_{\text{noise}} = \sqrt{1.3}$ for the GL-GP, leading to an RMSE of 0.853. For the GP with covariance (7), we obtain $A = 17, \rho = \sqrt{2.2}, \sigma_{\text{noise}} = \sqrt{0.7}$ and an RMSE of 1.967. The bottom two panels in Figure 7 show the predictions of the two different approaches. For better visualization, in Figure 8, we plot the predictions over the parameter $\theta_i$. The GL-GP greatly improves predictive performance. As an example, we provide an analysis over region 1. When we ignore the intrinsic geometry of the spiral, based on the response variables over the labelled data, there is a potential increase along the direction outward over region 1.
FIGURE 8 Let \( \{ \theta_1, \ldots, \theta_{60} \} \) = labelled inputs and \( \{ \theta_{61}, \ldots, \theta_{1560} \} \) = unlabelled inputs. Left: red points are response variables for labelled inputs, and blue points are ground truth for all \( \{ \theta_i \}, i = 1, \ldots, 1560 \). Middle: red points are prediction of Graph Laplacian based Gaussian process (GL-GP) over \( \{ \theta_i \} \) for \( i = 61, \ldots, 1560 \). Right: red points are prediction of GP over \( \{ \theta_i \} \) for \( i = 61, \ldots, 1560 \). [Colour figure can be viewed at wileyonlinelibrary.com]

Moreover, for the points that are close to region 1 in Euclidean distance on the inner neighbour arc, the labels are negative and large in magnitude. Hence, the prediction by the GP over region 1 is negative.

3.3 Temperature distribution on the right hand of a Raynaud’s disease patient

The 3D surface scan of a human hand is a 2-dimensional compact manifold with boundary isometrically embedded in \( \mathbb{R}^3 \). The dataset in this example is a discrete version of a scan of a right hand in Romero et al. (2017) and consists of 1950 points. There are 50 labelled points \( \{ x_i \}_{i=1}^{50} \) among which 25 points are on the fingers and 25 points on the palm. The remaining points \( \{ x_i \}_{i=51}^{1950} \) are unlabelled. Let \( \mathcal{X} = \{ x_i \}_{i=1}^{1950} \). For \( x_i, x_j \in \mathcal{X} \), \( W_{ij} = \exp \left( -\frac{||x_i - x_j||^2}{0.012} \right) \in \mathbb{R}^{1950 \times 1950} \). Define a diagonal matrix \( D_{ii} = \sum_{j=1}^{1950} W_{ij} \). Let \( L_{un} = D - W \) denote the unnormalized GL, which is different from the GL defined in the previous section. Let \( v_3, v_6, v_{13} \) be the eigenvectors normalized in \( \ell^2 \) corresponding to the 3rd, 6th and 13th smallest eigenvalues of \( L_{un} \). The labels are sampled under Equation (1) with \( \sigma_{\text{noise}} = 0.02 \) and

\[
f(x_i) = 8.118v_3(i) + 2.46v_6(i) + 0.738v_{13}(i) + 32.2,
\]

where \( v_{13}(i) = v_{13}(i) \) if \( v_{13}(i) > 0 \) and \( v_{13}(i) = 0 \) if \( v_{13}(i) \leq 0 \). The function \( f \) is used to simulate the temperature distribution over the samples. We plot the labels over \( x_i \) for \( i = 1, \ldots, 50 \), and the ground truth viewed from the back of the right hand in Figure 9.

Maximizing the marginal likelihood, we obtain \( K = 11, \epsilon = \sqrt{0.0015}, t = 0.002 \) and \( \sigma_{\text{noise}} = \sqrt{0.004} \) for the GL-GP, leading to an RMSE 0.062. For the GP with squared exponential covariance, the optimal parameters are \( A = 405, \rho = \sqrt{0.0145}, \sigma_{\text{noise}} = \sqrt{0.012} \), which leads to an RMSE 0.169. Figure 10 compares the prediction by GL-GP and GP over \( \{ x_i \} \), for \( i = 51, \ldots, 1950 \). For better visualization, in Figure 11, we compare the difference between the prediction and ground truth over \( \{ x_i \} \), for \( i = 51, \ldots, 1950 \). The GL-GP performs better over the regions indicated in the boxes. In particular, in region 3, since the squared exponential covariance in the GP
The response variables over labeled points

The ground truth over all points

**FIGURE 9** Let \( \{x_1, \ldots, x_{50}\} \) be the labels points and \( \{x_{51}, \ldots, x_{1950}\} \) be the unlabelled points. Left: The plot of the response variables over the labelled points \( \{x_i\}, \ i = 1, \ldots, 50 \). Right: The plot of the ground truth over all points \( \{x_i\}, \ i = 1, \ldots, 1950 \) [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 10** Left 2 panels: 2 different views of the prediction by the Gaussian process (GP) with square exponential covariance over over \( \{x_i\}, \ for \ i = 51, \ldots, 1950 \) with \( \text{RSME} = 0.169 \). Middle 2 panels: 2 different views of the prediction by the Graph Laplacian based GP over \( \{x_i\}, \ for \ i = 51, \ldots, 1950 \) with \( \text{RSME} = 0.062 \). Right 2 panels: 2 different views of the ground truth over \( \{x_i\}, \ for \ i = 51, \ldots, 1950 \) [Colour figure can be viewed at wileyonlinelibrary.com]
tries to smooth the predictive values between the ring finger and the little finger, the prediction in region 3 is lower than the true value. Based on the labelled information, there is a potential decrease along the index finger from the bottom part to the tip. However, there is a faster decrease from the middle finger to the index finger. The Euclidean distance term in the covariance of the GP reflects such a fast decrease; as a result, the predictive values of the GP are higher over region 2 and lower on the tip of the index finger. Similar explanation can be applied to regions 1 and 4.

4 | THEORY OF GRAPH BASED GAUSSIAN PROCESS

4.1 | Graph based Gaussian process on subsets of $\mathbb{R}^D$

In this section, we introduce theory of GLs when the dataset is sampled from a subset $S$ of $\mathbb{R}^D$. We are going to associate the GL with an integral operator on the set $S$. Then, we apply spectral theory to the integral operator to define the corresponding GL-GP covariance function. We first make the following assumption on the subset $S$ and the probability measure on $S$.

**Assumption 1** Suppose $S$ is a compact connected subset of $\mathbb{R}^D$. Suppose $(S, \mathcal{F}, P)$ is a probability space, where $P$ is a probability measure defined over the Borel sigma algebra $\mathcal{F}$ on $S$. Suppose $\mathcal{X} = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}\}$ are $m + n$ i.i.d. samples based on the measure $P$.

**Remark 4** There are natural measures on $S$ induced by metrics in the ambient Euclidean space $\mathbb{R}^D$, for example, the $d$ dimensional Hausdorff measure. However, we do not require $P$ to be absolutely continuous with respect to any of these measures.

Based on Assumption 1, we define an integral operator associated with the GL.
For any function \( f(x, x') \) on \( S \times S \), we define the operator \( P \) with respect to the probability measure \( P \) as follows, \( Pf(x) := \int_S f(x, x') dP(x') \). For the kernel \( k_i(x, x') \), we define \( d_i(x) := Pk_i(x) \) and \( Q_i(x, x') := \frac{k_i(x, x')}{d_i(x)d_i(x')} \). For \( f(x) \) on \( S \), we define the operator:

\[
T_\epsilon f(x) := \frac{PQ_\epsilon f(x)}{PQ_\epsilon(x)} = \int_S Q_i(x, x')f(x')dP(x') \int_S Q_i(x, x')dP(x').
\]

At last, we define \( l_\epsilon := \frac{I^n - T_\epsilon}{\epsilon^n} \).

The operator \( T_\epsilon \) has the following important property. The property motivates our covariance function for the GP. Hence, we state it as a theorem here. The proof is postponed to Section A of the Supplementary Materials.

**Theorem 1** Under Assumption 1, for any fixed \( \epsilon \), \( T_\epsilon \) is a linear, compact, self-adjoint operator from \( L^2(S, P) \) to \( L^2(S, P) \). Moreover, suppose \( A \) is the matrix in Equation (5) constructed from \( \mathcal{X} \), then for any \( f \in L^2(S, P) \), we have the following convergence result: for any \( x_k \in \mathcal{X} \),

\[
\lim_{m+n \to \infty} \sum_{i=1}^{m+n} A_k f(x_i) = T_\epsilon f(x_k), \quad a.s.
\]

We discuss the theoretical motivation for constructing a covariance function incorporating geometric information about the set \( S \). By Theorem 1, for a fixed \( \epsilon \), since \( T_\epsilon \) is compact and self-adjoint, the eigenvalues of \( l_\epsilon \) satisfy \( \lambda_{0, \epsilon} \leq \lambda_{1, \epsilon} \leq \cdots \). Suppose \( (\lambda_{i, \epsilon}, \varphi_{i, \epsilon}) \) is the \( i \)th eigenpair of \( l_\epsilon \) and each \( \varphi_{i, \epsilon} \) is normalized in \( L^2(S, P) \). Then \( \{ \varphi_{i, \epsilon} \} \) form an orthonormal basis of \( L^2(S, P) \). For any function \( f \in L^2(S, P) \), \( f \) can be expressed as a unique linear combination of \( \{ \varphi_{i, \epsilon} \} \), that is, \( f = \sum_{i=1}^{\infty} a_i \varphi_{i, \epsilon} \), a.e. Note that \( \sum_{i=1}^{\infty} a_i^2 = ||f||^2_{L^2(S, P)} < \infty \). Hence, \( a_i \to 0 \) as \( i \to \infty \) and a finite linear combination \( \sum_{i=1}^{L} a_i \varphi_{i, \epsilon} \) is an approximation of \( f \) in an \( L^2(S, P) \) sense. Based on the above observation, if our unknown regression function is \( f \in L^2(S, P) \), it is reasonable to propose a GP covariance function on \( S \) by using finitely many eigenpairs of \( T_\epsilon \). More precisely, fixing \( K \in \mathbb{N} \) and \( t > 0 \), we define

\[
C_{\epsilon, K}(x, x', t) = \sum_{i=0}^{K-1} e^{-\lambda_{i, \epsilon} t} \varphi_{i, \epsilon}(x) \varphi_{i, \epsilon}(x'),
\]

for \( x, y \in S \). This covariance function involves three parameters, \( t, \epsilon \) and \( K \). The parameter \( t \) can be regarded as controlling the bandwidth, and hence the choice of \( t \) should depend on the regularity of \( f \). The parameter \( \epsilon \) is used in the construction of the operator \( T_\epsilon \) over \( S \); \( \epsilon \) should depend on the regularity of \( S \). In contrast, \( K \) should depend on both \( S \) and the regularity of \( f \).

We give an intuitive justification to show that the GL-GP covariance matrix in Equation (6) approximates the GL-GP covariance function over \( \mathcal{X} \). The pointwise convergence of the matrix \( A \) to \( T_\epsilon \) in Theorem 1 suggests that when \( m + n \) is large enough, the \( i \)th smallest eigenvalue of \( -L \) approximates the \( i \)th smallest eigenvalue of \( l_\epsilon \). The corresponding eigenvector approximates \( \varphi_{i, \epsilon} \) over \( \mathcal{X} \) if the eigenvector is properly normalized. For any continuous \( f \in L^2(S, P) \), by the law of large numbers,
By the same argument as in Theorem 1, in Equation (3). Let

\[ f \] instead of \( f \)

for \( f \) and \( vi \) normalized in \( \ell^2 \), then \( (\sqrt{m+n})vi \) approximates \( \varphi_{i,t} \) over \( \mathcal{X} \). Hence, \( H^K_{i,t} \) approximates \( C_{i,K}(x,x',t) \) over \( \mathcal{X} \). To rigorously justify this intuition, we need to impose more structure on \( S \). We will discuss this in the next section.

\subsection{Graph based Gaussian process on manifolds}

In this section, we consider the special case in which \( S \) has a manifold structure and develop additional theory supporting the GL-GP in this case. Specifically, we make the following assumption.

\textbf{Assumption 2} Let \( M \) be a \( d \)-dimensional smooth, closed and connected Riemannian manifold isometrically embedded in \( \mathbb{R}^D \) through \( i : M \rightarrow \mathbb{R}^D \). Let \( S = i(M) \). Suppose \((M, \mathcal{F}, P)\) is a probability space, where \( P \) is a probability measure defined over the Borel sigma algebra \( \mathcal{F} \) on \( M \). We assume \( P \) is absolutely continuous with respect to the volume measure on \( M \), that is, \( dP = p dV \) by the Radon-Nikodym theorem, where \( p \) is the probability density function on \( M \) and \( dV \) is the volume form. We further assume \( p \) is smooth and is bounded from below by \( p_m > 0 \). \( \mathcal{X} = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}\} \) are i.i.d. sampled from \( P \).

Since \( i \) is an isometry, the function \( f \) in model (1) can be equivalently viewed as \( f : M \rightarrow \mathbb{R} \) instead of \( f : S \rightarrow \mathbb{R} \). We start by modifying the definition of GL and the operator \( T_\varepsilon \) for datasets and functions on \( M \). Based on Assumption 2, we define the kernel

\[ k_\varepsilon(x,x') = \exp \left( -\frac{||i(x) - i(x')||^2_{\mathbb{R}^D}}{4\varepsilon^2} \right) \]

for \( x, x' \in M \). We use the kernel \( k_\varepsilon(x,x') \) to construct an \( (m+n) \times (m+n) \) affinity matrix over \( \mathcal{X} \) as in Equation (3). Let \( L \) be the GL of the affinity graph \( G \) with vertices \( V = \mathcal{X} \).

In the manifold case, quantities in Definition 1 have the following expansions. For any function \( f(x,x') \) on \( M \times M \), we have

\[ Pf(x) := \int_M f(x,x') dP(x') = \int_M f(x,x') p(x') dV(x'). \]

For \( f(x) \) on \( M \), we have

\[ T_\varepsilon f(x) := \frac{PQ_\varepsilon f(x)}{PQ_\varepsilon(x)} = \frac{\int_M Q_\varepsilon(x,x') f(x') p(x') dV(x')}{\int_M Q_\varepsilon(x,x') p(x') dV(x')} \]

By the same argument as in Theorem 1, \( T_\varepsilon \) is a linear, compact, self-adjoint operator from \( L^2(M) \) to \( L^2(M) \). Since the manifold is compact and smooth, \( Q_\varepsilon(x,x') \) is smooth. Hence, \( T_\varepsilon f \in C^\infty(M) \). We have the following proposition about the eigenvalues of \( l_\varepsilon \).

\textbf{Proposition 3} Under Assumption 2, the eigenvalues of \( l_\varepsilon = \frac{l - T_\varepsilon}{\varepsilon^2} \) satisfy \( 0 \leq \lambda_{0,\varepsilon} \leq \lambda_{1,\varepsilon} \leq \cdots \).
Proof. Note that $T_\varepsilon$ is a linear, compact and self-adjoint operator from $L^2(M)$ to $L^2(M)$. It is sufficient to prove the eigenvalues of $T_\varepsilon$ are bounded by 1. Suppose $(\lambda, f)$ is an eigenpair of $T_\varepsilon$. Since the manifold is compact and smooth, $Q_\varepsilon(x, x')$ is smooth. Hence, $f$ is smooth and $\|f\|_\infty < \infty$. Since $Q_\varepsilon(x, x') > 0$, $T_\varepsilon f = \lambda f$ implies that

$$|\lambda| = \frac{\|T_\varepsilon f\|_\infty}{\|f\|_\infty} \leq \max_{x \in M} \frac{\int_M Q_\varepsilon(x, x')(f(x')/\|f\|_\infty)p(x')dV(x')}{\int_M Q_\varepsilon(x, x')p(x')dV(x')} \leq 1.$$ 

Suppose $(\lambda_i, \varphi_i)$ is the $i$th eigenpair of $l_\varepsilon$ and each $\varphi_i$ is normalized in $L^2(M)$. Then $\{\varphi_i\}$ form an orthonormal basis of $L^2(M)$. The covariance function $C_{\varepsilon,K}(x, x', t)$ under the manifold setup is defined in Equation (13) for $x, x' \in M$, but is constructed from the GL in a slightly different manner. When $S$ is a general subset of $\mathbb{R}^D$, we do not impose any other measures on $S$ except the probability measure. However, when $M$ is a Riemannian manifold, there is natural measure, namely the volume measure, induced by the Riemannian metric of the manifold. Hence, the eigenfunctions of $l_\varepsilon$ should be normalized in the $L^2(M)$ norm rather than $L^2(M, \nu)$ norm; proper normalization ensures that the eigenvectors of the GL can approximate the eigenfunctions of $l_\varepsilon$. Hence, we introduce the following definition.

Definition 2 Under Assumption 2, suppose $\tilde{v}$ is an eigenvector of $L$ which is normalized in the $\ell^2$ norm. Let $N(i) = |B_\varepsilon \cap \{t(x_1), \ldots, t(x_{m+n})\}|$, the number of points in an $\varepsilon$ ball in $\mathbb{R}^D$ around $x_i$. Then, we define the $\ell^2$ norm of $\tilde{v}$ with respect to the inverse estimated probability density $1/\hat{p}$ as:

$$\|\tilde{v}\|_{\ell^2(1/\hat{p})} := \sqrt{\frac{|S^{d-1}| \varepsilon^d}{d} \sum_{i=1}^{m+n} \tilde{v}^2(i) / N(i)},$$

where $|S^{d-1}|$ is the volume of the sphere $S^{d-1}$.

In the above definition, we approximate $p$ by using a simple $0 - 1$ kernel with bandwidth $\varepsilon$; refer to Dunson et al. (2021) for a detailed discussion.

Denote $\mu_{i,m+n,\varepsilon}$ to be the $i$th eigenvalue of $-L$ with the associated eigenvector $\tilde{v}_{i,m+n,\varepsilon}$ normalized in the $\ell^2$ norm, where $i = 0, \ldots, m + n - 1$. We order $\mu_{i,m+n,\varepsilon}$ so that $\mu_{0,m+n,\varepsilon} \leq \mu_{1,m+n,\varepsilon} \leq \ldots \leq \mu_{m+n-1,m+n,\varepsilon}$. If we define

$$v_{i,m+n,\varepsilon} = \frac{\tilde{v}_{i,m+n,\varepsilon}}{\|\tilde{v}_{i,m+n,\varepsilon}\|_{\ell^2(1/\hat{p})}}, \quad (14)$$

then $v_{i,m+n,\varepsilon}$ can be regarded as a discretization of some function that is normalized in the $L^2(M)$ norm. Fixing $K \in \mathbb{N}$ and $t > 0$, we define

$$\tilde{H}_{\varepsilon,K,t} = \sum_{i=0}^{K-1} e^{-\mu_{i,m+n,\varepsilon}t}v_{i,m+n,\varepsilon}v_{i,m+n,\varepsilon}^T \in \mathbb{R}^{(m+n) \times (m+n)} \quad (15)$$

to be the covariance matrix for GP regression over $\mathcal{X}$ on the manifold $M$. When the probability density is uniform, Equations (6) and (15) are equivalent. Hence, the general GL-GP algorithm in Equation (6) can be viewed as a simplification of Equation (15). The GL-GP algorithm on manifolds is described in Section B of the Supplementary Materials.
4.3 Convergence of $\tilde{H}_{e,K,t}$ and its Nyström extension under the manifold setup

First, we show convergence of the covariance matrix $\tilde{H}_{e,K,t}$ to the covariance function $C^K_e(x, x', t)$ over the dataset $\mathcal{X} = \{x_i\}_{i=1}^{m+n}$ under the manifold setup. More precisely, we provide the convergence rate of entry $i, j$ of the matrix $\tilde{H}_{e,K,t}$ to $C_e(x_i, x_j, t)$ as $m + n \to \infty$. To state our main theorem, we recall the Laplace–Beltrami operator $\Delta$ of the manifold $M$. Here, $\Delta$ is an intrinsic differential operator on the manifold generalizing the notion of the second order derivative on an interval, with the eigenpairs reflecting the geometric and topological structure of the manifold. Let $\sigma(-\Delta) = \{\lambda_i\}_{i=0}^\infty$ be the spectrum of $-\Delta$. By the standard elliptic theory, we have $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$ and each eigenvalue has finite multiplicity. Denote by $\varphi_i$ the eigenfunction normalized in $L^2(M)$ corresponding to $\lambda_i$; that is, for each $i \in \mathbb{N}$, we have $\Delta \varphi_i = -\lambda_i \varphi_i$. It is well known that $\{\varphi_i\}$ forms an orthonormal basis of $L^2(M)$. The main theorem of this section is as follows. The proof of the theorem is in Section C of the Supplementary Materials.

**Theorem 2** Under Assumption 2, let $\lambda_i$ be the $i$th eigenvalue of $-\Delta$. Fixing $K \in \mathbb{N}$, we let $\Gamma_K := \min_{1 \leq i \leq K} \text{dist}(\lambda_i, \sigma(-\Delta)\backslash\{\lambda_i\})$. If $\epsilon$ is small enough,

$$
\epsilon \leq \mathcal{H}_1 \min \left( \frac{\min(\Gamma_K, 1)}{\mathcal{H}_2 + \lambda_K^{d/2+5}}, \frac{1}{(\mathcal{H}_3 + \lambda_K^{(5d+7)/4})^2} \right),
$$

and $m + n$ is sufficiently large so that $\left( \frac{\log(m+n)}{m+n} \right)^{\frac{1}{m+n+1}} \leq \epsilon$, then for any $t$ less than the diameter of $M$, with probability greater than $1 - (m + n)^{-2}$,

$$
\max_{x_i, x_j \in \mathcal{X}} |\tilde{H}_{e,K,t}(i, j) - C_e(x_i, x_j, t)| \leq \mathcal{H}_4 K^2 \epsilon^{1/2}.
$$

$
\mathcal{H}_1$ and $
\mathcal{H}_2, \mathcal{H}_3 > 1$ are constants depending on $d$, $p_m$, the $C^2$ norm of $\rho$, and the volume, injectivity radius, sectional curvature and second fundamental form of the manifold. $
\mathcal{H}_4$ depends on $d$, $p_m$, the $C^2$ norm of $\rho$, and the diameter, volume and Ricci curvature of $M$.

Note that $\tilde{H}_{e,K,t}$ is composed of the eigenpairs of $-L$, while $C_e(x)$ is composed of the eigenpairs of $l_e$. Therefore, at first glance, the convergence rate of $\tilde{H}_{e,K,t}$ to $C_e(x)$ seems to be irrelevant to the Laplace–Beltrami operator. However, to control the convergence rate, we need to bound the eigen-gaps, the magnitude of the eigenvalues and the $L^\infty$ norm of the eigenfunctions of $l_e$, and those terms can be controlled by the eigenvalues of $-\Delta$. The relationship between $l_e$ to $-\Delta$, particularly the spectral convergence when $\epsilon \to 0$, is detailed in Lemma S.3 in Section C of the Supplementary Materials. On the other hand, the eigenvalues of $-\Delta$ are completely determined by the geometry of the manifold. Hence, bounding the convergence rate by the eigenvalues of $-\Delta$ shows that the error between $\tilde{H}_{e,K,t}$ and $C_e(x)$ is determined by the geometry of the manifold.

The GL-GP covariance matrix $\tilde{H}_{e,K,t}$ can also be used to recover the heat kernel of the manifold $M$. The heat kernel on a manifold $M$ is the fundamental solution to the heat equation with an appropriate boundary condition. It describes the diffusion process of the heat flow out of a point on the manifold along the intrinsic distance. Hence, the result that the GL-GP matrix approximates the heat kernel implies that the covariance structure of GL-GP also varies with respect to the intrinsic distance of the manifold as the bandwidth $t$ in $H_{e,K,t}$ changes. We refer the readers to Theorem 3 in Dunson et al. (2021) for the convergence rate from $H_{e,K,t}$ to the heat kernel.
Remark 5 Suppose we consider a fixed manifold $M$. As the sample size $m + n$ increases, the relation between $\varepsilon$ and $m + n$ and the lower bound on the p.d.f implies that the sampling points become increasingly dense in $M$. Hence, Theorem 2 matches with the fixed-domain asymptotics regime considered in spatial statistics. Let $|\text{Rm}(M)|$ denote the norm of the curvature tensor of $M$. Consider the set of all manifolds $\mathcal{M} := \{ M : \text{dim}(M) = d, \ \text{diam}(M) < D, ~ \text{vol}(M) > v, |\text{Rm}(M)| \leq \tau \}$, where $d \in \mathbb{N}, D > 0, v > 0, \tau \geq 0$. Suppose we have a sequence of manifolds $\{ M_i \}_{i=1}^{\infty} \subset \mathcal{M}$, and sample $i = m + n$ points on each $M_i$. Then, Theorem 2 still holds for each $M_i$, when $i$ is sufficiently large. A future direction is removing the diameter upper bound on the manifold. If Theorem 2 holds for manifolds without this bound, then the theorem holds in the mixed-domain asymptotics regime (Chang et al., 2017) in spatial statistics.

If we treat $K$, $\lambda_k$ and the eigengaps as constants and focus on the relation between $\varepsilon$ and $m + n$, then we have the following corollary by applying the results in Dunson et al. (2021) and the same proof of the above theorem. The corollary says that if we treat $K$, $\lambda_k$ and the eigengaps as constants, then the convergence rate of $\tilde{H}_{\varepsilon,K,i}(i,j)$ to $C_{\varepsilon,K}(x_i, x_j, t)$ is $O \left( \left( \frac{\log(m+n)}{m+n} \right)^{\frac{1}{2m+n}} \right)$.

**Corollary 1** Under Assumption 2, let $\lambda_i$ be the $i$th eigenvalue of $-\Delta$. Fixing $K \in \mathbb{N}, \gamma$ we let $\Gamma_K \equiv \min_{1 \leq i \leq K} \text{dist}(\lambda_i, \sigma(-\Delta) \setminus \{ \lambda_i \})$. If $\varepsilon$ is small enough and $m + n$ is large enough so that

$$\left( \frac{\log(m+n)}{m+n} \right)^{\frac{1}{2m+n}} \leq \varepsilon,$$

then for any $t$ less than the diameter of $M$, with probability greater than $1 - (m + n)^{-2},$

$$\max_{x_i, x_j \in \mathcal{X}} |\tilde{H}_{\varepsilon,K,i}(i,j) - C_{\varepsilon,K}(x_i, x_j, t)| \leq \mathcal{K} \left( \frac{\log(m+n)}{m+n} \right)^{\frac{1}{2m+n}},$$

where $\mathcal{K}$ depends on $K$, $\Gamma_K$, $\lambda_K$, $d$, $p_m$, the $C^2$ norm of $p$ and the diameter, volume, injectivity radius, curvature and second fundamental form of the manifold.

Remark 6 The convergence rates in the above theorem and corollary are not optimal. We expect that the optimal convergence rate is better than what we have reported. Finding the optimal convergence rate will be explored in our future work. Even if the optimal convergence rate, and hence the asymptotic relation between $\varepsilon$ and $n + m$ is known, this relationship does not provide a practical approach for choosing $\varepsilon$ due to the unknown constant.

Next, we come back to the Nyström extension. Suppose we have $\ell$ additional i.i.d. samples, $\{x_{m+n+1}, \ldots, x_{m+n+\ell}\}$, from density $P$. Let $\mathcal{X}^∗ = \{ x_1, \ldots, x_{m+n}, x_{m+n+1}, \ldots, x_{m+n+\ell} \}$. Let $E \in \mathbb{R}^{(m+n+\ell) \times (m+n)}$ be the extension matrix defined in Equation (9). Similar to Equation (10), the Nyström extension of the GL-GP covariance matrix over $\mathcal{X}^*$ on the manifold is defined as

$$\tilde{H}_{\varepsilon,K,i}^* = E \left( \sum_{l=0}^{K-1} \frac{e^{-\mu_{i,m+n+\ell} t}}{1 - \varepsilon^2 \mu_{i,m+n+\ell}} v_{i,m+n,\ell} v_{i,m+n,\ell}^T \right) E^T \in \mathbb{R}^{(m+n+\ell) \times (m+n+\ell)}.$$  

The following theorem shows that $\tilde{H}_{\varepsilon,K,i}^*$ is an approximation to $C_{\varepsilon,K}(x,x', t)$ over $\mathcal{X}^*$.

**Theorem 3** Under Assumption 2, suppose we sample $\ell$ additional times i.i.d. from density $P$ to obtain $\{x_{m+n+1}, \ldots, x_{m+n+\ell}\}$. Let $\mathcal{X}^* = \{ x_1, \ldots, x_{m+n}, x_{m+n+1}, \ldots, x_{m+n+\ell} \}$.
Let $\tilde{\mathbf{H}}_{\varepsilon,K,t}$ and $\tilde{\mathbf{H}}_{\varepsilon,K,t}^*$ be defined in Equations (15) and (17) respectively. Then $\tilde{\mathbf{H}}_{\varepsilon,K,t}(i,j) = \tilde{\mathbf{H}}_{\varepsilon,K,t}^*(i,j)$, for $1 \leq i,j \leq m+n$.

Let $\lambda_i$ be the $i$th eigenvalue of $-\Delta$. Fixing $K \in \mathbb{N}$, we let $\Gamma_K := \min_{1 \leq i \leq K} \text{dist}(\lambda_i, \sigma(-\Delta) \setminus \{\lambda_i\})$. If $\varepsilon$ is small enough and $\varepsilon$ satisfies Equation (16), and $m+n$ satisfies
\[
\left(\frac{\log(m+n)}{m+n}\right)^{\frac{1}{4+\varepsilon}} \leq \varepsilon, \text{ then for any } t \text{ less than the diameter of } M, \text{ with probability greater than } 1 - (m+n)^{-2},
\]
\[
\max_{x_i,x_j \in \mathcal{X}} |\tilde{\mathbf{H}}_{\varepsilon,K,t}(i,j) - \mathbf{C}_{\varepsilon,K}(x_i,x_j,t)| \leq \mathcal{K}_4 K^2 \varepsilon^{1/2},
\]
where $\mathcal{K}_4$ is the same constant defined in Theorem 2.

The proof is similar to the proof of Theorem 2 and is sketched in Section C of the Supplementary Materials. Part (a) says that $\tilde{\mathbf{H}}_{\varepsilon,K,t}$ is an extension of $\mathbf{H}_{\varepsilon,K,t}$ to $\mathcal{X}^*$. In the trivial case when there is no additional sample points, that is, $\ell = 0$, we have $\tilde{\mathbf{H}}_{\varepsilon,K,t}^* = \mathbf{H}_{\varepsilon,K,t}$. Although we have more samples to be used to construct a covariance matrix, the eigenvectors we use are an extension of the eigenvectors of the GL constructed from $\mathcal{X}$ while the eigenvalues remain the same. Hence, part (b) says the accuracy of the approximation of $\mathbf{C}_{\varepsilon,K}(x,x',t)$ by $\tilde{\mathbf{H}}_{\varepsilon,K,t}$ over $\mathcal{X}^*$ is on the same level as the approximation of $\mathbf{C}_{\varepsilon,K}(x',x',t)$ by $\mathbf{H}_{\varepsilon,K,t}$ over $\mathcal{X}$. If we treat $K$ and $\lambda_K$ as constants and focus on the relation between $\varepsilon$ and $m+n$, then by using the same argument as Corollary 1, the error between $\tilde{\mathbf{H}}_{\varepsilon,K,t}(i,j)$ and $\mathbf{C}_{\varepsilon,K}(x_i,x_j,t)$ is $O\left(\left(\frac{\log(m+n)}{m+n}\right)^{\frac{1}{4+\varepsilon}}\right)$.

### 4.4 The predictive error of the covariance matrix

Suppose $\tilde{\mathbf{H}}_{\varepsilon,t}^K$ is the GL-GP covariance matrix by the GL over $\{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}\}$. If we rewrite $\tilde{\mathbf{H}}_{\varepsilon,t}^K$ as
\[
\begin{bmatrix}
H_1 & H_2 \\
H_3 & H_4
\end{bmatrix},
\]
where $H_1$ is an $m \times m$ matrix, then the prediction at $\{x_{m+1}, \ldots, x_{m+n}\}$ using the GL-GP covariance matrix is $\mathbf{f}_e := H_3 (H_1 + \sigma^2_{\text{noise}} I)^{-1} y$. We define an $(m+n) \times (m+n)$ matrix $\Sigma$ so that
\[
\Sigma_{ij} = \mathbf{C}_{\varepsilon}^K (x_i, x_j, t).
\]
Let $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, where $\Sigma_{11}$ is an $m \times m$ symmetric matrix.

The predictive at $\{x_{m+1}, \ldots, x_{m+n}\}$ using the exact GL-GP covariance function is then $\mathbf{f}_{\text{GLGP}} := \Sigma_{21} (\Sigma_{11} + \sigma^2_{\text{noise}} I)^{-1} y$. The following theorem describes the difference between the predictions of the GP under the matrix $\tilde{\mathbf{H}}_{\varepsilon,t}^K$ and the exact GL-GP covariance function. The proof of the theorem is in Section D of the Supplementary Materials.

**Theorem 4** Suppose $\|\tilde{\mathbf{H}}_{\varepsilon,t}^K - \Sigma\|_{\max} \leq \delta < 1$. Then we have
\[
\|\mathbf{f}_e - \mathbf{f}_{\text{GLGP}}\|_{\max} \leq \frac{m\|y\|_{\max}(c + 3\lambda)\|\Sigma_{21}\|_{\max} + \lambda)}{\lambda(\lambda + c\delta)},
\]
where $c$ is a constant depending on $\frac{1}{\delta}(\tilde{\mathbf{H}}_{\varepsilon,t}^K - \Sigma)$ and $\lambda$ is the smallest eigenvalue of $\Sigma_{11} + \sigma^2_{\text{noise}} I$.

If $\Sigma$ is a discretization of the covariance function over $\{x_1, \ldots, x_{m+n}, x_{m+n+1}, \ldots, x_{m+n+\ell}\}$ and we divide $\Sigma$ so that $\Sigma_{11} \in \mathbb{R}^{m \times m}$ and $\Sigma_{21} \in \mathbb{R}^{(n+\ell) \times m}$, then the above theorem also holds for the Nyström extension of $\tilde{\mathbf{H}}_{\varepsilon,K,t}$. The result of Theorem 4 can be combined with Theorem 2 or Theorem 3.
to estimate the error between the prediction of the GL-GP covariance matrix or its Nyström extension and the prediction of the GL-GP covariance function.

4.5 Measurement error

In this subsection, we discuss the stability of GL-GP when there are measurement errors, so that the data do not fall exactly on the manifold.

Assumption 3 In addition to Assumption 2, due to measurement error, we assume that the data we observe are \( \{ x_1', \ldots, x_{m+n}' \} \subset \mathbb{R}^D \) such that \( \| \iota(x_i) - x_i' \|_0 < \delta \), for \( i = 1, \ldots, m+n \).

Denote \( L' \) to be the GL associated with \( \{ x_1', \ldots, x_{m+n}' \} \). The following theorem shows that if the measurement errors are not too large, then one can still control the eigenvalues and eigenvectors of \( L' \) by those of \( L \). The proof of the theorem is in Section E of the Supplementary Materials.

Theorem 5 Suppose Assumption 3 holds. For any \( \epsilon \) small enough, if \( \delta < C_{\epsilon} d+1 \) and \( m+n \) is large enough so that \( \left( \frac{\log(m+n)}{m+n} \right)^{1/2} \leq \epsilon \), then with probability greater than \( 1 - (m+n)^{-2} \),

\[
\| L - L' \|_2 \leq C' \delta,
\]

where \( \| \cdot \|_2 \) is the spectral norm of a matrix, and \( C \) and \( C' \) are constants depending on \( \rho_m \) and the \( C^0 \) norm of \( p \).

It will be interesting to relax Assumption 3 and allow unbounded measurement error in future work. We refer the reader to El Karoui and Wu (2016) for more discussion and relevant results.

Since the GL-GP covariance matrix is constructed by using the eigenpairs of the GL, its stability follows from the above stability theorem of the GL.

4.6 Posterior contraction rate of GL-GP on manifold under the fixed design setup

Let \( M \) be a \( d \)-dimensional smooth, closed and connected Riemannian manifold with diameter bounded by 1. Assume \( M \) is isometrically embedded in \( \mathbb{R}^D \) through \( \iota : M \to \mathbb{R}^D \). Let \( \lambda_0, \epsilon \leq \lambda_1, \epsilon \leq \cdots \) be the eigenvalues of the operator \( I_{\iota} \). Suppose \( \varphi_{i, \epsilon} \) is the \( i \)th eigenfunction of \( I_{\iota} \) and each \( \varphi_{i, \epsilon} \) is normalized in \( L^2(M) \). We define the following finite dimensional inner product space \( H_{\iota, K, t} \).

Definition 3 For fixed \( \epsilon, K \) and \( t \geq 0 \), we define \( H_{\iota, K, t} \) to be an inner product space on \( M \) such that

\[
H_{\iota, K, t} = \left\{ h(x) = \sum_{i=0}^{K-1} a_i e^{-\lambda_{i, \epsilon} t} \varphi_{i, \epsilon}(x), \sum_{i=0}^{K-1} a_i^2 \right\},
\]

with the inner product

\[
\left\langle \sum_{i=0}^{K-1} a_i e^{-\lambda_{i, \epsilon} t} \varphi_{i, \epsilon}(x), \sum_{i=0}^{K-1} b_i e^{-\lambda_{i, \epsilon} t} \varphi_{i, \epsilon}(x) \right\rangle_{H_{\iota, K, t}} = \sum_{i=0}^{K-1} a_i b_i.
\]

Denote \( E_{i, \epsilon, \iota} := e^{-\lambda_{i, \epsilon} \frac{t}{2}} \varphi_{i, \epsilon} \). Then, \( \{ E_{i, \epsilon, \iota} \}_{i=0}^{K-1} \) is an orthonormal basis of \( H_{\iota, K, t} \). Denote \( B_R^{H_{\iota, K, t}} \) to be the ball of radius \( R \) in \( H_{\iota, K, t} \). Also, let \( H_{\iota, K} := \bigcup_{0 \leq t \leq 1} H_{\iota, K, t} \).
Refer to Section F of the Supplementary Materials for basic properties of the spaces $\mathbb{H}_{\varepsilon,K}^s$ and $\mathbb{H}_{t,K}^s$. We explain the reason that we focus our discussion on the space $\mathbb{H}_{t,K}^s$. Any function in $L^\infty(M)$ with certain regularity, for example, a function in the Besov space, can be approximated by the functions in $\cup_x \cup_K \mathbb{H}_{t,K}^s$. The Besov space $B_{s,\infty}^s$ is a subspace of $L^\infty(M)$ with some regularity quantitatively characterized by the parameter $s$; refer to Definition 2 in the Supplementary Materials. We state our result as the following approximation proposition. The proposition implies that $\cup_x \cup_K \mathbb{H}_{t,K}^s$ is a dense subset of $B_{s,\infty}^s$.

**Proposition 4** Suppose $f_0 \in B_{s,\infty}^s$ with $s > 0$ and $\|f_0\|_\infty \leq 1$. For any $\gamma > 0$ small enough, if $K = \lfloor D_1\gamma^{-\frac{d}{4}} \rfloor$, $\varepsilon$ satisfies Equation (16) and $\varepsilon \leq D_2\gamma^{-\frac{d}{4}}$, then there is a function $h \in \mathbb{H}_{t,K}^s$ such that $\|h - f_0\|_\infty \leq \gamma$ and $\|h\|_{\mathbb{H}_{t,K}^s} \leq \nu(M) \exp(D_3\gamma^{-\frac{d}{2}}t)$. $D_1$ is a constant depending on $s$, $d$, the diameter and Ricci curvature of the manifold, $D_2$ depends on $s$, $d$, $\rho$ and the $C^s$ norm of $\rho$, and the volume, injectivity radius, curvature and second fundamental form of the manifold and $D_3$ depends on $s$, $d$, the diameter, volume and Ricci curvature of the manifold.

Through the above Proposition, we construct a stratification indexed by $\varepsilon$ and $K$, a dense subset of $B_{s,\infty}^s$. We associate each layer $\mathbb{H}_{t,K}^s$ in the stratification with a GP. Suppose \{$Z_i$\} are independent random variables with mean 0 and variance 1. Fix $\varepsilon > 0$ and $K \in \mathbb{N}$. We define a Gaussian process for $t \geq 0$ indexed by $x \in M$,

$$W_{t,K}^t(x) := \sum_{i=0}^{K-1} Z_i E_{i,t}(x) = \sum_{i=0}^{K-1} e^{-\lambda_i t} Z_i \varphi_{i,t}(x).$$

(18)

By a straightforward calculation, the covariance function of $W_{t,K}^t(x)$, $C_{t,K}(x,x',t)$, satisfies

$$C_{t,K}(x,x',t) = \mathbb{E}[W_{t,K}^t(x)W_{t,K}^t(x')] = \sum_{i=0}^{K-1} e^{-\lambda_i t} \varphi_{i,t}(x) \varphi_{i,t}(x').$$

Suppose $T$ is a random variable with probability density function $g$ on $[0, 1]$, with

$$\mathbb{E}_1 t^{-p} e^{-t^{-q}} \leq g(t) \leq \mathbb{E}_2 t^{-p} e^{-t^{-q}},$$

(19)

for constants $0 < \mathbb{E}_1 \leq \mathbb{E}_2$ and $p, q > 0$. These bounds are satisfied when $T$ follows an inverse transformed gamma density, induced by raising a gamma random variable to the power $-1/q$; in Theorem 6, $q$ is shown to relate to the dimension of the manifold. Using $g$ as a prior for $T = t$ in the GL-GP corresponding to $W_{t,K}^T(x)$ leads to a prior $\Pi$ on $\mathbb{H}_{t,K}^s$.

**Assumption 4** Let $M$ be a $d$-dimensional smooth, closed and connected Riemannian manifold isometrically embedded in $\mathbb{R}^D$ through $t : M \rightarrow \mathbb{R}^D$. We assume that the diameter of $M$ is bounded by 1. We consider the regression model (1) with \{${x_1, \ldots, x_n}$\} $\subset M$. We propose the prior for the regression function $f$ as

$$W_{t,K}^T|T \sim GP(0, C_{t,K}(x,y, T)), \quad T \sim g(t),$$

where $g(t)$ is a density function on $[0, 1]$ satisfying Equation (19).

**Remark 7** The assumption that the diameter of $M$ is bounded by 1 is for notation simplicity and the below result holds without this assumption.
Following Ghosal et al. (2000); Ghosal and van der Vaart (2007), given a true function $f_0 \in \mathbb{H}_{\varepsilon,K}$ and a data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we say that the posterior contraction rate of the GP prior in the fixed design is at least $\gamma_n$ if

$$\Pi \left( \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f_0(x_i)) \geq \gamma_n \mid \{(x_1, y_1), \ldots, (x_n, y_n)\} \right) \to 0$$

when $n \to \infty$. In the fixed design case, $\{x_1, \ldots, x_n\}$ are deterministic rather than sampled based on some p.d.f on $M$.

**Remark 8** The random variable $T$ can be regarded as a bandwidth parameter. Our bounds on the probability density function $g$ in Equation (19) are motivated by van der Vaart and van Zanten (2009) and Bhattacharya et al. (2014). In showing minimax rates in the fixed design setup, they choose inverse transformed gamma priors for bandwidth parameters in squared exponential GPs, with the power equal to the Euclidean space dimension.

We have the following theorem about the posterior contraction rate.

**Theorem 6** Fix $K \in \mathbb{N}$ and fix any $\varepsilon$ small enough so that Equation (16) holds. Under Assumption 4, suppose $f_0 \in \mathbb{H}_{\varepsilon,K}$ with $\|f_0\|_{\infty} \leq 1$ and $p \geq 1, q \geq \frac{d}{2}$. If we choose $\gamma_n = \left( \frac{n}{\log n} \right)^{-\frac{d}{2p+2d}}$ with $\gamma_n \leq \frac{1}{K}$, then the posterior contraction rate of the GP prior in the fixed design is at least $\gamma_n$.

**Remark 9** For notation simplicity, we assume $\|f_0\|_{\infty} \leq 1$. A similar result holds with the coefficients in the relation between $\gamma$ and $n$ depending on $\|f_0\|_{\infty}$.

## 5 DISCUSSION

In this article, we were motivated by the problem of nonparametric regression with predictors on an unknown subset $S$ of $\mathbb{R}^D$ that might have a complicated geometric and topological structure. The proposed GL-GP approach is appealing in allowing the GP covariance to reflect the intrinsic geometry of $S$. Although we have taken substantial first steps theoretically, while showing promising empirical results in illustrative examples, there are multiple areas for future research.

A first direction relates to developing scalable implementations of the GL-GP. There is a rich literature developing scalable GP algorithms in other contexts, such as for huge spatial and/or temporal datasets modelled via GPs with traditional squared exponential or Matérn covariance functions; for example, refer to Datta et al. (2016a, b), Peruzzi et al. (2020) and Ambikasaran et al. (2015). It is not straightforward to extend such algorithms to our context. One possibility is to adapt subset-of-regressor (Rasmussen, 2003) and predictive process (Banerjee et al., 2008) algorithms via the Nyström extension idea of Section 2.3, or the Roseland algorithm (Shen & Wu, 2020) that could be viewed as a generalization of Nyström via diffusion.

Another promising direction is to consider broader classes of GL-GPs by using more flexible kernels in constructing the GL; for example, a Matérn kernel could be used in place of the Gaussian or the Gaussian kernel could be modified to have bandwidth parameters for each dimension. It is interesting to consider the theoretical and practical behavior of such kernels, and the induced smoothness and covariance behavior for a variety of geometric structures. The Matérn case may be of particular relevance in applications to spatial statistics, in which our approach may provide
a competitor to the current literature on spatial barriers and the so-called coastline problem; refer, for example to Bakka et al. (2019).

Finally, there are several natural next steps theoretically. One general direction is to build on our rate results, attempting to obtain the optimal rate in the manifold case. Another is to study consistency in estimating the covariance parameters, a problem of particular relevance in spatial statistics; refer, for example, to Li (2020) and Tang et al. (2021). We would also like to obtain an improved understanding for much broader classes of $S$, including extensions beyond manifolds to stratified spaces and various metric spaces.

6 | CODE AND SUPPLEMENTARY MATERIALS

The MATLAB implementation of the examples in Sections 3 are available on the GitHub repository https://github.com/wunan3/Diffusion-based-Gaussian-Process. The proofs of the main theorems are in the Supplementary Materials.

ACKNOWLEDGEMENTS

The authors thank the associate editor and the referees for their comments to improve the quality of the paper. David B Dunson and Nan Wu acknowledge the support from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 856506).

ORCID

Hau-Tieng Wu https://orcid.org/0000-0002-0253-3156
Nan Wu https://orcid.org/0000-0001-6064-6370

REFERENCES

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D.W. & O’Neil, M. (2015) Fast direct methods for Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 252–265.
Bakka, H., Vanhatalo, J., Illian, J.B., Simpson, D. & Rue, H. (2019) Non-stationary Gaussian models with physical barriers. Spatial Statistics, 29, 268–288.
Banerjee, S., Gelfand, A.E., Finley, A.O. & Sang, H. (2008) Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70, 825–848.
Belkin, M., Niyogi, P. & Sindhwani, V. (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7, 2399–2434.
Bhattacharya, A., Pati, D. & Dunson, D. (2014) Anisotropic function estimation using multi-bandwidth Gaussian processes. The Annals of Statistics, 42, 352–381.
Borovitskiy, V., Terenin, A., Mostowsky, P. & Deisenroth, M.P. (2020) Matérn Gaussian processes on Riemannian manifolds. Advances in Neural Information Processing Systems, 33, 12426–12437.
Castillo, I., Kerkyacharian, G. & Picard, D. (2014) Thomas Bayes’ walk on manifolds. Probability Theory and Related Fields, 158, 665–710.
Chang, C.-H., Huang, H.-C. & Ing, C.-K. (2017) Mixed domain asymptotics for a stochastic process model with time trend and measurement error. Bernoulli, 23, 159–190.
Cheng, M.-Y. & Wu, H.-T. (2013) Local linear regression on manifolds and its geometric interpretation. Journal of the American Statistical Association, 108, 1421–1434.
Chung, F. (1996) Spectral graph theory. Providence: American Mathematical Society.
Coifman, R.R. & Lafon, S. (2006) Diffusion maps. Applied and Computational Harmonic Analysis, 21, 5–30.
Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. (2016a) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association, 111, 800–812.
Datta, A., Banerjee, S., Finley, A.O., Hamm, N.A. & Schaap, M. (2016b) Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. The Annals of Applied Statistics, 10, 1286–1316.

Dunlop, M.M., Slepčev, D., Stuart, A.M. & Thorpe, M. (2020) Large data and zero noise limits of graph-based semi-supervised learning algorithms. Applied and Computational Harmonic Analysis, 49, 655–697.

Dunson, D.B., Wu, H.-T. & Wu, N. (2021) Spectral convergence of graph Laplacian and Heat kernel reconstruction in $L^\infty$ from random samples. Applied and Computational Harmonic Analysis, 55, 282–336.

El Karoui, N. & Wu, H.-T. (2016) Graph connection Laplacian methods can be made robust to noise. The Annals of Statistics, 44, 346–372.

Ghosal, S. & van der Vaart, A. (2007) Convergence rates of posterior distributions for noniid observations. The Annals of Statistics, 35, 192–223.

Ghosal, S., Ghosh, J.K. & van der Vaart, A.W. (2000) Convergence rates of posterior distributions. The Annals of Statistics, 28, 500–531.

Li, C. (2020) Bayesian fixed-domain asymptotics: Bernstein-von Mises theorem for covariance parameters in a Gaussian process model. arXiv preprint arXiv:2010.02126.

Li, D., Tang, W. & Banerjee, S. (2021) Fixed-domain inference for Gaussian processes with Matérn covariogram on compact Riemannian manifolds. arXiv preprint arXiv:2104.03529.

Lin, L., Mu, N., Cheung, P. & Dunson, D. (2019) Extrinsic Gaussian processes for regression and classification on manifolds. Bayesian Analysis, 14, 887–906.

Nadler, B., Srebro, N. & Zhou, X. (2009) Semi-supervised learning with the graph Laplacian: the limit of infinite unlabelled data. Advances in Neural Information Processing Systems, 22, 1330–1338.

Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N. & Dunson, D. (2019) Intrinsic Gaussian processes on complex constrained domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81, 603–627.

Peruzzi, M., Banerjee, S. & Finley, A.O. (2020) Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. Journal of the American Statistical Association, 0, 1–14.

Rasmussen, C. E. (2003) Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U. & Ratsch, G. (Eds.) Summer school on machine learning. Berlin, Heidelberg: Springer, pp. 63–71.

Romero, J., Tzionas, D. & Black, M.J. (2017) Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG), 36, 1–17.

Shen, C. & Wu, H.-T. (2020) Scalability and robustness of spectral embedding: landmark diffusion is all you need. arXiv preprint arXiv:2001.00801.

Tang, W., Zhang, L. & Banerjee, S. (2021) On identifiability and consistency of the nugget in Gaussian spatial process models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). Eprint.

van der Vaart, A.W. & van Zanten, J.H. (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. The Annals of Statistics, 37, 2655–2675.

Wang, X. & Lerman, G. (2015) Nonparametric Bayesian regression on manifolds via Brownian motion. arXiv preprint arXiv:1507.06710.

Zhu, X.J. (2005) Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Zhu, X., Ghahramani, Z. & Lafferty, J.D. (2003) Semi-supervised learning using Gaussian fields and harmonic functions. In Proceedings of the 20th international conference on machine learning (ICML-03), pp. 912–919.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Dunson, D.B., Wu, H.-T. & Wu, N. (2022) Graph based Gaussian processes on restricted domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 84(2), 414–439. Available from: https://doi.org/10.1111/rssb.12486