Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Equivariant Projected Kernels

Michael Hutchinson[*1] **Alexander Terenin**[*2, 3] **Viacheslav Borovitskiy**[*4]
So Takao[*5] **Yee Whye Teh**[1] **Marc Peter Deisenroth**[5]

[1]University of Oxford [2]University of Cambridge [3]Imperial College London
[4]St. Petersburg State University [5]Centre for Artificial Intelligence, University College London

Abstract

Gaussian processes are machine learning models capable of learning unknown functions in a way that represents uncertainty, thereby facilitating construction of optimal decision-making systems. Motivated by a desire to deploy Gaussian processes in novel areas of science, a rapidly-growing line of research has focused on constructively extending these models to handle non-Euclidean domains, including Riemannian manifolds, such as spheres and tori. We propose techniques that generalize this class to model vector fields on Riemannian manifolds, which are important in a number of application areas in the physical sciences. To do so, we present a general recipe for constructing gauge equivariant kernels, which induce Gaussian vector fields, i.e. vector-valued Gaussian processes coherent with geometry, from scalar-valued Riemannian kernels. We extend standard Gaussian process training methods, such as variational inference, to this setting. This enables vector-valued Gaussian processes on Riemannian manifolds to be trained using standard methods and makes them accessible to machine learning practitioners.

1 Introduction

Gaussian processes are an effective model class for learning unknown functions. They are particularly attractive for use within data-efficient decision systems, including Bayesian optimization [3, 32, 39], model-based reinforcement learning [34, 7], and active learning [21]. In these settings, Gaussian processes can represent and propagate uncertainty, as well as encode inductive biases as prior information in order to drive data efficiency. A key aspect of prior information is the geometry of the domain on which the Gaussian process is defined, which often encodes key properties, such as symmetry. Following the growing deployment of Gaussian processes, a number of recent works have focused on how to define Gaussian processes on non-Euclidean domains in ways that reflect their geometric structure [2, 1].

In many applications, such as climate science, quantities of interest are vector-valued. For example, global wind velocity modeling must take into account both speed and direction, and is represented by a vector field. On geometric domains, the mathematical properties of vector fields can differ noticeably from their Euclidean counterparts: for instance, one can prove that every smooth vector field on a sphere must vanish in at least one point [26]. Behavior such as this simultaneously highlights the need to represent geometry correctly when modeling vector-valued data, and presents a number of non-trivial technical challenges in constructing models that are mathematically sound.

*Equal contribution. Code at https://github.com/MJHutchinson/ExtrinsicGaugeEquivariantVectorGPs. For a general implementation, see https://github.com/GPflow/GeometricKernels/.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In particular, even the classical definition of a vector-valued Gaussian process—that is, a random function with multivariate Gaussian marginals at any finite set of points—already fails to be a fully satisfactory notion when considering smooth vector fields on a sphere. This is because tangent vectors at distinct points live within different tangent spaces, and it is not clear how to construct a cross-covariance between them that does not depend on a completely arbitrary choice of basis vectors within each space. Constructions that are independent of this choice of basis are called *gauge equivariant*, and a large body of recent work [5, 15] in geometric machine learning has focused on satisfying this key property for convolutional neural networks that deal with non-Euclidean data.

Our contributions include the following. We (a) present a differential-geometric formalism for defining Gaussian vector fields on manifolds in a coordinate-free way, suitable for Gaussian process practitioners with minimal familiarity with differential geometry, (b) present a universal and fully constructive technique for defining prior Gaussian vector fields on Riemannian manifolds, which we term the *projected kernel* construction, and (c) discuss how to adapt key components in the computational Gaussian process toolkit, such as inducing point methods, to the vector field setting.

The structure of the paper is as follows. In Section 2, we define vector-valued Gaussian processes on smooth manifolds. We start by reviewing the multi-output Gaussian process set-up, which is typically used in machine learning. We then detail a differential-geometric formalism for defining vector-valued Gaussian processes on smooth manifolds. In Section 3, we provide a concrete construction for these Gaussian processes on Riemannian manifolds and discuss how they can be trained using variational sparse approximations. Section 4 showcases Gaussian vector fields on two tasks, namely weather imputation from satellite observations and learning the dynamics of a mechanical system.

2 Vector-valued Gaussian Processes on Smooth Manifolds

A vector-valued Gaussian process (GP) is a random function $\boldsymbol{f} : X \to \mathbb{R}^d$ such that, for any finite set of points $\boldsymbol{x} \in X^n$, the random variable $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{n \times d}$ is jointly Gaussian. Every such GP is characterized by its mean function $\boldsymbol{\mu} : X \to \mathbb{R}^d$ and matrix-valued covariance kernel $k : X \times X \to \mathbb{R}^{d \times d}$, which is a positive-definite function in the matrix sense. These functions satisfy $\mathbb{E}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{\mu}(\boldsymbol{x})$ and $\mathrm{Cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')) = k(\boldsymbol{x}, \boldsymbol{x}')$ for any $\boldsymbol{x}, \boldsymbol{x}' \in X$. Here, dependence between function values is encoded in the kernel's variability along its input domain, and correlations between different dimensions of the vector-valued output are encoded in the matrix that the kernel outputs.

Consider a function $\boldsymbol{f}$ with $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and training data $(\boldsymbol{x}, \boldsymbol{y})$. Placing a GP prior $\boldsymbol{f} \sim \mathrm{GP}(0, k)$ on the unknown function results in a GP posterior, whose mean and covariance are given by

$$\mathbb{E}(\boldsymbol{f} \mid \boldsymbol{y}) = \mathbf{K}_{(\cdot)\boldsymbol{x}}(\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} + \sigma^2 \mathbf{I})^{-1} \boldsymbol{y} \quad \mathrm{Cov}(\boldsymbol{f} \mid \boldsymbol{y}) = \mathbf{K}_{(\cdot,\cdot)} - \mathbf{K}_{(\cdot)\boldsymbol{x}}(\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_{\boldsymbol{x}(\cdot)}. \tag{1}$$

Here, $(\cdot)$ denotes an arbitrary set of test locations, $\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} = k(\boldsymbol{x}, \boldsymbol{x})$ is the kernel matrix, and $\mathbf{K}_{(\cdot)\boldsymbol{x}} = k(\boldsymbol{x}, \cdot)$ is the cross-covariance matrix between function values evaluated at the training and test inputs. The GP posterior can also be written as

$$(\boldsymbol{f} \mid \boldsymbol{y})(\cdot) = \boldsymbol{f}(\cdot) + \mathbf{K}_{(\cdot)\boldsymbol{x}}(\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} + \sigma^2 \mathbf{I})^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{2}$$

where $\boldsymbol{f}(\cdot)$ is the prior GP, and equality holds in distribution [46, 47]. These expressions form the foundation upon which Gaussian-process-based methods in machine learning are built.

Recent works have studied techniques for working with the expressions (1) and (2) when the input domain X is a Riemannian manifold, focusing both on defining general classes of kernels [2], and on efficient computational techniques [46, 47]. In this setting, namely for $f : X \to \mathbb{R}$, defining kernels already presents technical challenges: the seemingly-obvious first choice one might consider, namely the geodesic squared exponential kernel, is ill-defined in general [13]. We build on these recent developments to model vector fields on manifolds using GPs. We do not consider manifold-valued generalizations of Gaussian processes, for instance $f : \mathbb{R} \to X$: various constructions in this setting are instead studied by Stroock [40], Émery [12], Mallasto and Feragen [29], and Mallasto et al. [30]. To begin, we review what a vector field on a manifold actually is.

2.1 Vector Fields on Manifolds

Let X be a d-dimensional smooth manifold with $T_x X$ denoting its tangent space at x. Let $TX = \{(x, v) \mid x \in X, v \in T_x X\}$ be its *tangent bundle*, and let $T^*X = \{(x, \phi) \mid x \in X, \phi \in T_x^* X\}$

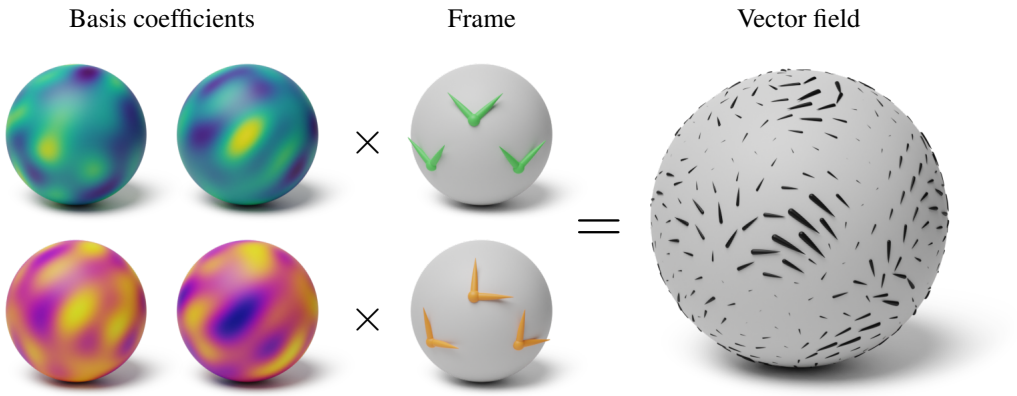

Figure 1: Illustration on $\mathbb{S}^2$. Here we illustrate two possible bases (also called frames), consisting of green and orange basis vectors (center), that can be chosen locally on the manifold $\mathbb{S}^2$. The vector field on $\mathbb{S}^2$ (right) can be produced by taking two scalar fields (left) in each respective color, and combining them with the basis vectors (center) to form the vector field.

be its *cotangent bundle*—endow both spaces with the structure of smooth manifolds. Define the projection map $\mathrm{proj}_X : TX \to X$ by $\mathrm{proj}_X(x, v) = x$. A *vector field* on X is a map that assigns each point in X to a tangent vector attached to that point. More formally, a vector field is a *cross-section*, or simply a *section*, of the tangent bundle, which is a map $f : X \to TX$, such that $\mathrm{proj}_X f(x) = x$ for all x.[1] A vector field is called *smooth* if this map f is smooth.

To represent a vector field on a manifold numerically, one must choose a basis in each tangent space, which serves as a coordinate system for vectors in the tangent space. On many manifolds it is *impossible* to choose these basis vectors in a way that they vary smoothly in space.[2] This can be handled by working with local coordinates, or with bases that are non-smooth. Any chosen set of basis vectors is arbitrary, so objects constructed using them should not depend on this choice. Constructions that satisfy this property are called *gauge equivariant*. This notion is illustrated in Figure 1, and will play a key role in the sequel.

2.2 Gaussian Vector Fields

Upon reflecting on the above considerations in the context of GPs, the first issue one encounters is that, for a random vector field $f : X \to TX$, it is not clear what it means for finite-dimensional marginal distributions to be *multivariate Gaussian* given that f takes its values in a bundle rather than a vector space. The first step towards constructing Gaussian vector fields, therefore, involves adapting the notion of *finite-dimensional marginal distributions* appropriately.

Definition 1. *Let X be a smooth manifold. We say that a random vector field f is* Gaussian *if for any finite set of locations $x_1, \ldots, x_n \in X$, the random vector $(f(x_1), \ldots, f(x_n)) \in T_{x_1}X \oplus \ldots \oplus T_{x_n}X$ is Gaussian (either in the sense of duality or in any basis: see Appendix A for details).*

This definition is near-identical to the Euclidean case: the only difference is that finite-dimensional marginals are now supported in a direct sum of tangent spaces, instead of $\mathbb{R}^{n \times d}$. With this definition, the standard multi-output GP properties, such as conditioning, carry over, virtually unmodified. Definition 1 is a natural choice: if we embed our manifold into Euclidean space, the induced GP is a vector-valued GP as defined in the beginning of Section 2.

[1] A vector field is *not the same as* a map $\tilde{f} : X \to \mathbb{R}^d$: an output value $f(x) \in TX$ formally consists of both a copy of the input point x, and vector within the tangent space T_xX at this point. This encodes the geometric structure of the underlying manifold. The algebraic requirement $\mathrm{proj}_X f(x) = x$ for all x ensures that the tangent vector chosen correctly corresponds to the point at which it is attached.

[2] If a smooth choice of basis vectors existed, it would define a smooth non-vanishing vector field. On the sphere, by the *hairy ball theorem*, all smooth vector fields vanish in at least one point, so no such bases exist.

Proposition 2. *Let X be a manifold and* $\mathrm{emb} : X \to \mathbb{R}^p$ *be a smooth embedding. Let f be a Gaussian vector field (as defined in Definition 1), and let $f_{\mathrm{emb}} : \mathrm{emb}(X) \to \mathbb{R}^p$ be its pushforward along the embedding. Then f_{emb} is a vector-valued Gaussian process in the Euclidean sense.*

All proofs in this work can be found in Appendix A. Having established the notion of a vector-valued Gaussian process on a smooth manifold, we proceed to deduce what mathematical objects play the role of a mean function and kernel, so that it is clear what ingredients are needed to construct and determine such a process.

The former is clear: the mean of a Gaussian vector field should be an ordinary vector field, and will determine the mean vector at all finite-dimensional marginals. The kernel, on the other hand, is less obvious: because distinct tangent vectors live in different tanget spaces, it is unclear whether or not a Gaussian vector field is characterized by an appropriate notion of a matrix-valued kernel, or by something else. Now, we define the right notion for kernel in this setting.

Definition 3. *We call a symmetric function $k : T^*X \times T^*X \to \mathbb{R}$* fiberwise bilinear *if for all pairs of points $x, x' \in X$*

$$k(\lambda\alpha_x + \mu\beta_x, \gamma_{x'}) = \lambda k(\alpha_x, \gamma_{x'}) + \mu k(\beta_x, \gamma_{x'}) \tag{3}$$

holds for any $\alpha_x, \beta_x \in T^*_x X$, $\gamma_{x'} \in T^*_{x'} X$ and $\lambda, \mu \in \mathbb{R}$, and positive semi-definite *if for any set of covectors $\alpha_{x_1}, \ldots, \alpha_{x_n} \in T^*X$, we have $\sum_{i=1}^n \sum_{j=1}^n k(\alpha_{x_i}, \alpha_{x_j}) \geq 0$. We call a symmetric fiberwise bilinear positive semi-definite function a* cross-covariance kernel.

This coordinate-free function should be viewed as analogous to $((x, \boldsymbol{v}), (x', \boldsymbol{v}')) \mapsto \boldsymbol{v}^T \mathbf{K}_{x,x'} \boldsymbol{v}'$ in the Euclidean setting, where $\boldsymbol{v}, \boldsymbol{v}'$ multiply the matrix-valued kernel from both sides. Its coordinate representation, which more closely matches the Euclidean case, will be explored in the sequel. To show that this is indeed the right notion, we prove the following result.

Theorem 4. *The system of marginal distributions of a Gaussian vector field on a smooth manifold X is uniquely determined by a mean vector field $\mu : X \to TX$ and a cross-covariance kernel $k : T^*X \times T^*X \to \mathbb{R}$. Moreover, this correspondence is one-to-one.*

By virtue of defining and characterizing all Gaussian vector fields, Theorem 4 assures us the definition of a kernel introduced is the correct mathematical notion. The constructions presented here are all intrinsic or, in other words, coordinate-free, and do not involve the use of bases. To understand how to perform numerical calculations with these kernels we proceed to study their coordinate representations with respect to a specific choice of basis.

2.3 Gauge Equivariant Kernels

In Section 2.2, we defined what a Gaussian vector field on a manifold is. However, by nature of the manifold setting, the resulting objects are more abstract than usual and do not describe how it can be represented numerically. We now develop a point of view suitable for this task.

To this end, we introduce a *frame* F on X, also known as a *gauge* in physical literature, which is a collection of (not necessarily smooth) vector fields $e_1, \ldots, e_d$ on X such that at each point $x \in X$, the set of vectors $e_1(x), \ldots, e_d(x)$ forms a basis of $T_x X$. The frame allows us to express a vector field f on X as simply a vector-valued function $\boldsymbol{f} = (f^1, \ldots, f^d) : X \to \mathbb{R}^d$, such that $f(x) = \sum_{i=1}^d f^i(x) e_i(x)$ for all $x \in X$. The corresponding *coframe* F^* is defined as a collection $e^1, \ldots, e^d$ of covector fields (one-forms) on X such that $\langle e^i(x) | e_j(x) \rangle = \delta_{ij}$ for all $x \in X$, where δ_{ij} is the Kronecker delta. In the following proposition, we show that if f is a Gaussian vector field on X (in the sense of Definition 1), then the corresponding vector representation $\boldsymbol{f}$ expressed in a given frame is a vector-valued GP in the standard sense.

Proposition 5. *Let f be a Gaussian vector field defined on X with cross-covariance kernel $k : T^*X \times T^*X \to \mathbb{R}$. Given a frame $F = (e_1, \ldots, e_d)$ on X, define $\boldsymbol{f} : X \to \mathbb{R}^d$ as above. Then $\boldsymbol{f}$ is a vector-valued GP in the usual sense with kernel $\mathbf{K}_F : X \times X \to \mathbb{R}^{d \times d}$ given by*

$$\mathbf{K}_F(x, x') = \begin{bmatrix} k(e^1(x), e^1(x')) & \ldots & k(e^1(x), e^d(x')) \\ \vdots & \ddots & \vdots \\ k(e^d(x), e^1(x')) & \ldots & k(e^d(x), e^d(x')) \end{bmatrix}, \tag{4}$$

where (e^i), with raised indices, is the coframe corresponding to (e_i). Conversely, given a vector-valued GP $\boldsymbol{f} = (f^1, \ldots, f^d) : X \to \mathbb{R}^d$ and a frame $F = (e_1, \ldots, e_d)$ on X, $f(\cdot) := \sum_{i=1}^d f^i(\cdot) e_i(\cdot)$ defines a Gaussian vector field on X.

This result shows precisely how numerical representations of a Gaussian vector field depends on the choice of frame. While this representation is not *invariant* under this choice, it is *equivariant*, meaning that a transformation in the frame results in an appropriate transformation of the kernel. To make this notion precise, we introduce a matrix subgroup $G = \mathrm{GL}(d, \mathbb{R})$, called the *gauge group*, that acts on $\mathbb{R}^d$ by a standard matrix-vector multiplication. Given two frames F, F' on X, an abstract vector $f_x \in T_x X$ has two vector representations $\boldsymbol{f}_x, \boldsymbol{f}'_x$ in the respective frames. We say that F' is obtained from F by a *gauge transformation* with respect to a matrix field $\mathbf{A} : X \to G \subseteq \mathbb{R}^{d \times d}$, if

$$\boldsymbol{f}'_x = \mathbf{A}(x) \boldsymbol{f}_x \tag{5}$$

holds for all $x \in X$, and we write $F' = \mathbf{A}F$. Note that $\mathbf{A}(x)$ need not be smooth in x. We see that the gauge transformation is therefore just a linear change of basis of the frame F at each point for which one can identify vectors in $T_x X$ as elements in $\mathbb{R}^d$. The corresponding matrix-valued kernels must also respect this transformation rule, which is the statement of *gauge equivariance*.

Corollary 6. *Let F be a frame on X and $\mathbf{K}_F : X \times X \to \mathbb{R}^{d \times d}$ be the corresponding matrix representation* (4) *of a cross-covariance kernel $k : T^*X \times T^*X \to \mathbb{R}$. This satisfies the* gauge equivariance condition

$$\mathbf{K}_{\mathbf{A}F}(x, x') = \mathbf{A}(x) \mathbf{K}_F(x, x') \mathbf{A}(x')^T, \tag{6}$$

where $\mathbf{A} : X \to G \subseteq \mathbb{R}^{d \times d}$ is a gauge transformation. All cross-covariance kernels in the sense of Proposition 4 arise this way.

Hence, one way to define a Gaussian vector field on a manifold is to find a gauge equivariant kernel. In summary, we have described Gaussian vector fields in a coordinate-free differential-geometric language, and deduced enough properties to confirm the objects defined truly deserve to be called GPs. In doing so, we have both introduced the necessary formalism to the GP community, and obtained a recipe for defining kernels, through a simple gauge equivariance condition atop standard matrix-valued kernels. To proceed towards practical machine learning methods, we therefore study techniques for constructing such kernels explicitly.

3 Model Construction and Bayesian Learning for Riemannian Manifolds

In Section 2, we introduced a notion of a Gaussian vector field. We now study how to use vector fields for machine learning purposes. This entails two primary issues: (a) how to construct practically useful kernels, and (b) once a kernel is constructed, how to train Gaussian processes.

To construct a Gaussian vector field prior, the preceding theory tells us that we need to specify a mean vector field and a cross-covariance kernel. From the definition, it is not at all obvious how to specify a natural kernel, and experience with the scalar-valued case—where the innocuous-looking geodesic squared exponential kernel is generally not positive semi-definite on most Riemannian manifolds [13]—suggests that the problem is delicate, i.e., simply guessing the kernel's form is unlikely to succeed. Our goal, therefore, is to introduce a general construction for building wide classes of kernels from simple building blocks.

The same issues are present if we consider variational approximations to posterior GPs, such as the inducing point framework of Titsias [43]: these are formulated using matrix-vector expressions involving kernel matrices, and it is important for the approximate posterior covariance to be gauge equivariant in order to lead to a valid approximate process. We proceed to address these issues.

3.1 Gauge Equivariant Projected Kernels

Here, we introduce a general technique for defining cross-covariance kernels $k : T^*X \times T^*X \to \mathbb{R}$ and for working with such functions numerically. Section 2 gives us a promising strategy to construct a suitable kernel—namely, it suffices to find a *gauge equivariant matrix-valued kernel*. At first glance, it is not obvious how to construct such a kernel in the manifold setting. On many manifolds, such as the sphere, owing to the hairy ball theorem, every frame must be discontinuous: therefore, constructing a continuous kernel in such a choice of frame appears difficult.

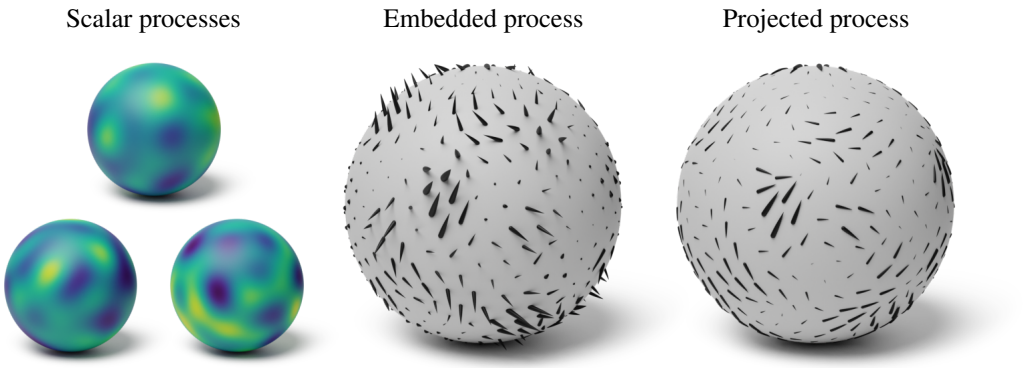

Figure 2: Illustration of the construction process of the projected process. The manifold $\mathbb{S}^2$ is embedded into $\mathbb{R}^3$. Three identical scalar GPs (left) are placed on the manifold. These three scalar GPs are combined to construct a vector-valued GP in the ambient Euclidean space (center). This GP is then projected onto the tangent space of $\mathbb{S}^2$ as a subspace of the tangent space of $\mathbb{R}^3$ (right).

To both get around these obstacles, and aid numerical implementation, we propose to isometrically embed the manifold into Euclidean space.[3] Doing so greatly simplifies these issues by virtue of making it possible to represent the manifold using a single global coordinate system. On the other hand, the main trade-off from this choice is that by its extrinsic nature, the construction can make theoretical analysis more difficult. To proceed, we need two ingredients.

1. An isometric embedding $\mathrm{emb} : X \to \mathbb{R}^{d'}$ of the manifold.
2. A vector-valued Gaussian process $\boldsymbol{f}' : X \to \mathbb{R}^{d'}$ in the standard sense.[4]

A simple choice which reflects the geometry of the manifold is to take $\boldsymbol{f}'$ to be d' independent scalar-valued GPs on X.

By standard results in differential geometry, any smooth map $\phi : X \to X'$ between two manifolds induces a corresponding linear map on the tangent spaces $\mathrm{d}_x\phi : T_xX \to T_{\phi(x)}X'$, which can loosely be thought of as mapping ϕ to its first-order Taylor expansion at x. Thus, an embedding $\mathrm{emb} : X \to \mathbb{R}^{d'}$, induces a map $\mathrm{d}_x\mathrm{emb} : T_xX \to T_{\mathrm{emb}(x)}\mathbb{R}^{d'}$. Now fixing a frame F on X, each tangent space T_xX can be identified with $\mathbb{R}^d$, so without loss of generality, the map $\mathrm{d}_x\mathrm{emb}$ can be expressed simply as a position-dependent matrix $\mathbf{P}_x^T \in \mathbb{R}^{d'\times d}$. Taking the transpose, we obtain $\mathbf{P}_x \in \mathbb{R}^{d\times d'}$, which we call the *projection matrix*. The desired Gaussian vector field on X, with respect to F, is then constructed as $\boldsymbol{f}(x) = \mathbf{P}_x\boldsymbol{f}'(x)$. This procedure is illustrated in Figure 2: there we see that to get a vector field on an $\mathbb{R}^3$-embedded sphere, we may take a vector-valued function on it and project its values to make vectors tangential to this sphere, thus obtaining a valid vector field. Since the projection operator preserves smoothness and since we can take a smooth vector-valued GP to begin with, it is clear that this approach may be used to build smooth vector fields.

We prove that (a) the resulting expression is, indeed, a kernel, and that (b) no expressivity is lost via the construction because all cross-covariance kernels arise this way.

Proposition 7. *Let (X, g) be a Riemannian manifold, $\mathrm{emb} : X \to \mathbb{R}^{d'}$ be an isometric embedding and F be a frame on X. We denote by $\mathbf{P}_{(\cdot)} : X \to \mathbb{R}^{d\times d'}$ the associated projection matrix under F, and let $\boldsymbol{f}' : X \to \mathbb{R}^{d'}$ be any vector-valued Gaussian process with matrix-valued kernel $\boldsymbol{\kappa} : X \times X \to \mathbb{R}^{d'\times d'}$. Then, the vector-valued function $\boldsymbol{f} = \mathbf{P}\boldsymbol{f}'$ defines a Gaussian vector field f on X using the construction in Proposition 5, whose kernel under the frame F has matrix representation*

$$\mathbf{K}_F(x, x') = \mathbf{P}_x\boldsymbol{\kappa}(x, x')\mathbf{P}_{x'}^T. \tag{7}$$

[3] An embedding $\mathrm{emb} : X \to \mathbb{R}^{d'}$ is called *isometric* if it preserves the metric tensor. By Nash's Theorem [26], such embeddings exist for any d-dimensional manifold, with an embedded dimension $d' \leq 2d + 1$.

[4] We emphasize again that $\boldsymbol{f}'$ is *not* a Gaussian vector field because it is not a random section. In particular, note that $d' > d$ for most embeddings.

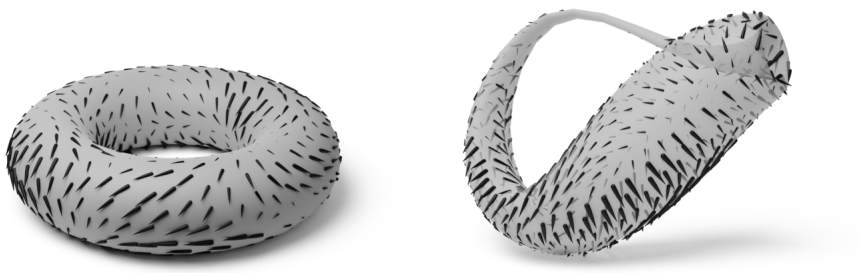

Figure 3: Random samples from Gaussian processes with gauge equivariant projected kernels, on the torus and Klein bottle, respectively. The latter is a non-orientable manifold: the ability to handle such cases highlights the generality of projected kernels.

Moreover, all cross-covariance kernels $k : T^*X \times T^*X \to \mathbb{R}$ arise this way. We call a kernel defined this way a projected kernel.

To construct these kernels we require scalar-valued kernels on manifolds to use as a basic building block. These are studied in the general Riemannian setting by Lindgren et al. [28] and Borovitskiy et al. [2]: relying on these kernels is the only reason we require the Riemannian structure. It is also possible to obtain such kernels using embeddings, following Lin et al. [27]. Similar techniques to those we consider are used by Freeden and Schreiner [14] to construct vector-valued zonal kernels on the sphere: in contrast, we work with arbitrary manifolds. The projection kernel idea is a very general way to build kernels for vector fields by combining scalar kernels, but effective scalar kernels, naturally, rely on Riemannian structure.[5] Figure 3 shows random samples from Gaussian processes constructed with the described kernels.

The projected kernel construction both makes it easy to define cross-covariance kernels on general manifolds, and describes a straightforward way to implement them numerically by representing the embedded manifold in coordinates and calculating the resulting matrix-vector expressions. The constructed kernel depends on the embedding, but can be transformed appropriately if switching to a different embedding. Embeddings, in turn, are available for most manifolds of practical interest, and are obtained automatically for manifolds approximated numerically as meshes. Everything described is constructive and fully compatible with the modern automatic-differentiation-based machine learning toolkit, and most operations for constructing and/or sampling from specialized priors [44, 22, 23], including on spaces such as the sphere where specific analytic tools are available [6, 8, 11, 10]. With these kernels in hand, we thus proceed to study training methods.

3.2 Gauge Equivariant Variational Approximations

We now discuss variational inference for training GPs in the Riemannian vector field setting. Approximations, such as the inducing-point framework by Titsias [43] and Hensman et al. [17], approximate the posterior GP with another GP, termed the *variational approximation*. The latter is typically constructed by specifying a multivariate Gaussian at a set of test locations with a parameterized mean and kernel matrix. For example, Opper and Archambeau [31] consider $\mathrm{N}(\mathbf{m}, \mathbf{S})$, where

$$\mathbf{m} = \mathbf{K}_{(\cdot)\boldsymbol{z}}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} \qquad \mathbf{S} = \mathbf{K}_{(\cdot,\cdot)} - \mathbf{K}_{(\cdot)\boldsymbol{z}}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \boldsymbol{\Sigma})^{-1}\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \boldsymbol{\Sigma})^{-1}\mathbf{K}_{\boldsymbol{z}(\cdot)}. \quad (8)$$

The variational parameters include a set of inducing locations $\boldsymbol{z}$, a mean vector $\boldsymbol{\mu}$, and a block-diagonal cross-covariance matrix $\boldsymbol{\Sigma}$. Training proceeds by optimizing these parameters to minimize the Kullback–Leibler divergence of the variational distribution from the true posterior, typically using mini-batch stochastic gradient descent.

In the last decade, a wide and diverse range of inducing point approximations suited for many different settings have been proposed [43, 31, 25, 44, 48]. The vast majority of them employ coordinate-

[5] The structure of a smooth manifold is not rigid enough to define natural kernels. For instance, smooth structure of the sphere is indistinguishable from the smooth structure of an ellipsoid or even of the dragon manifold from Borovitskiy et al. [2], but their Riemannian structures differ considerably.

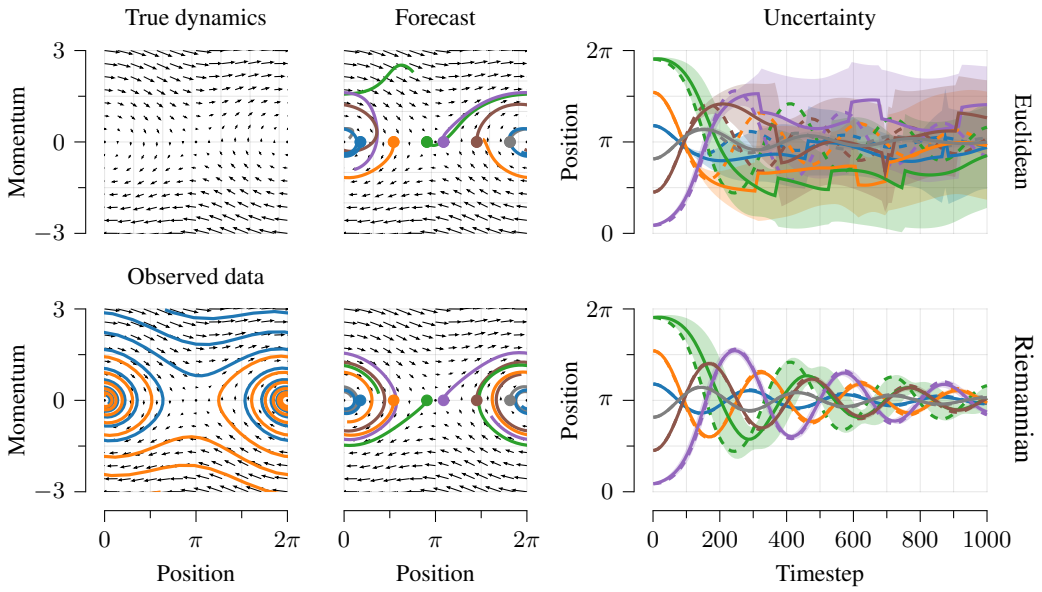

Figure 4: Upper left: Pendulum with friction state space. Lower left: Two rollouts used to train the GP. Upper middle: State space plot of rollouts from a standard Euclidean GP condtioned on the training data. Upper right: Temporal plot of rollouts from a standard Euclidean vector GP. Solid line is the true rollout, dashed line and shade is the mean and $\pm$ 1 std of the GP rollouts. Lower middle and right: same for a geometric manifold vector field kernel on $\mathbb{S}^1 \times \mathbb{R}$.

dependent matrix-vector expressions. This raises the question, which of these constructions can be adapted to define valid variational approximations in the vector field setting?

To proceed, one can choose a frame and formulate a given variational approximation using matrices defined with respect to this frame. To ensure well-definedness, one must ensure that all these matrices, such as the kernel matrix and the variational parameter $\mathbf{\Sigma}$ in (8), are gauge equivariant. These considerations can be simplified adopting the pathwise view of GPs, and examining the random variables directly. For example, the variational approximation of Opper and Archambeau [31] shown previously in (8) can be reinterpreted pathwise as the GP

$$(f \mid y)(\cdot) \approx f(\cdot) + \mathbf{K}_{(\cdot)\boldsymbol{z}}(\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} + \mathbf{\Sigma})^{-1}(\boldsymbol{\mu} - f(\boldsymbol{z}) - \boldsymbol{\varepsilon}) \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{9}$$

where we view the matrices $\mathbf{K}_{(\cdot)\boldsymbol{z}}, \mathbf{K}_{\boldsymbol{z}\boldsymbol{z}}, \mathbf{\Sigma}$ as *linear operators* between direct sums of tangent spaces: $\mathbf{K}_{(\cdot)\boldsymbol{z}} : T_{z_1}X \oplus \ldots \oplus T_{z_m}X \to T_{(\cdot)}X$ and $\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}}, \mathbf{\Sigma} : T_{z_1}X \oplus \ldots \oplus T_{z_m}X \to T_{z_1}X \oplus \ldots \oplus T_{z_m}X$. By virtue of being defined at the level of vector fields using components that are all intrinsically valid, the posterior covariance of the resulting variational approximation is *automatically* gauge equivariant. Hence, checking gauge equivariance is then equivalent to deducing the domains and ranges of these operators from their coordinate representations, and checking if they are compatible. This applies to any variational family that can be constructed in the given manner.

The vast majority of inducing point constructions can be interpreted in this manner and thus extend readily to the Riemannian vector field setting by simply representing the necessary matrices in a chosen frame. In particular, the classical approach of Titsias [43] is gauge equivariant.

4 Illustrated Examples

Here, we showcase a number of examples that illustrate potential use cases of the models developed.

4.1 Dynamical Systems Modeling

Here, we show how Gaussian vector fields can be used to learn the equations of motion of a physical system—an important task in imitation learning, model-based reinforcement learning, and robotics. GPs are an attractive model class in this area owing to their ability to represent and propagate

uncertainty, which enables them to separate what is known about an environment from what is not, thereby driving data-efficient exploration.

For a prototype physical system, we consider an ideal pendulum, whose configuration space is the circle $\mathbb{S}^1$, representing the angle of the pendulum, with zero being at the bottom of the loop, and whose position-momentum state-space is the cylinder $\mathbb{S}^1 \times \mathbb{R}$. We consider conservative dynamics with additional friction applied at the pivot. Since this system is non-conservative, we cannot just learn the Hamiltonian of the system, but must learn the vector field over the state space that defines the dynamics of the system. The true dynamics of the system are given by the differential equations

$$\mathcal{H} = \frac{p^2}{2ml^2} + mgl(1 - \cos(q)) \qquad \frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial p} \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial q} - \frac{b}{m}p, \tag{10}$$

where $\mathcal{H}$ is the Hamiltonian of the system defining the conservative part of the dynamics, q and p are the position and momentum of the pendulum, m is the mass, l is the length, g is the gravitational field strength and b is a friction parameter. Experimental details can be found in Appendix B.

To learn this model, we initialise the system at two start points, and evolve the system using leapfrog integration. From these observations of position, we backward Euler integrate the momentum of the system, and from these position-momentum trajectories we estimate observations of the dynamics field. Using these observations, we condition a sparse GP. The result is an estimate of the system dynamics with suitable uncertainty estimates. In order to compute rollouts of these dynamics, we use pathwise sampling of this sparse GP [46, 47] for speed together with leapfrog integration.

Results can be seen in Figure 4. While the Euclidean GP performs reasonably well at the start of the rollouts, once the trajectory crosses the discontinuity caused by looping the angle back around to zero, the system starts to make incoherent predictions: this is due to the discontinuity arising from wrap-around condition of the angle. The manifold vector-valued GP does not have this issue as the learned and sampled dynamics fields are continuous throughout the state-space.

4.2 Weather Modeling

In this experiment, we show how vector-valued GPs on manifolds can be used in the context of meteorology, where geometric information often plays an important role in accurately modeling global weather fields [45, 38, 19]. *Data assimilation* in numerical weather forecasting refers to the practice of using observed data to update predictions of the atmosphere closer to the truth. Uncertainty plays a critical role here: it is not usually possible to observe the weather at all locations on the globe simultaneously, and taking into account observation uncertainty is crucial in numerical weather forecasting during the data assimilation step [24, 36]. In this section, we explore Gaussian processes as a tool for carrying out global interpolation of wind fields, while simultaneously performing uncertainty quantification, mirroring *optimal interpolation* techniques in data assimilation [20].

We consider a simplified setting, where the goal is to interpolate the wind velocity observed by the Aeolus satellite [37], which uses LiDAR sensors to measure wind velocity directly. To mimic this setting, we use an hour of the Aeolus satellite track during the period 2019/01/01 09:00-10:00 for the input locations and the wind velocity data (10m above ground) from the ERA5 atmospheric reanalysis data [18] interpolated at these locations, to simulate measurements taken from the Aeolus satellite. We subtract the weekly historical average wind velocity from the observations, before training the GP models, where the historical mean is computed from the hourly wind data (10 m above ground) from the WeatherBench dataset [35], available from 1979–2018. Further details can be found in Appendix B. We compare the results of a Matérn-3/2 manifold vector-valued GP regression model fitted on the wind anomaly observations along the Aeolus trajectory, with the results from a Euclidean Matérn-3/2 multi-output GP trained on the same data, except projected onto a latitude-longitude map.

Results are shown in Figure 5, where the benefits of using a manifold vector-valued GP become clear. When the satellite crosses the left/right boundary in the lat/lon projection, the outputs from the Euclidean vector-valued GP give rise to a spurious discontinuity in the uncertainty along the solid pink line. In addition, predictions become less certain in the Euclidean case as the satellite approaches the poles, which is simply an artifact of the distortion caused by projecting the spherical data onto the plane. By construction, the manifold vector-valued GP is able to avoid both of these issues, resulting in a more realistic prediction with much more uniform uncertainty along the satellite trajectory from pole to pole. In addition, the predictions from the manifold GP are more certain overall, due to the useful structural bias embedded in the kernel.

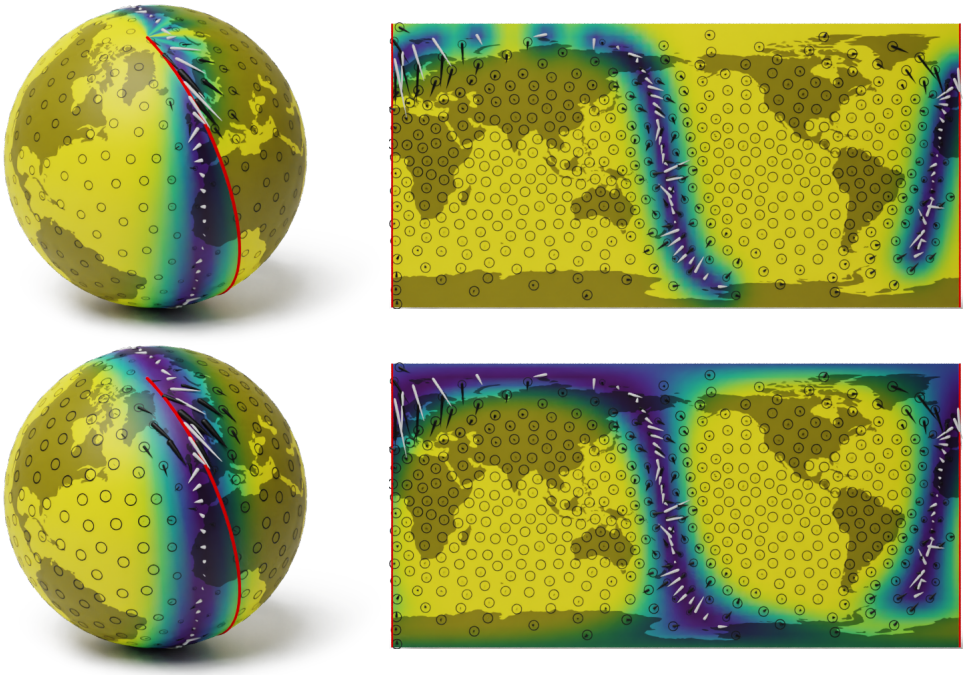

Figure 5: Top row: Euclidean GP trained on wind measurements along the chosen Aeolus satellite trajectory, viewed as deviation from normal with respect to the historical average vector field. White arrows are the satellite measurements, black arrows and ellipsoids are the posterior mean and cross-covariance of the vector field, colors indicate the posterior standard deviation norm, and the solid red line indicates the latitudinal boundary when the sphere is projected onto the plane using the lat/lon projection. Bottom row: Same as above except using a manifold kernel on $\mathbb{S}^2$.

5 Conclusion

In this paper, we propose techniques that generalize Gaussian processes to model vector fields on Riemannian manifolds. This is done by first providing a well-defined notion of such processes on manifolds and then introducing an explicit method to construct them in a way that respects the underlying geometry. By virtue of satisfying the key condition of gauge equivariance, our construction is coordinate-free and thus meaningful on manifolds. In addition to this, we extend standard Gaussian process training methods, such as variational inference, to this setting, and verify that such methods are also compatible with gauge equivariance. This theoretical work gives practitioners additional tools for stochastic modeling of vector fields on manifolds. As such, its societal impact will be mainly determined by the applications that belong to the domain of future work. We demonstrate our techniques on a series of examples in modeling dynamical systems and weather science, and show that incorporating geometric structural bias into probabilistic modeling is beneficial in these settings to obtain coherent predictions and uncertainties.

6 Acknowledgements and Funding Disclosure

Michael is supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). Section 3 and Appendix A of this paper were solely financially supported by the RSF grant №21-11-00047 (https://rscf.ru/en/project/21-11-00047/), but was contributed to by all authors in equal proportion with the remaining work. Other sections were supported by the remaining funding. We also acknowledge support from Huawei Research. We thank Peter Dueben for useful suggestions in Section 4.2, including, in particular, making us aware of the Aeolus satellite. We are grateful to Nick Sharp for introducing us to the process used for making the three-dimensional figures appearing in this work.

References

[1] V. Borovitskiy, I. Azangulov, A. Terenin, P. Mostowsky, M. P. Deisenroth, and N. Durrande. Matérn Gaussian Processes on Graphs. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021. Cited on page 1.

[2] V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Matern Gaussian Processes on Riemannian Manifolds. In *Advances in Neural Information Processing Systems*, 2020. Cited on pages 1, 2, 7, 22–24.

[3] E. Brochu, V. M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. Technical report TR-2009-023, Department of Computer Science, University of British Columbia, 2009. Cited on page 1.

[4] Y. Canzani. Analysis on Manifolds via the Laplacian. *Lectures notes, Harvard University*, 2013. Cited on page 22.

[5] T. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling. Gauge Equivariant Convolutional Networks and the Icosahedral CNN. In *Proceedings of the International Conference on Machine Learning*, 2019. Cited on page 2.

[6] P. E. Creasey and A. Lang. Fast generation of isotropic Gaussian random fields on the sphere. *Monte Carlo Methods and Applications*, 2018. Cited on page 7.

[7] M. P. Deisenroth and C. E. Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *Proceedings of the International Conference on Machine Learning*, 2011. Cited on page 1.

[8] V. Dutordoir, N. Durrande, and J. Hensman. Sparse Gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, 2020. Cited on page 7.

[9] ECMWF. The ESA ADM-Aeolus Doppler Wind Lidar Mission. 2017. URL: https://www.ecmwf.int/sites/default/files/elibrary/2016/16851-esa-adm-aeolus-doppler-wind-lidar-mission-status-and-validation-strategy.pdf. Cited on page 24.

[10] X. Emery, R. Furrer, and E. Porcu. A turning bands method for simulating isotropic Gaussian random fields on the sphere. *Statistics & Probability Letters*, 2019. Cited on page 7.

[11] X. Emery and E. Porcu. Simulating isotropic vector-valued Gaussian random fields on the sphere through finite harmonics approximations. *Stochastic Environmental Research and Risk Assessment*, 2019. Cited on page 7.

[12] M. Émery. *Stochastic calculus in manifolds*. Springer, 2012. Cited on page 2.

[13] A. Feragen, F. Lauze, and S. Hauberg. Geodesic Exponential Kernels: When Curvature and Linearity Conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3032–3042, 2015. Cited on pages 2, 5.

[14] W. Freeden and M. Schreiner. *Spherical functions of mathematical geosciences: a scalar, vectorial, and tensorial setup*. Springer, 2008. Cited on page 7.

[15] P. D. Haan, M. Weiler, T. Cohen, and M. Welling. Gauge Equivariant Mesh CNNs: Anisotropic Convolutions on Geometric Graphs. In *Proceedings of the International Conference on Learning Representations*, 2021. Cited on page 2.

[16] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, 2006. Cited on page 23.

[17] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *Uncertainty in Artificial Intelligence*, 2013. Cited on page 7.

[18] H. Hersbach. The ERA5 Atmospheric Reanalysis. In *AGU Fall Meeting Abstracts*, 2016. Cited on page 9.

[19] I. L. Jover. *Geometric Deep Learning for Medium Range Weather Prediction*. Master's thesis, EPFL, 2020. Cited on page 9.

[20] E. Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003. Cited on page 9.

[21] A. Krause, A. Singh, and C. Guestrin. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Journal of Machine Learning Research*, 2008. Cited on page 1.

[22] M. Lange-Hegermann. Algorithmic Linearly Constrained Gaussian Processes. *Advances in Neural Information Processing Systems*, 2018. Cited on page 7.

[23] M. Lange-Hegermann. Linearly Constrained Gaussian Processes with Boundary Conditions. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on page 7.

[24] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation*. Springer, 2015. Cited on page 9.

[25] M. Lázaro-Gredilla and A. R. Figueiras-Vidal. Inter-domain Gaussian Processes for Sparse Inference Using Inducing Features. In *Advances in Neural Information Processing Systems*, 2009. Cited on page 7.

[26] J. M. Lee. Smooth Manifolds. In *Introduction to Smooth Manifolds*. Springer, 2013. Cited on pages 1, 6.

[27] L. Lin, N. Mu, P. Cheung, and D. Dunson. Extrinsic Gaussian processes for regression and classification on manifolds. *Bayesian Analysis*, 2019. Cited on page 7.

[28] F. Lindgren, H. Rue, and J. Lindström. An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011. Cited on page 7.

[29] A. Mallasto and A. Feragen. Wrapped Gaussian process regression on Riemannian manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. Cited on page 2.

[30] A. Mallasto, S. Hauberg, and A. Feragen. Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019. Cited on page 2.

[31] M. Opper and C. Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 2009. Cited on pages 7, 8.

[32] M. A. Osborne, R. Garnett, and S. J. Roberts. Gaussian Processes for Global Optimization. In *Proceedings of the International Conference on Learning and Intelligent Optimization*, 2009. Cited on page 1.

[33] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, 2008. Cited on page 22.

[34] C. E. Rasmussen and M. Kuss. Gaussian Processes in Reinforcement Learning. In *Advances in Neural Information Processing Systems*. 2004. Cited on page 1.

[35] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 2020. Cited on pages 9, 23.

[36] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015. Cited on page 9.

[37] O. Reitebuch. The Spaceborne Wind LiDAR Mission ADM-Aeolus. In *Atmospheric Physics*. 2012. Cited on page 9.

[38] S. Scher and G. Messori. Spherical Convolution and Other Forms of Informed Machine Learning for Deep Neural Network Based Weather Forecasts. *arXiv:2008.13524*, 2020. Cited on page 9.

[39] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 2016. Cited on page 1.

[40] D. W. Stroock. *An introduction to the analysis of paths on a Riemannian manifold*. American Mathematical Society, 2000. Cited on page 2.

[41] D. J. Sutherland and J. Schneider. On the error of random Fourier features. In *Uncertainty in Artificial Intelligence*, 2015. Cited on page 22.

[42] T. Tao. *An Introduction to Measure Theory*. American Mathematical Society, 2011. Cited on page 16.

[43] M. K. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, 2009. Cited on pages 5, 7, 8.

[44] M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman. A Framework for Interdomain and Multioutput Gaussian Processes. arXiv:2003.01115, 2020. Cited on page 7.

[45] J. A. Weyn, D. R. Durran, and R. Caruana. Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems*, 2020. Cited on page 9.

[46] J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Pathwise Conditioning of Gaussian Processes. In *Proceedings of the International Conference on Machine Learning*, 2020. Cited on pages 2, 9, 22, 23.

[47] J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Pathwise Conditioning of Gaussian Processes. *Journal of Machine Learning Research*, 2021. Cited on pages 2, 9, 22, 23.

[48] L. Wu, A. Miller, L. Anderson, G. Pleiss, D. Blei, and J. Cunningham. Hierarchical Inducing Point Gaussian Process for Inter-domian Observations. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on page 7.

A Theory

Preliminaries on Gaussian measures

Since we are working in a setting beyond $\mathbb{R}^d$, we need a suitable notion of a multivariate Gaussian that can be employed in a coordinate-free manner. We employ the notion of a Gaussian in the sense of duality, given below. These notions are standard and classical, but since they are not well-known in machine learning, and for completeness, we prove the necessary properties ourselves.

Definition 8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let V by a finite-dimensional real topological vector space, equipped with the standard topology, Borel σ-algebra, and the canonical pairing $\langle \cdot \mid \cdot \rangle : V^* \times V \to \mathbb{R}$ with its topological dual V^*. A random vector $v : \Omega \to V$ is called* Gaussian *if, for all $\phi \in V^*$, the random variable $\langle \phi \mid v \rangle : \Omega \to \mathbb{R}$ is univariate Gaussian.*

Remark. It is not hard to show that in the setting of the definition above, the random variables $\langle \phi_1 \mid v \rangle, \ldots, \langle \phi_k \mid v \rangle$ are jointly Gaussian for any finite collection $\phi_1, \ldots, \phi_k \in V^*$. Indeed, this is equivalent to the Gaussianity of every linear combination $\alpha_1 \langle \phi_1 \mid v \rangle + \ldots + \alpha_k \langle \phi_k \mid v \rangle = \langle \alpha_1 \phi_1 + \ldots + \alpha_k \phi_k \mid v \rangle$, which is also ensured by the definition since $\alpha_1 \phi_1 + \ldots + \alpha_k \phi_k \in V^*$.

We begin by showing that a Gaussian random vector in the sense of duality is characterized by a mean and a covariance, just like Gaussians in the standard, coordinate-dependent sense, starting with defining appropriate analogs of both notions in this setting.

Lemma 9. *For every Gaussian random vector v, there is a unique vector $\mu \in V$ and unique symmetric positive semi-definite bilinear form $k : V^* \times V^* \to \mathbb{R}$ such that for all $\phi \in V^*$, we have $\mathbb{E}\langle \phi \mid v \rangle = \langle \phi \mid \mu \rangle$ and $k(\phi, \psi) = \mathrm{Cov}(\langle \phi \mid v \rangle, \langle \psi \mid v \rangle)$. We say that μ is its* mean *and k is its* covariance form*, and write $v \sim \mathrm{N}(\mu, k)$.*

Proof. Consider the map $\mathbb{E}\langle \cdot \mid v \rangle : V^* \to \mathbb{R}$. This map is a linear functional on the space V^*. Since V is finite-dimensional, V is reflexive, so there is exactly one vector $\mu \in V$ such that

$$\langle \phi \mid \mu \rangle = \mathbb{E}\langle \phi \mid v \rangle \tag{11}$$

for all $\phi \in V^*$. Next, define k as

$$k(\phi, \psi) = \mathrm{Cov}(\langle \phi \mid v \rangle, \langle \psi \mid v \rangle) \tag{12}$$

for all $\phi, \psi \in V^*$. Clearly, k is bilinear and positive semi-definite, that is $k(\phi, \phi) \geq 0$ for all $\phi \in V^*$. Thus the claim follows. □

This tells us that every Gaussian random vector admits a mean and covariance: we now show that such Gaussians exist and are uniquely determined by this pair. Recall that for a measure π, and a measurable function ϕ, the *pushfoward measure* $\phi_* \pi$ is defined as $(\phi_* \pi)(A) = \pi(\phi^{-1}(A))$ for all measurable sets A.

Lemma 10. *For any vector $\mu \in V$ and any positive semi-definite bilinear form $k : V^* \times V^* \to \mathbb{R}$, there exists a random vector $v \sim \mathrm{N}(\mu, k)$. Moreover, if $w : \Omega \to V$ is another Gaussian random vector in the sense of Definition 8 with $w \sim \mathrm{N}(\mu, k)$, then v and w are identically distributed.*

Proof. Choose a basis (e_i) on V, and let (e^i) be the dual basis. Define the vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and matrix $\mathbf{K} \in \mathbb{R}^{d \times d}$ by

$$\boldsymbol{\mu} = \begin{bmatrix} \langle e^1 \mid \mu \rangle \\ \vdots \\ \langle e^d \mid \mu \rangle \end{bmatrix} \qquad \mathbf{K} = \begin{bmatrix} k(e^1, e^1) & \ldots & k(e^1, e^d) \\ \vdots & \ddots & \vdots \\ k(e^d, e^1) & \ldots & k(e^d, e^d) \end{bmatrix}. \tag{13}$$

By positive semi-definiteness of k, the matrix $\mathbf{K}$ is a positive semi-definite matrix, so there exists a random vector $\boldsymbol{v} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$ in the classical Euclidean sense. Let $\mathcal{E} : V \to \mathbb{R}^n$ be the continuous linear isomorphism induced by the basis and define

$$v = \mathcal{E}^{-1} \boldsymbol{v}. \tag{14}$$

We claim that (a) v is Gaussian, that is, if we test it against any covector, we obtain a univariate Gaussian, (b) the mean vector of v is μ, and (c) the covariance form of v is k. To show (a), let v^i

denote the components of $\boldsymbol{v}$ (scalar Gaussian random variables) so that $v = \sum_{i=1}^d v^i e_i$ and for any $\phi \in V^*$, write $\phi = \sum_{i=1}^d \phi_i e^i$, where $\phi_i = \langle \phi \mid e_i \rangle$. Then we have

$$\langle \phi \mid v \rangle = \left\langle \sum_{i=1}^d \phi_i e^i \,\middle|\, \sum_{j=1}^d v^j e_j \right\rangle = \sum_{i=1}^d \sum_{j=1}^d \phi_i v^j \underbrace{\langle e^i \mid e_j \rangle}_{\delta_{ij}} = \sum_{i=1}^d \phi_i v^i. \tag{15}$$

Since each v^i is a univariate Gaussian, the linear combination on the right hand side is also a univariate Gaussian, which proves (a). To prove (b) and (c), we see that for any $\phi \in V^*$,

$$\mathbb{E}\langle \phi \mid v \rangle = \mathbb{E}\langle \phi \mid \sum_{i=1}^d v^i e_i \rangle = \mathbb{E} \sum_{i=1}^d v^i \langle \phi \mid e_i \rangle \tag{16}$$

$$= \sum_{i=1}^d \underbrace{(\mathbb{E}\, v^i)}_{\langle e^i \mid \mu \rangle} \langle \phi \mid e_i \rangle = \langle \phi \mid \sum_{i=1}^d \langle e^i \mid \mu \rangle e_i \rangle = \langle \phi \mid \mu \rangle. \tag{17}$$

Thus v has the right mean. Now take an additional $\psi \in V^*$ and write

$$\operatorname{Cov}\Big(\langle \phi \mid v \rangle, \langle \psi \mid v \rangle\Big) = \mathbb{E}\big((\langle \phi \mid v \rangle - \langle \phi \mid \mu \rangle)(\langle \psi \mid v \rangle - \langle \psi \mid \mu \rangle)\big) \tag{18}$$

$$= \mathbb{E}\left(\sum_{i=1}^d (v^i - \langle e^i \mid \mu \rangle) \langle \phi \mid e_i \rangle\right)\left(\sum_{j=1}^d (v^j - \langle e^j \mid \mu \rangle) \langle \psi \mid e_j \rangle\right) \tag{19}$$

$$= \sum_{i=1}^d \sum_{j=1}^d \langle \phi \mid e_i \rangle \underbrace{\mathbb{E}\big((v^i - \langle e^i \mid \mu \rangle)(v^j - \langle e^j \mid \mu \rangle)\big)}_{k(e^i, e^j)} \langle \psi \mid e_j \rangle \tag{20}$$

$$= k\left(\sum_{i=1}^d \langle \phi \mid e_i \rangle e^i, \sum_{j=1}^d \langle \psi \mid e_j \rangle e^j\right) = k(\phi, \psi), \tag{21}$$

hence v has the right covariance form.

Now let $w : \Omega \to V$ be another Gaussian random vector with $w \sim \mathrm{N}(\mu, k)$, and let π_w be its pushforward measure. Similarly, let π_v be the pushforward measure of v. Reversing the above argument, we see that pushforwards of measures π_v and π_w through $\mathcal{E}$, which we denote by $\pi_{\boldsymbol{v}}$ and $\pi_{\boldsymbol{w}}$, are both Gaussian distributions (in the classical sense) in $\mathbb{R}^d$ with the same mean vectors $\boldsymbol{\mu}$ and covariance matrices $\mathbf{K}$. Hence $\pi_{\boldsymbol{v}} = \pi_{\boldsymbol{w}}$ in distribution, but since $\mathcal{E}$ is a measurable space isomorphism,[6] we have $\pi_v = \pi_w$, which proves the claim. □

Lemmas 9 and 10 show that a pair μ, k defines a unique probability distribution on V which we call the Gaussian distribution with mean vector μ and covariance form k on the vector space V and denote by $\mathrm{N}(\mu, k)$. This establishes a notion of Gaussianity that is suitable and natural for describing finite-dimensional marginals in a coordinate-free manner.

Existence and uniqueness (Proof of Theorem 4)

Here, we prove that Gaussian vector fields exist and are uniquely determined by their mean vector field and cross-covariance kernel. Our goal now is, from a cross-covariance kernel, to construct a projective family of finite-dimensional marginals.

Definition 11 (Preliminaries)**.** *Let X be a smooth manifold. Let*

$$\Gamma_{\mathrm{nns}}(TX) = \{f : X \to TX : \mathrm{proj}_X \circ f = \mathrm{id}_X\} \tag{22}$$

be the vector space of not necessarily smooth sections.

Definition 12 (Cross-covariance kernel)**.** *A symmetric function $k : T^*X \times T^*X \to \mathbb{R}$ is called* fiberwise bilinear *if at any pair of points $x, x' \in X$, we have*

$$k(\lambda\alpha_x + \mu\beta_x, \gamma_{x'}) = \lambda k(\alpha_x, \gamma_{x'}) + \mu k(\beta_x, \gamma_{x'}) \tag{23}$$

[6] A measurable space isomorphism is a measurable bijection with a measurable inverse.

for any $\alpha_x, \beta_x \in T_x^*X$, $\gamma_{x'} \in T_{x'}^*X$ and $\lambda, \mu \in \mathbb{R}$, where we note by symmetry that the same requirement applies to its second argument. A fiberwise bilinear function k is called positive semi-definite *if for any set of covectors $\alpha_{x_1}, \ldots, \alpha_{x_n} \in T^*X$, we have*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} k(\alpha_{x_i}, \alpha_{x_j}) \geq 0. \tag{24}$$

We call a symmetric fiberwise bilinear positive semi-definite function a cross-covariance kernel.

We show in the following example that this definition of the cross-covariance kernel is compatible with the notion of matrix-valued kernels used in classical vector-valued GPs and extends it naturally.

Example 13 (Euclidean case)**.** *Consider $X = \mathbb{R}^d$ with a fixed inner product and an orthonormal basis, under which $\mathbb{R}^d$ is identified with $\left(\mathbb{R}^d\right)^*$. Consider a matrix-valued kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ in the standard sense. Let $k((x,v),(x',v')) = v^T\kappa(x,x')v'$. Then $k : T^*\mathbb{R}^d \times T^*\mathbb{R}^d \to \mathbb{R}$ is a cross-covariance kernel in the above sense.*

Indeed, k is symmetric and fiberwise bilinear. Moreover, since κ is positive semi-definite in the regular sense, we have that for arbitrary $x_1, \ldots, x_n \in \mathbb{R}^d$, the $nd \times nd$ matrix

$$\Gamma(x_1, \ldots, x_n) = \begin{bmatrix} \kappa(x_1,x_1) & \ldots & \kappa(x_1,x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n,x_1) & \ldots & \kappa(x_n,x_n) \end{bmatrix} \tag{25}$$

is positive semi-definite, meaning that for an arbitrary collection $v_1, \ldots, v_n \in \mathbb{R}^d$, we have

$$0 \leq \begin{bmatrix} v_1^T & \ldots & v_n^T \end{bmatrix} \begin{bmatrix} \kappa(x_1,x_1) & \ldots & \kappa(x_1,x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n,x_1) & \ldots & \kappa(x_n,x_n) \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{n} \underbrace{v_i^T\kappa(x_i,x_j)v_j}_{k((x_i,v_i),(x_j,v_j))}. \tag{26}$$

Condition (24) *thus follows, proving that this is a valid cross-covariance kernel.*

We proceed to introduce the system of coordinate-free finite-dimensional marginals that will be used to construct the vector-valued GP.

Definition 14. *Let $\mu \in \Gamma_{\text{nns}}(TX)$ and $k : T^*X \times T^*X \to \mathbb{R}$ be a cross-covariance kernel. For any $x_1, \ldots, x_n \in X$, let $V_{x_1,\ldots,x_n} = T_{x_1}X \oplus \ldots \oplus T_{x_n}X$ and $V^*_{x_1,\ldots,x_n} = T^*_{x_1}X \oplus \ldots \oplus T^*_{x_n}X$. Define $\mu_{x_1,\ldots,x_n} \in V_{x_1,\ldots,x_n}$ and $k_{x_1,\ldots,x_n} : V^*_{x_1,\ldots,x_n} \times V^*_{x_1,\ldots,x_n} \to \mathbb{R}$ by*

$$\mu_{x_1,\ldots,x_n} = (\mu(x_1), \ldots, \mu(x_n)) \qquad k_{x_1,\ldots,x_n}(\alpha,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} k(\alpha_{x_i}, \beta_{x_j}) \tag{27}$$

for any $\alpha = (\alpha_{x_1}, \ldots, \alpha_{x_n})$, $\beta = (\beta_{x_1}, \ldots, \beta_{x_n}) \in V^*_{x_1,\ldots,x_n}$. We denote $\pi_{x_1,\ldots,x_n} = \mathrm{N}(\mu_{x_1,\ldots,x_n}, k_{x_1,\ldots,x_n})$ the system of marginals induced by k.

We now prove existence and uniqueness of a measure on $\Gamma_{\text{nns}}(TX)$ from the Gaussian measures defined on $V_{x_1,\ldots,x_n}$ for any $\{x_1, \ldots, x_n\} \subseteq X$. We do this by means of the general form of the Kolmogorov extension theorem formulated below. Recall again that for a measure π, and a measurable function ϕ, the *pushfoward measure* $\phi_*\pi$ is defined as $(\phi_*\pi)(A) = \pi(\phi^{-1}(A))$ for all measurable sets A.

Result 15 (Kolmogorov Extension Theorem)**.** *Let $(X_\alpha, \mathcal{B}_\alpha, \mathcal{O}_\alpha)_{\alpha\in A}$ be a family of measurable spaces, each equipped with a topology. For each finite $B \subseteq A$, let μ_B be an inner regular probability measure on $X_B = \prod_{\alpha\in B} X_\alpha$ with σ-algebra $\mathcal{B}_B$ and with the product topology $\mathcal{O}_B$ obeying*

$$(\mathrm{proj}_C)_*\mu_B = \mu_C \tag{28}$$

whenever $C \subseteq B \subseteq A$ are two nested finite subsets of A. Here projections $\mathrm{proj}_C : X_B \to X_C$ are defined by $\mathrm{proj}_C(\{x_\alpha\}_{\alpha\in B}) = \{x_\alpha\}_{\alpha\in C}$ and $(\mathrm{proj}_C)_*$ denotes the pushforward by proj_C. Then there exists a unique probability measure μ_A on $\mathcal{B}_A$ with the property that $(\mathrm{proj}_B)_*\mu_A = \mu_B$ for all finite $B \subseteq A$.

Proof. Tao [42], Theorem 2.4.3. □

By showing the existence of a probability measure on the space $\Gamma_{\text{nns}}(TX)$, one can start speaking about random variables $f : \Omega \to \Gamma_{\text{nns}}(TX)$ with said measure as their distribution: these are the Gaussian vector fields we seek. However, in order to apply the above result, we first need to verify condition (28). This is done in the following.

Proposition 16. *The family of measures $(\pi_{x_1,\dots,x_n})_{\{x_1,\dots,x_n\}\subseteq X}$ is a projective family in the sense that for any $\{x_1,\dots,x_m\} \subseteq \{x_1,\dots,x_n\} \subseteq X$, we have*

$$(\text{proj}_{x_1,\dots,x_m})_* \pi_{x_1,\dots,x_n} = \pi_{x_1,\dots,x_m} \tag{29}$$

where $\text{proj}_{x_1,\dots,x_m} : V_{x_1,\dots,x_n} \to V_{x_1,\dots,x_m}$ is the canonical projection induced by the direct sum.

Proof. Take two random variables $v_{x_1,\dots,x_n} : \Omega \to V_{x_1,\dots,x_n}$ and $v_{x_1,\dots,x_m} : \Omega \to V_{x_1,\dots,x_m}$ with $v_{x_1,\dots,x_n} \sim \pi_{x_1,\dots,x_n}$ and $v_{x_1,\dots,x_m} \sim \pi_{x_1,\dots,x_m}$. It suffices to show that for the random variable $v_{x_1,\dots,x_n} : \Omega \to V_{x_1,\dots,x_m}$ we have

$$v_{x_1,\dots,x_m} \overset{\text{d}}{=} \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n} \tag{30}$$

where $\overset{\text{d}}{=}$ denotes the equality of distributions. We first show that $\text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n}$ is Gaussian. Let $\phi \in V^*_{x_1,\dots,x_m}$ and write

$$\left\langle \phi \;\middle|\; \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n} \right\rangle = \langle (\phi, 0) \mid v_{x_1,\dots,x_n} \rangle \tag{31}$$

where $(\phi, 0) \in V^*_{x_1,\dots,x_n}$ is the natural inclusion of $\phi \in V^*_{x_1,\dots,x_m}$ in the space $V^*_{x_1,\dots,x_n}$ by padding with the zero vector over all components of the direct sum whose indices are not $x_1,\dots,x_m$. This identity holds for all vectors, hence it holds for random vectors, and $\text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n}$ is Gaussian. Now, we compute its moments: write

$$\begin{aligned}
\mathbb{E}\left\langle \phi \;\middle|\; \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n} \right\rangle &= \mathbb{E}\langle (\phi, 0) \mid v_{x_1,\dots,x_n} \rangle && (32)\\
&= \langle (\phi, 0) \mid \mu_{x_1,\dots,x_n} \rangle && (33)\\
&= \left\langle \phi \;\middle|\; \text{proj}_{x_1,\dots,x_m} \mu_{x_1,\dots,x_n} \right\rangle && (34)\\
&= \langle \phi \mid \mu_{x_1,\dots,x_m} \rangle && (35)
\end{aligned}$$

where the last line follows by definition of $\mu_{x_1,\dots,x_m}$, and

$$\begin{aligned}
&\text{Cov}(\left\langle \phi \;\middle|\; \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n} \right\rangle, \left\langle \psi \;\middle|\; \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n} \right\rangle) && (36)\\
&\quad = \text{Cov}\left(\langle (\phi, 0) \mid v_{x_1,\dots,x_n} \rangle, \langle (\psi, 0) \mid v_{x_1,\dots,x_n} \rangle\right) && (37)\\
&\quad = k_{x_1,\dots,x_n}((\phi, 0), (\psi, 0)) && (38)\\
&\quad = k_{x_1,\dots,x_m}(\phi, \psi) && (39)
\end{aligned}$$

where the last line follows by bilinearity and the definition of $k_{x_1,\dots,x_m}$.

So far we have shown that $\text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n}$ is Gaussian over $V_{x_1,\dots,x_m}$ and its mean vector and covariance form coincide with those of $v_{x_1,\dots,x_m}$. Hence, by the uniqueness part of Lemma 10 we have $v_{x_1,\dots,x_m} \overset{d}{=} \text{proj}_{x_1,\dots,x_m} v_{x_1,\dots,x_n}$. This finishes the proof. □

We are now ready to apply the Kolmogorov extension theorem to show existence of the desired distribution.

Proposition 17. *There exists a unique measure π_∞ on the infinite product space $\prod_{x\in X} T_x X$.*[7]

Proof. We apply the prior result 15. Let X be the index set, and take $(T_x X)_{x\in X}$, equipped with the standard topology and Borel σ-algebra as our measurable spaces. For each finite $\{x_1,\dots,x_n\} \subseteq X$, take $\pi_{x_1,\dots,x_n}$ as our probability measure, and note that since each $\pi_{x_1,\dots,x_n}$ is a finite measure on a finite-dimensional real vector space $V_{x_1,\dots,x_n}$, it is automatically inner regular. Moreover, the family of measures $(\pi_{x_1,\dots,x_n})_{\{x_1,\dots,x_n\}\subseteq X}$ is projective by Proposition 16. The claim follows. □

This gives our GP as a measure on an infinite Cartesian space: we now map this measure into the space of sections.

[7] Note that this is the Tychonoff product of topological spaces rather than a direct product of linear spaces.

Corollary 18. *There exists a unique measure $\pi_{\Gamma_{\mathrm{nns}}(TX)}$ on $\Gamma_{\mathrm{nns}}(TX)$ equipped with the pushforward σ-algebra.*

Proof. Define the operator $\mathcal{I} : \prod_{x \in X} T_x X \to \Gamma_{\mathrm{nns}}(TX)$ by

$$(\mathcal{I}s)(x) = (x, \mathrm{proj}_x\, s) \tag{40}$$

for all $x \in X$ and $s \in \prod_{x \in X} T_x X$. Take $\pi_{\Gamma_{\mathrm{nns}}(TX)} = \mathcal{I}_* \pi_\infty$. □

This is the probability distribution of our Gaussian process. We are now ready to define Gaussian vector fields, and show that each Gaussian vector field in turn possesses a mean vector field and cross-covariance kernel.

Definition 19. *Let X be a manifold. We say that a random vector field $f : \Omega \to \Gamma_{\mathrm{nns}}(TX)$ is* Gaussian *if for any finite set of locations $(x_1, \ldots, x_n) \in X^n$, the random vector $f(x_1), \ldots, f(x_n) \in T_{x_1}X \oplus \ldots \oplus T_{x_n}X$ is Gaussian in the sense of Definition 8.*

Definition 20. *Let $f : \Omega \to \Gamma_{\mathrm{nns}}(TX)$ be a Gaussian vector field. Define μ to be the unique vector field for which, for any $x \in X$ and any $\phi \in T_x^*$, we have that*

$$\langle \phi \mid \mu(x) \rangle = \mathbb{E}\langle \phi \mid f(x) \rangle. \tag{41}$$

Next taking an additional $x' \in X$ and $\psi \in T_{x'}^*$, define the cross-covariance kernel k by

$$k(\phi, \psi) = \mathrm{Cov}(\langle \phi \mid f(x) \rangle, \langle \psi \mid f(x') \rangle). \tag{42}$$

Summarizing, we obtain the following claim.

Theorem 21. *Every pair consisting of a mean vector field and symmetric fiberwise bilinear positive definite function $k : T^*X \times T^*X \to \mathbb{R}$, which we call a* cross-covariance kernel*, defines a unique (distribution-wise) Gaussian vector field in the sense of Definition 19. Conversely, every Gaussian vector field admits and is characterized uniquely by this pair.*

Proof. Corollary 18, Definition 19, and Definition 20. □

Embeddings (Proof of Proposition 2)

Proposition 22. *Let $\mathrm{emb} : X \to \mathbb{R}^p$ be an embedding, let f be a Gaussian vector field on X, and denote by $\boldsymbol{f}_{\mathrm{emb}} : \mathrm{emb}(X) \to \mathbb{R}^p$ its pushforward along the embedding, that is, for any $x \in X$,*

$$\boldsymbol{f}_{\mathrm{emb}}(\mathrm{emb}(x)) = \mathrm{d}_x\mathrm{emb}(f(x)), \tag{43}$$

where $\mathrm{d}_x\mathrm{emb} : T_x X \to T_{\mathrm{emb}(x)}\mathbb{R}^p$ is the differential of emb. Then $\boldsymbol{f}_{\mathrm{emb}}$ is a vector-valued Gaussian process in the standard sense.

Proof. Let $x_1, \ldots, x_n \in X^n$ be a finite set of arbitrary locations. In what follows, we use a slight abuse of notation by letting x_i denote both x_i and $\mathrm{emb}(x_i)$ for simplicity. We claim that the random vector $(\boldsymbol{f}_{\mathrm{emb}}(x_1), \ldots, \boldsymbol{f}_{\mathrm{emb}}(x_n)) \in \mathbb{R}^{np}$ is multivariate Gaussian, which is sufficient to prove our result. Since f is a Gaussian vector field, we have that

$$(f(x_1), \ldots, f(x_n)) \sim \mathrm{N}(\mu_{x_1,\ldots,x_n}, k_{x_1,\ldots,x_n}) \tag{44}$$

is a Gaussian random vector on $T_{x_1}X \oplus \ldots \oplus T_{x_n}X$. Now consider the map $\phi_{x_1,\ldots,x_n} : T_{x_1}X \oplus \ldots \oplus T_{x_n}X \to T_{\mathrm{emb}(x_1)}\mathbb{R}^p \oplus \ldots \oplus T_{\mathrm{emb}(x_n)}\mathbb{R}^p \cong \mathbb{R}^{np}$ defined as

$$\phi_{x_1,\ldots,x_n}(f_{x_1}, \ldots, f_{x_n}) = (\mathrm{d}_{x_1}\mathrm{emb}(f_{x_1}), \ldots, \mathrm{d}_{x_n}\mathrm{emb}(f_{x_n})), \tag{45}$$

for all $(f_{x_1}, \ldots, f_{x_n}) \in T_{x_1}X \oplus \ldots \oplus T_{x_n}X$, which is linear, owing to the linearity of $\mathrm{d}_x\mathrm{emb}$. Since linear maps preserve Gaussianity, it follows that the vector $\phi_{x_1,\ldots,x_n}(f(x_1), \ldots, f(x_n)) = (\boldsymbol{f}_{\mathrm{emb}}(x_1), \ldots, \boldsymbol{f}_{\mathrm{emb}}(x_n)) \in \mathbb{R}^{np}$ is multivariate Gaussian and the claim follows. □

Coordinate Expressions (Proof of Proposition 5)

We recall the definition of a *frame* on X and its dual object, namely, the *coframe*.

Definition 23. *A* frame *F on X is defined as a collection $(e_i)_{i=1}^d$ of not necessarily smooth sections of TX such that at each point $x \in X$, the vectors $(e_i(x))_{i=1}^d$ form a basis of T_xX. The corresponding* coframe *F^* is defined as a collection $(e^i)_{i=1}^d$ of not necessarily smooth sections of T^*X such that $\langle e^i(x) | e_j(x) \rangle = \delta_{ij}$ for all $x \in X$.*

Proposition 24. *Let $f : \Omega \to \Gamma_{\text{nns}}(TX)$ be a Gaussian vector field on X with cross-covariance kernel $k : T^*X \times T^*X \to \mathbb{R}$. Given a frame $F = (e_1, \dots, e_d)$ on X and $F^* = (e^1, \dots, e^d)$ be its coframe, define $f^i = \langle e^i \mid f \rangle$ for all $i = 1, \dots, d$. Then $\boldsymbol{f} = (f^1, \dots, f^d) : \Omega \times X \to \mathbb{R}^d$ is a vector-valued GP in the usual sense with matrix-valued kernel $\mathbf{K}_F : X \times X \to \mathbb{R}^{d \times d}$ given by*

$$\mathbf{K}_F(x, x') = \begin{bmatrix} k(e^1(x), e^1(x')) & \dots & k(e^1(x), e^d(x')) \\ \vdots & \ddots & \vdots \\ k(e^d(x), e^1(x')) & \dots & k(e^d(x), e^d(x')) \end{bmatrix}. \tag{46}$$

Conversely, given a vector-valued GP $\boldsymbol{f} = (f^1, \dots, f^d) : \Omega \times X \to \mathbb{R}^d$ and a frame $F = (e_1, \dots, e_d)$ on X, $f(\cdot) := \sum_{i=1}^d f^i(\cdot) e_i(\cdot)$ defines a Gaussian vector field on X.

Proof. First, we note that $f^i(x) = \langle e^i(x) \mid f(x) \rangle$ are jointly Gaussian for all $i = 1, \dots, d$ and all $x \in X$. Thus for any $x_1, \dots, x_n \in X$, the vector $(\boldsymbol{f}(x_1), \dots, \boldsymbol{f}(x_n)) \in \mathbb{R}^{n \times d}$ is multivariate Gaussian and therefore $\boldsymbol{f}$ is a vector-valued GP in the usual sense. Now for any $x, x' \in X$, the kernel of $\boldsymbol{f}$ evaluated at these points reads

$$\mathbf{K}_F(x, x') = \begin{bmatrix} \mathrm{Cov}(f^1(x), f^1(x')) & \dots & \mathrm{Cov}(f^1(x), f^d(x')) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(f^d(x), f^1(x')) & \dots & \mathrm{Cov}(f^d(x), f^d(x')) \end{bmatrix} \tag{47}$$

$$= \begin{bmatrix} k(e^1(x), e^1(x')) & \dots & k(e^1(x), e^d(x')) \\ \vdots & \ddots & \vdots \\ k(e^d(x), e^1(x')) & \dots & k(e^d(x), e^d(x')) \end{bmatrix}, \tag{48}$$

which follows from Definition 20. This concludes the first part of the proof.

To prove the converse direction, for any collection of points $x_1, \dots, x_n \in X$, define the random vector $v_{x_1,\dots,x_n} = (f(x_1), \dots, f(x_n))$, where f is given by $f(x) = \sum_{i=1}^d f^i(x) e_i(x)$. Now for any $\phi_{x_1,\dots,x_n} = (\phi_{x_1}, \dots, \phi_{x_n}) \in V^*_{x_1,\dots,x_n}$, we have

$$\langle \phi_{x_1,\dots,x_n} \mid v_{x_1,\dots,x_n} \rangle = \sum_{i=1}^n \langle \phi_{x_i} \mid f(x_i) \rangle \tag{49}$$

$$= \sum_{i=1}^n \langle \phi_{x_i} \mid \sum_{j=1}^d f^j(x_i) e_j(x_i) \rangle \tag{50}$$

$$= \sum_{i=1}^n \sum_{j=1}^d f^j(x_i) \langle \phi_{x_i} \mid e_j(x_i) \rangle. \tag{51}$$

Since $f^j(x_i)$ is univariate Gaussian for all $i = 1, \dots, n$ and $j = 1, \dots, d$, the above linear combination is univariate Gaussian and therefore $v_{x_1,\dots,x_n}$ is Gaussian in the sense of Definition 8. Since $x_1, \dots, x_n$ were chosen arbitrarily, f is a Gaussian vector field. □

Gauge Equivariance (Proof of Corollary 6)

Given two frames F, F' on X, an abstract vector $f_x \in T_xX$ has two vector representations $\boldsymbol{f}_x, \boldsymbol{f}'_x$ in the respective frames. Recall that F' is said to be obtained from F by a *gauge transformation* with respect to a matrix field $\mathbf{A} : X \to \mathrm{GL}(d, \mathbb{R})$, if

$$\boldsymbol{f}'_x = \mathbf{A}(x) \boldsymbol{f}_x \tag{52}$$

holds for all $x \in X$, and we write $F' = \mathbf{A}F$. In the following, we compute an explicit expression for the gauge-transformed frame $\mathbf{A}F$ and its coframe.

Lemma 25. *Let $F = (e_1, \ldots, e_d)$ be a frame on X, $\mathbf{A} : X \to \mathrm{GL}(d, \mathbb{R})$ be a matrix field of gauge transformations, $\mathbf{A}F = (\varepsilon_1, \ldots, \varepsilon_d)$ be the gauge transformed frame as above and let $(\mathbf{A}F)^* = (\varepsilon^1, \ldots, \varepsilon^d)$ be the corresponding coframe. Then we have the following explicit expressions*

$$\varepsilon_i(x) = \sum_{j=1}^{d} e_j(x)[\mathbf{A}^{-1}(x)]_{ji}, \qquad \varepsilon^i(x) = \sum_{j=1}^{d} [\mathbf{A}(x)]_{ij} e^j(x). \tag{53}$$

Proof. For any $x \in X$, let $f_x \in T_x X$ be an abstract vector, which has the vector representations $\boldsymbol{f}_x$ and $\mathbf{A}(x)\boldsymbol{f}_x$ in the frames F and $\mathbf{A}F$ respectively. Letting $\boldsymbol{f}_x = (f_x^1, \ldots, f_x^d)$, we have

$$f_x = \sum_{i=1}^{d} f_x^i e_i(x) = \sum_{i=1}^{d}\sum_{j=1}^{d} ([\mathbf{A}(x)]_{ji} f_x^i)\varepsilon_j(x) = \sum_{i=1}^{d} f_x^i \left(\sum_{j=1}^{d} \varepsilon_j(x)[\mathbf{A}(x)]_{ji} \right). \tag{54}$$

Thus, $e_i(x) = \sum_{j=1}^{d} \varepsilon_j(x)[\mathbf{A}(x)]_{ji}$, or identically, $\varepsilon_i(x) = \sum_{j=1}^{d} e_j(x)[\mathbf{A}^{-1}(x)]_{ji}$. We now claim that $\varepsilon^i(x) = \sum_{j=1}^{d} [\mathbf{A}(x)]_{ij} e^j(x)$, which we prove by showing that it satisfies the relation $\left\langle \varepsilon^i(x) \,\middle|\, \varepsilon_j(x) \right\rangle = \delta_{ij}$ as follows:

$$\begin{aligned}
\left\langle \varepsilon^i(x) \,\middle|\, \varepsilon_j(x) \right\rangle &= \left\langle \sum_{k=1}^{d} [\mathbf{A}(x)]_{ik} e^k(x) \,\middle|\, \sum_{l=1}^{d} e_l(x)[\mathbf{A}^{-1}(x)]_{lj} \right\rangle && (55)\\
&= \sum_{k=1}^{d}\sum_{l=1}^{d} [\mathbf{A}(x)]_{ik} \underbrace{\left\langle e^k(x) \,\middle|\, e_l(x) \right\rangle}_{\delta_{kl}} [\mathbf{A}^{-1}(x)]_{lj} && (56)\\
&= \sum_{k=1}^{d} [\mathbf{A}(x)]_{ik} [\mathbf{A}^{-1}(x)]_{kj} && (57)\\
&= \underbrace{[\mathbf{A}(x)\mathbf{A}^{-1}(x)]_{ij}}_{\delta_{ij}}. && (58)
\end{aligned}$$

This concludes the proof. □

The following is, then, straightforward to show.

Corollary 26. *Let F be a frame on X and $\mathbf{K}_F : X \times X \to \mathbb{R}^{d \times d}$ be the corresponding matrix representation of a cross-covariance kernel $k : T^*X \times T^*X \to \mathbb{R}$. This satisfies the* gauge equivariance condition

$$\mathbf{K}_{\mathbf{A}F}(x, x') = \mathbf{A}(x)\mathbf{K}_F(x, x')\mathbf{A}(x')^T, \tag{59}$$

where $\mathbf{A} : X \to \mathrm{GL}(d, \mathbb{R})$ is a gauge transformation applied to each point on X. All cross-covariance kernels in the sense of Proposition 4 arise this way.

Proof. Let $F = (e_1, \ldots, e_d)$ and $\mathbf{A}F = (\varepsilon_1, \ldots, \varepsilon_d)$. Then by the previous lemma, we have

$$\begin{aligned}
[\mathbf{K}_{\mathbf{A}F}(x, x')]_{ij} &= k(\varepsilon^i(x), \varepsilon^j(x')) && (60)\\
&= \sum_{k=1}^{d}\sum_{l=1}^{d} k\left([\mathbf{A}(x)]_{ik}\, e^k(x), [\mathbf{A}(x')]_{jl}\, e^l(x') \right) && (61)\\
&= \sum_{k=1}^{d}\sum_{l=1}^{d} [\mathbf{A}(x)]_{ik}\, [\mathbf{K}_F(x, x')]_{kl}\, [\mathbf{A}(x')]_{jl}, && (62)
\end{aligned}$$

which proves the identity (59).

The second claim is obvious: take some cross-covariance kernel in the sense of Proposition 4 and some frame—this induces a gauge equivariant matrix-valued kernel that correspond to the cross-covariance kernel in the sense of Proposition 4 from which it was constructed in the first place. □

Projected Kernels (Proof of Proposition 7)

Here we formally describe the projected kernel construction. We start by noting some properties of the projection matrices associated with differentials of isometric embeddings.

Lemma 27. *Let (X, g) be a Riemannian manifold and $\mathrm{emb} : X \to \mathbb{R}^{d'}$ be an isometric embedding. Given a frame $F = (e_1, \ldots, e_d)$ on X, denote by $\mathbf{P}_{(\cdot)} : X \to \mathbb{R}^{d \times d'}$ its associated projection matrix, defined for every x as the matrix representation of $\mathrm{d}_x\mathrm{emb}$ within F, and $\mathbf{\Gamma} : X \to \mathbb{R}^{d \times d}$, the matrix field representation of the Riemannian metric g, that is, $[\mathbf{\Gamma}(x)]_{ij} = g_x(e_i(x), e_j(x))$ for all $i, j = 1, \ldots, d$ and $x \in X$. Then we have*

$$\mathbf{P}_x \mathbf{P}_x^T = \mathbf{\Gamma}(x). \tag{63}$$

Proof. Since the embedding is isometric, for any $v, v' \in T_x X$, we have

$$g_x(u, v) = \langle \mathrm{d}_x\mathrm{emb}(v), \mathrm{d}_x\mathrm{emb}(v') \rangle, \tag{64}$$

which, in the corresponding vector representation with respect to a frame F, reads

$$\boldsymbol{v}^T \mathbf{\Gamma}(x) \boldsymbol{v}' = \left\langle \mathbf{P}_x^T \boldsymbol{v}, \mathbf{P}_x^T \boldsymbol{v}' \right\rangle = \boldsymbol{v}^T (\mathbf{P}_x \mathbf{P}_x^T) \boldsymbol{v}'. \tag{65}$$

for any $\boldsymbol{v}, \boldsymbol{v}'$. This implies that $\mathbf{\Gamma}(x) = \mathbf{P}_x \mathbf{P}_x^T$ for all x and proves the claim. □

We proceed to describe the projected kernel construction, which lets us transform a matrix-valued kernel on an ambient space into a cross-covariance kernel on the manifold.

Proposition 28. *Let (X, g) be a Riemannian manifold, $\mathrm{emb} : X \to \mathbb{R}^{d'}$ be an isometric embedding and F be a frame on X. We denote by $\mathbf{P}_{(\cdot)} : X \to \mathbb{R}^{d \times d'}$ the associated projection matrix under F, and let $\boldsymbol{f}' : X \to \mathbb{R}^{d'}$ be any vector-valued Gaussian process with matrix-valued kernel $\boldsymbol{\kappa} : X \times X \to \mathbb{R}^{d' \times d'}$. Then, the vector-valued function $\boldsymbol{f} = \mathbf{P}\boldsymbol{f}' : X \to \mathbb{R}^d$ defines a Gaussian vector field f on X using the construction in Proposition 24, whose kernel under the frame F has matrix representation*

$$\mathbf{K}_F(x, x') = \mathbf{P}_x \boldsymbol{\kappa}(x, x') \mathbf{P}_{x'}^T. \tag{66}$$

Moreover, all cross-covariance kernels $k : T^*X \times T^*X \to \mathbb{R}$ arise this way.

Proof. We demonstrate the first part by computing the covariance of $\boldsymbol{f}$. For any $x, x' \in X$, we have

$$\begin{aligned}
\mathbf{K}_F(x, x')_{ij} &= \mathrm{Cov}(f^i(x), f^j(x')) && (67)\\
&= \mathrm{Cov}([\mathbf{P}_x \boldsymbol{f}'(x)]_i, [\mathbf{P}_{x'} \boldsymbol{f}'(x')]_j) && (68)\\
&= \sum_{k=1}^{d} \sum_{l=1}^{d} \mathrm{Cov}([\mathbf{P}_x]_{ik} f'_k(x), [\mathbf{P}_{x'}]_{jl} f'_l(x')) && (69)\\
&= \sum_{k=1}^{d} \sum_{l=1}^{d} [\mathbf{P}_x]_{ik} \, \mathrm{Cov}(f'_k(x), f'_l(x')) \, [\mathbf{P}_{x'}]_{jl} && (70)\\
&= \sum_{k=1}^{d} \sum_{l=1}^{d} [\mathbf{P}_x]_{ik} \, \boldsymbol{\kappa}(x, x')_{kl} \, [\mathbf{P}_{x'}]_{jl}, && (71)
\end{aligned}$$

which proves the identity (66).

Conversely, let $k : T^*X \times T^*X \to \mathbb{R}$ be a cross-covariance kernel. We first construct a matrix-valued kernel $\mathbf{K}_F$ as in Proposition 24. Define

$$\boldsymbol{\kappa}(x, x') = \mathbf{P}_x^T \mathbf{K}_{\mathbf{\Gamma}^{-1}F}(x, x') \mathbf{P}_{x'}, \tag{72}$$

where $\mathbf{\Gamma} : X \to \mathbb{R}^{d \times d}$ is the matrix field representation of the metric g as given in the statement of Lemma 27. Then by the same lemma, we have

$$\begin{aligned}
\mathbf{P}_x \boldsymbol{\kappa}(x, x') \mathbf{P}_{x'}^T &= (\mathbf{P}_x \mathbf{P}_x^T) \mathbf{K}_{\mathbf{\Gamma}^{-1}F}(x, x') (\mathbf{P}_{x'} \mathbf{P}_{x'}^T) && (73)\\
&= \mathbf{\Gamma}(x) \mathbf{K}_{\mathbf{\Gamma}^{-1}F}(x, x') \mathbf{\Gamma}(x') && (74)\\
&= \mathbf{K}_F(x, x'), && (75)
\end{aligned}$$

where we used that $\mathbf{\Gamma}(x)^T = \mathbf{\Gamma}(x)$ and Corollary 26 to deduce the last equality. Thus, any cross-covariance kernel k can be obtained from a matrix-valued kernel $\boldsymbol{\kappa}$ on the ambient space and therefore we do not lose any generality by working with the latter. □

B Experimental details

Here, we include further details about the experiments conducted in Section 4. All experiments were conducted on a single workstation with 64GB RAM, using CPU-based computation.

Fourier features for product kernels

Throughout this paper we use the sparse GP formulation of Wilson et al. [46, 47] to work with GPs. In order to apply this method we need to be able to sample a Fourier feature approximation of the kernel. For stationary kernels supported on Euclidean space one typically uses a random Fourier feature (RFF) approximation [33]

$$\widetilde{f}(\cdot) = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} w_i \phi_i(x), \qquad w_i \sim \mathrm{N}(0, 1), \tag{76}$$

where the ϕ_i are Fourier basis functions sampled from the spectral density of the kernel—see Sutherland and Schneider [41] for details. The resulting random function $\widetilde{f}(\cdot)$ is then a Gaussian process with zero mean and kernel $l^{-1}\mathbf{\Phi}(\cdot)^T\mathbf{\Phi}(\cdot)$, where $\mathbf{\Phi}$ is a vector of the l basis functions. This approximates the true GP with a dimension-free error of the order $l^{-1/2}$.

For kernels supported on compact spaces we use a Karhunen–Loéve (KL) expansion. If we have a Gaussian process $f(\cdot)$ on a compact space, then we can optimally approximate this function (in terms of L^2-norm) by truncating its KL expansion

$$f(\cdot) = \sum_{i=1}^{\infty} w_i \psi_i(x) \qquad w_i \sim \mathrm{N}(0, \lambda_i) \tag{77}$$

where ϕ_i, λ_i are the ith eigenfunctions and values of the kernel, $\int_X \psi(x)k(x, \cdot)\,\mathrm{d}x = \lambda_i \psi_i(\cdot)$, sorted in descending order of the eingenvalues. For the squared exponential and Matérn kernels on compact manifolds, these eigenfunctions are the eigenfunctions of the Laplacian of the manifold, and the eigenvalues are given by a transformation of the Laplacian eigenvalues [2].

The question then arises of what to do in the case of a product of kernels, each taking as input some different space, where some are suited to RFF approximation, and some to a KL approximation. We propose the following approach.

1. All the RFF-appropriate kernels can be combined into one approximation by sampling the basis functions from the product measure of their Fourier transforms.
2. All of the KL-appropriate kernels can be combined into one approximation by computing the k largest eigenvalues of the product manifold the kernels are defined on. If we have two compact manifolds with eigenvalue-function pairs $(\alpha_i, f_i(\cdot))_{i=1}^{\infty}$ and $(\beta_j, g_j(\cdot))_{j=1}^{\infty}$, then the eigenvalue-function pairs on the product manifold are $(\alpha_i + \beta_j, f_i(\cdot)g_j(\cdot))_{i,j=1}^{\infty,\infty}$ [4]. We can repeatedly apply this to find the approximation for the kernel on arbitrary products of compact manifolds.
3. Define the Fourier feature approximation of the combination of this RFF and KL approximations as

$$f(e, m) = \frac{1}{\sqrt{l}} \sum_{i=1}^{l} \sum_{j=1}^{k} w_{i,j} \phi_i(e) \psi_j(m) \qquad w_{i,j} \sim \mathcal{N}(0, \lambda_j) \tag{78}$$

   where e, m are the inputs to the RFF and KL appropriate kernels respectively, ϕ_i are the basis functions of the Euclidean kernels sampled from the product measure, and λ_j, ψ_j, are the product eigenpairs on the product manifold. In the limit of infinite basis functions in both l and k this will give the correct kernel, and therefore the true prior.

Dynamics experiment

In this experiment, the base manifold is the state space of the single pendulum system. The position of the pendulum is represented by a single angle in $[0, 2\pi)$, which corresponds to the circle $\mathbb{S}^1$. The

momentum lies then in its respective cotangent space. The phase space is the product of these, $\mathbb{S}^1 \times \mathbb{R}^1$.

This product manifold naturally embeds into $\mathbb{R}^3$ by embedding the circle into $\mathbb{R}^2$ in the canonical way, and leaving $\mathbb{R}^1$ unchanged. The embedding is then

$$\mathrm{emb}(q, p) = (\cos q, \sin q, p), \tag{79}$$

where q is the position and p is the angular momentum. The global projection matrix given by

$$\mathbf{P}_{q,p} = \begin{bmatrix} -\sin q & \cos q & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{80}$$

The Euclidean vector kernel we use is a separable kernel, produced by taking the product of an intrinsic squared exponential manifold kernel with the identity matrix to give a matrix valued kernel, $\boldsymbol{\kappa} = k_{\mathbb{S}^1 \times \mathbb{R}^1} \mathbf{I}_{3\times 3}$. The intrinsic manifold kernel is produced by the product of a typical Euclidean squared exponential kernel with a squared exponential kernel defined on $\mathbb{S}^1$ by Borovitskiy et al. [2], so that $k_{\mathbb{S}^1 \times \mathbb{R}^1} = k_{\mathbb{S}^1} k_{\mathbb{R}^1}$. The length scales of these kernels are set to 0.3 and 1.2 respectively, and the amplitude set to give $k(x, x) = 1$.

To learn the dynamics, we initialise the system at two start points, and evolve the system using leapfrog integration. From these observations of position, we backward Euler integrate the momentum of the system, $p_i = \frac{1}{2} ml(q_{i+1} - q_i)$, and from these position-momentum trajectories we estimate observations of the dynamics field

$$\nabla_t (q, p)_i = \left(\frac{q_{i+1} - q_i}{h}, \frac{p_{i+1} - p_i}{h} \right) \tag{81}$$

where $h = 0.01$ is the step size. Using these observations, we condition a sparse GP using all the data using the analytic expression for the sparse posterior kernel matrix. The result is an estimate of the system dynamics with suitable uncertainty estimates. In order to compute rollouts of these dynamics, we follow Wilson et al. [46, 47] and employ linear-time pathwise sampling of this sparse GP together with leapfrog integration [16].

Wind interpolation experiment

In this experiment, the base manifold is the sphere $\mathbb{S}^2$, which we embed naturally in $\mathbb{R}^3$ as

$$\mathrm{emb}(\phi, \theta) = (\cos\theta \sin\phi,\ \sin\theta \sin\phi,\ \cos\phi), \tag{82}$$

where we used spherical coordinates $\phi \in (0, \pi), \theta \in [0, 2\pi)$ to parametrise the sphere

$$(\phi, \theta) \in \{(0, 0)\} \cup \{(\pi, 0)\} \cup (0, \pi) \times [0, 2\pi) \tag{83}$$

We choose a frame $F = (e_1, e_2)$, where $e_1(\phi, \theta) = \hat{\phi}$ and $e_2(\phi, \theta) = \hat{\theta}$ are the unit vectors in the ϕ, θ directions respectively for all $\phi \in (0, \pi), \theta \in (0, 2\pi)$. The choice of points on the North and South poles determines the choice of gauge at these points. The corresponding orthonormal projection matrix reads

$$\mathbf{P}_{\phi,\theta} = \begin{bmatrix} \cos\theta\cos\phi & \sin\theta\cos\phi & -\sin\phi \\ -\sin\theta & \cos\theta & 0 \end{bmatrix}, \tag{84}$$

for all points, with the choice of $\theta = 0$ giving the choice of frame at the poles.

For the data, we used the following publicly available data sets.

- The ERA5 atmospheric reanalysis data. In particular, the variables 10M-U-COMPONENT-OF-WIND and 10M-V-COMPONENT-OF-WIND from the REANALYSIS-ERA5-SINGLE-LEVELS dataset for the date 01/01/2019 09:00-10:00, regridded from 0.25° to 5.625° resolution using python's XESMF package.
- The WeatherBench dataset [35], which can be found at https://github.com/pangeo-data/WeatherBench. In particular the variables 10M-U-COMPONENT-OF-WIND and 10M-V-COMPONENT-OF-WIND at 5.625° resolution for the entire available period 1979/01/01 - 2018/12/31.

- The Aeolus trajectory data, which can be read using Python's SKYFIELD API from Aeolus' two-line element set given below.
  1 43600U 18066A 21153.73585495 .00031128 00000-0 12124-3 0 9990
  2 43600 96.7150 160.8035 0006915 90.4181 269.7884 15.87015039160910

Instead of using actual observations from the Aeolus satellite, we generated our own by interpolating the ERA5 data along the satellite track, whose locations are available minutely. This is so that we can compare the predictions against the ground truth to assess the performance. We use one hour of data, and hence 60 data points, to perform a spatial interpolation instead of a space-time interpolation, which is reasonable as the atmosphere hardly moves during that time period at the spatial scale of interest. Moreover, we include the weekly climatology as prior information (computed by taking the temporal average of historical global wind patterns for each of the 52 calendar weeks during the period 1979-2018 in WeatherBench), which captures general circulation patterns such as trade winds in the poles and the equator. This is equivalent to training the GP on the difference of the wind velocity from the weekly climatology.

For the kernel, we used Matérn-3/2 on the sphere and the Euclidean space (see Borovitskiy et al. [2] for the construction of Matérn kernels on the sphere), where the prior amplitude parameter was set to a fixed value (11.5 in the spherical case and 2.2 in the Euclidean case) and the length scale parameter was learnt from data. We have tried to learn the length scale initially by fitting the GP on the satellite observations and maximizing the marginal likelihood. However, this gave an unrealistically small value, likely due to the observations being too sparse: so, instead, we first trained a sparse GP on 150 randomly chosen time slices of the weatherbench historical wind reanalysis data and minimizing the Kullback–Leibler divergence of the variational distribution from the posterior (using the Adam optimizer with learning rate `1e-2`). The mean of the learnt length scales of the 150 samples was then used as the final length scale. Denoting by $k_{\mathbb{S}^2}$ the scalar Matérn-3/2 kernel on the sphere, we construct a matrix-valued kernel on the ambient space $\mathbb{R}^3$ by taking $\boldsymbol{\kappa} = k_{\mathbb{S}^2}\mathbf{I}_{3\times 3}$ as in the dynamics experiment, which is then used to construct the projected kernel with the projection given by (84). Finally, we note that when fitting the GP on the satellite observations, we use an observation error of 1.7m/s, which reflects the sum of the random and systematic error in the Aeolus satellite, as detailed by its technical specifications [9].