11696C

Gaussian Process Regression Constrained by Boundary Value Problems

Mamikon Gulian[1], Ari Frankel[2], Laura Swiler[3]

[1] Quantiative Modeling and Analysis, Sandia National Laboratories, Livermore, CA
[2] Computational Science and Analysis, Sandia National Laboratories, Livermore, CA
[3] Center for Computing Research, Sandia National Laboratories, Albuquerque, NM

Exceptional service in the national interest

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND number:

Introduction & Summary

- Gaussian process regression (GPR) is a widely used Bayesian technique for inference in scientific applications with limited scattered data.
- Several physical processes are described by a well-posed boundary value problem (BVP) of the form

$$\begin{cases} Lu(x) = f(x), & x \in \Omega, \\ \mathcal{B}u(x) = g(x), & x \in \partial\Omega, \end{cases} \tag{1}$$

  where L denotes a linear partial differential operator, Ω a domain with boundary $\partial\Omega$, and $\mathcal{B}$ a general mixed boundary operator.
- We develop a framework for Gaussian processes regression constrained by boundary value problems, which can infer the BVP solution when only scattered observations of the source term are available.
- The framework benefits from a reduced-rank property of covariance matrices, so it scales well to large data regimes.
- We demonstrate more accurate and stable solution inference as compared to physics-informed (PDE-only) Gaussian process regression without BCs.

But first: a brief survey of constrained GPR

- Since we'll combine two types of constraints, let's start with a survey of the evolving field of constrained GPR.
- Why constrained GPR? In many scientific applications a large amount of data may not be available for training.
- Unlike data from internet or text searches, computational and physical experiments are typically extremely expensive.
- Moreover, even if ample data exists, the machine learning model may yield behaviors that are inconsistent with what is expected physically when queried in an extrapolatory regime.
- To aid and improve the process of building machine learning models for scientific applications, it is desirable to have a framework that allows the incorporation of physical principles and other a priori information to supplement the limited data and regularize the behavior of the model.

Basics of GPR: prior and likelihood

- In GPR, a function of interest $u(x)$ is modeled by a Gaussian process with a given mean function $m(x)$ and covariance function given by $K(x, x') = \mathrm{Cov}(u(x), u(x'))$:
$$u \sim \mathcal{GP}(m, K). \tag{2}$$
- That is, the vector of values $u(X)$ over a finite collection of locations X has a multivariate normal density
$$u(X) \sim \mathcal{N}(m(X), K(X, X)), \tag{3}$$
where $m(X)$ is a vector of mean values of u and $K(X, X)$ is the covariance matrix between the values.
- One common choice of the covariance function is the squared-exponential kernel given by
$$K(x, x') = s^2 \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) \tag{4}$$
where s^2 and ℓ^2 are magnitude and length-scale parameters that control the behavior of the covariance function, i.e., the hyperparameters.
- We assume that data or observations y at the X locations are contaminated by independently and identically distributed Gaussian noise with variance σ^2, giving a likelihood function
$$p(y|u, X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - u_i(X_i))^2}{2\sigma^2}\right). \tag{5}$$

Basics of GPR: posterior prediction

- Gaussian process regression proceeds by invoking Bayes' rule to compute the posterior distribution of f as

$$p(u|y,X) = \frac{p(y|u,X)p(u|X)}{p(y|X)}, \tag{6}$$

with log-marginal-likelihood

$$\begin{aligned} \log p(y|X) &= \int p(y|u,X)p(u|X)du \\ &= -\frac{1}{2}y^\top (K(X,X)+\sigma^2 I_N)^{-1} y - \frac{1}{2}\log|K(X,X)+\sigma^2 I_N| - \frac{N}{2}\log 2\pi, \end{aligned} \tag{7}$$

using the prior (3) and the Gaussian likelihood (5).

- Here, I_N denotes the identity matrix of size $N \times N$. The predictive distribution for $u^* = u(x^*)$ at a new point x^* is a Gaussian with mean

$$\mathbb{E}[u^*] = K(x^*,X)(K(X,X)+\sigma^2 I_N)^{-1} y \tag{8}$$

and variance

$$\mathrm{Var}[u^*] = K(x^*,x^*) - K(x^*,X)(K(X,X)+\sigma^2 I_N)^{-1} K(X,x^*). \tag{9}$$

- The most common way to obtain hyperparameters to use maximum likelihood optimization of the log-marginal-likelihood with respect to the covariance hyperparameters.

GPR: A Complete Example

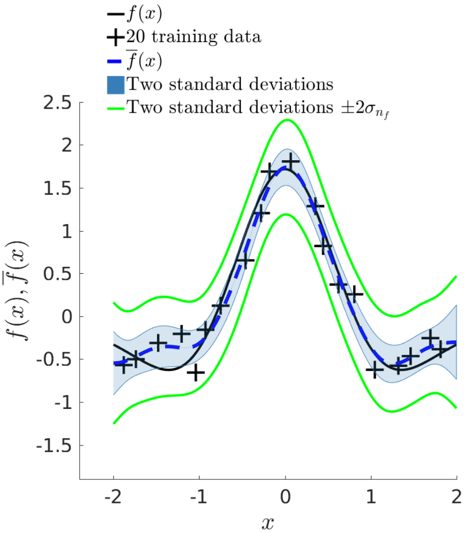


Figure: Noise is added to some locations on the black curve to generate data (black crosses).
GPR fits a mean posterior to the data after filtering out some noise with a Gaussian likelihood, with the posterior variance giving an esimate of uncertainty in the prediction.
The Gaussian likelihood allows us to infer white noise in the data.

Strategies & Differences to look for

- Each step of GPR – sample space/prior, likelihood, posterior - gives opportunities to enforce constraints.
- The difficulty with applying constraints to a GP is that a constraint typically calls for a condition to hold *globally* – that is, for *all* points x in an interval I – for all realizations or predictions of the process.
- *A priori*, this amounts to an infinite set of point constraints for an infinite dimensional sample space of functions.
- This raises a numerical feasibility issue, which each method circumvents in some way.
- Some methods relax the global constraints to constraints at a finite set of "virtual" points.
- Other methods transform the output of the GP to guarantee the predictions satisfy constraints,
- Further methods construct a sample space of predictions in which every realization satisfies the constraints.
- These distinctions should be kept in mind when surveying constrained GPs.

Bound constraints: warping functions and non-Gaussian likelihoods

- Bound constraints of the form $a \leq f(\mathbf{x}) \leq b$ over some region of interest arise naturally in many applications, such as chemical concentration data.
- **Warping functions** can be used to transform bounded observations z_i to unbounded observations u_i which can be treated with unconstrained GPR, then transformed back (Jensen et al.).
- E.g., the probit function (the inverse of the CDF Φ of a standard normal random variable) transforms bounded values $z \in [0, 1]$ to unbounded values $u \in (-\infty, \infty)$ via $u = \Phi^{-1}(z_i)$.
- In addition to using warping functions, bound constraints can also be enforced using **non-Gaussian likelihood functions** $p(\mathbf{y}|X, \mathbf{f}, \theta)$ that are constructed to produce GP observations which satisfy the constraints (Jensen et al.).
- There are a number of parametric distribution functions with finite support that can be used for the likelihood function to constrain the GP model, such as the truncated Gaussian or the beta distribution.
- Other approaches involve **truncated MVNs** and **spline expansions**.

Bound constraints via spline expansions

- Assume that a 1D process being modeled is restricted to the domain [0,1]. Let $h(x)$ be the standard tent function, i.e., the piecewise linear spline function defined by

$$h(x) = \max(1 - |x|, 0) \tag{10}$$

  and define the locations of the knots to be $x_i = i/M$ for $i = 0, 1, ...M$, with $M + 1$ total spline functions.

- For any set of spline basis coefficients ξ_i, the function representation is given by

$$f(x) = \sum_{i=0}^{M} \xi_i h(M(x - x_i)) = \sum_{i=0}^{M} \xi_i h_i(x). \tag{11}$$

  This function representation gives a C^0 piecewise linear interpolant of the point values (x_i, ξ_i) for all $i = 0, 1, ..., M$.

- $a \leq f(x) \leq b$ if $a \leq \xi_i \leq b$ – a finite-dimensional constraint.

- Suppose we are given a set of N data points at unique locations (x_j, y_j). Define the matrix A such that

$$A_{ij} = h_i(x_j). \tag{12}$$

Bound constraints via spline expansions

- Then any set of spline coefficients $\boldsymbol{\xi}$ that satisfy the equation

$$A\boldsymbol{\xi} = \mathbf{y} \tag{13}$$

  will interpolate the data exactly. Solutions to this system of equations will exist only if the rank of A is greater than N.

- We now assume the knot values $\boldsymbol{\xi}$ to be governed by a Gaussian process with covariance function K.

- **Because a linear function of a GP is also a GP**, the values of $\boldsymbol{\xi}$ and $\mathbf{y}$ are governed jointly by a GP prior in the form

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\xi} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} AKA^\top & KA^\top \\ AK & K \end{bmatrix} \right) \tag{14}$$

  where each entry of the covariance matrix is understood to be a matrix.

Linear PDE Constraints via co-kriging or block covariance approach

- Gaussian processes may be constrained to satisfy linear operator constraints of the form

$$\mathcal{L}u = f \tag{15}$$

given data on f and u. When $\mathcal{L}$ is a linear partial differential operator of the form

$$\mathcal{L} = \sum_{\boldsymbol{\alpha}} C_{\boldsymbol{\alpha}}(\mathbf{x}) \frac{\partial^{\boldsymbol{\alpha}}}{\partial \mathbf{x}^{\boldsymbol{\alpha}}}, \quad \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d), \quad \frac{\partial^{\boldsymbol{\alpha}}}{\partial \mathbf{x}^{\boldsymbol{\alpha}}} = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_3^{\alpha_d}}, \tag{16}$$

the equation (15) can be used to constrain GP predictions to satisfy known physical laws expressed as linear partial differential equations.

- If $u(\mathbf{x})$ is a GP with mean function $m(\mathbf{x})$ and covariance kernel $k(\mathbf{x}, \mathbf{x}')$,

$$u \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{17}$$

and if $m(\cdot)$ and $k(\cdot, \mathbf{x}')$ belong to the domain of $\mathcal{L}$, then $\mathcal{L}_{\mathbf{x}} \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}')$ defines a valid covariance kernel for a GP with mean function $\mathcal{L}_{\mathbf{x}} m(\mathbf{x})$. This Gaussian process is denoted $\mathcal{L}u$:

$$\mathcal{L}u \sim \mathcal{GP}(\mathcal{L}_{\mathbf{x}} m(\mathbf{x}), \mathcal{L}_{\mathbf{x}} \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}')). \tag{18}$$

Linear PDE Constraints via co-kriging or block covariance approach

- The notation "$\mathcal{L}u$" for the GP $\mathcal{GP}(\mathcal{L}_{\mathbf{x}}m(\mathbf{x}), \mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}'))$ is suggested by noting that if one could apply $\mathcal{L}$ to the samples of the GP u, then the mean of the resulting stochastic process $\mathcal{L}[u]$ would indeed be given by

$$\operatorname{mean}\left(\mathcal{L}[u](\mathbf{x})\right) = \mathbb{E}\left[\mathcal{L}[u](\mathbf{x})\right] = \mathcal{L}\mathbb{E}\left[u(\mathbf{x})\right] = \mathcal{L}m(\mathbf{x}). \tag{19}$$

- The covariance would be given by

$$\begin{aligned} \operatorname{cov}\left(\mathcal{L}[u](\mathbf{x}), \mathcal{L}[u](\mathbf{x}')\right) &= \mathbb{E}\left[\mathcal{L}_{\mathbf{x}}[u(\mathbf{x})]\mathcal{L}_{\mathbf{x}'}[u(\mathbf{x}')]\right] \\ &= \mathbb{E}\left[\mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}\left[u(\mathbf{x})u(\mathbf{x}')\right]\right] \\ &= \mathcal{L}_{\mathbf{x}}\mathbb{E}\left[\mathcal{L}_{\mathbf{x}'}\left[u(\mathbf{x})u(\mathbf{x}')\right]\right] \\ &= \mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}\mathbb{E}\left[u(\mathbf{x})u(\mathbf{x}')\right] \\ &= \mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}\left[\operatorname{cov}\left(u(\mathbf{x}), u(\mathbf{x}')\right)\right] \\ &= \mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}'). \end{aligned} \tag{20}$$

- This justification is formal, as in general the samples of the process $\mathcal{L}u \sim \mathcal{GP}(\mathcal{L}_{\mathbf{x}}m(\mathbf{x}), \mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}'))$ cannot be identified as $\mathcal{L}$ applied to the samples of u.

Linear PDE Constraints via co-kriging or block covariance approach

- If scattered measurements $\mathbf{y}_f$ on the source term f in (15) are available at domain points X_f, then this can be used to train and obtain predictions for $\mathcal{L}u$ in the standard way.

- If, in addition, measurements $\mathbf{y}_u$ of u are available at domain points X_u a GP co-kriging procedure can be used, forming the joint Gaussian process $[u; f]$.

- Given the covariance kernel $k(\mathbf{x}, \mathbf{x}')$ for u, the covariance kernel of this joint GP is

$$k\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}, \begin{bmatrix}\mathbf{x}_1'\\ \mathbf{x}_2'\end{bmatrix}\right) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1') & \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}_1, \mathbf{x}_2') \\ \mathcal{L}_{\mathbf{x}} k(\mathbf{x}_2, \mathbf{x}_1') & \mathcal{L}_{\mathbf{x}} \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}_2, \mathbf{x}_2') \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}. \tag{21}$$

- In this notation, the joint Gaussian process for $[u; f]$ is then

$$\begin{bmatrix} u(X_1) \\ f(X_2) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} m(X_1) \\ \mathcal{L}m(X_2) \end{bmatrix}, \begin{bmatrix} K_{11}(X_1, X_1) & K_{12}(X_1, X_2) \\ K_{21}(X_2, X_1) & K_{22}(X_2, X_2) \end{bmatrix} \right), \tag{22}$$

Linear PDE Example

Comparison of unconstrained and PDE constrained GP. The PDE is $-1 = d^2u/dx^2$ on the interval $[0, 1]$. Data is generated from sampling the solution $u = \frac{1}{8}[(2x-1)^2 - 1]$.

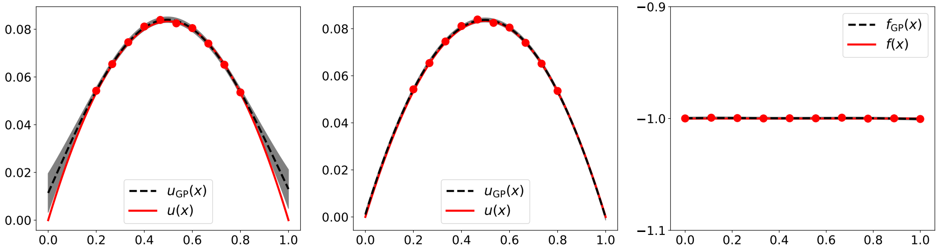


Figure: **Left:** Reconstruction of u (red line) with an unconstrained GP (black line) using 10 data points (red dots) in $[0.2, 0.8]$. **Center:** Reconstruction of u (red line) with a PDE constrained GP (black line) using the same 10 data points (red dots) in $[0.2, 0.8]$. **Right:** Right-hand side f of the PDE, with 10 additional data points in $[0, 1]$ used for the PDE constraint. Note the improved accuracy of the constrained GP outside $[0.2, 0.8]$ due to this constraint data.

Monotonicity and convexity: exploiting linearity and bound constraints

- Roughly speaking, given a method to enforce bound constraints, monotonicity constraints can be enforced by utilizing this method to enforce $\mathbf{f}' \geq \mathbf{0}$ on the derivative of the Gaussian process in a "co-kriging" setup for the joint GP $[\mathbf{f}; \mathbf{f}']$.
- Since monotonicity constraints are positivity (bound) constraints on the derivative part of such a joint GP, the "co-kriging" setup can be combined with methods for bound constraints to implement monotonicity constraints.
- The spline approach and truncated multivariate normal approach we reviewed for bound constraints have both been applied to monotonicity constraints.
- The story is similar for convexity constraints in one dimension, which can be expressed as $f'' \geq 0$, but more complicated in higher dimensions, where convexity becomes a *nonlinear* constraint between the second partials of a GP.

Curl-free and div-free constraints for vector-valued GPs: exploting linearity again

- Curl-free and divergence-free vector-valued GPR was developed by Narcowich & Ward and Fusilier Jr.
- Curl-free constraint $\mathcal{L}_{\mathbf{x}} f = \nabla \times \mathbf{f} = 0$ for $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^3$; $\mathbf{f}$ can be written $\mathbf{f} = \nabla g$.
- Divergence-free constraint $\nabla \cdot \mathbf{f} = 0$ for $\mathbf{f}$; $\mathbf{f}$ can be written $\mathbf{f} = \nabla \times \mathbf{g}$.
- Putting a GP prior on $\mathbf{g}$ with a square-exponential covariance kernel, curl-free and div-free covariance kernels for the GP $\mathbf{f}$ can be derived analytically.

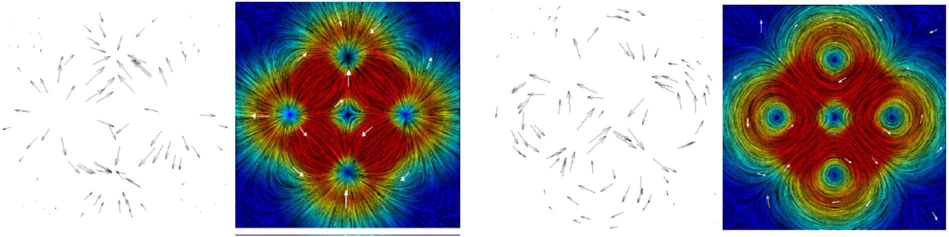


Figure: Curl-free (left) and div-free (right) GP vector field regression, from Macedo and Castro.

Boundary Value Constraints

- In many experimental setups, measurements can be taken at the boundaries of a system in a cheap and non-invasive way that permits nearly complete knowledge of the boundary values.
- The work of Solin et al. introduced a method based on the spectral expansion of a desired stationary isotropic covariance kernel $k(\mathbf{x}, \mathbf{x}') = k(|\mathbf{x} - \mathbf{x}'|)$ in eigenfunctions of the Laplacian.
- For enforcing zero Dirichlet boundary values on a domain Ω, we use the *spectral density* (Fourier transform) of the kernel,

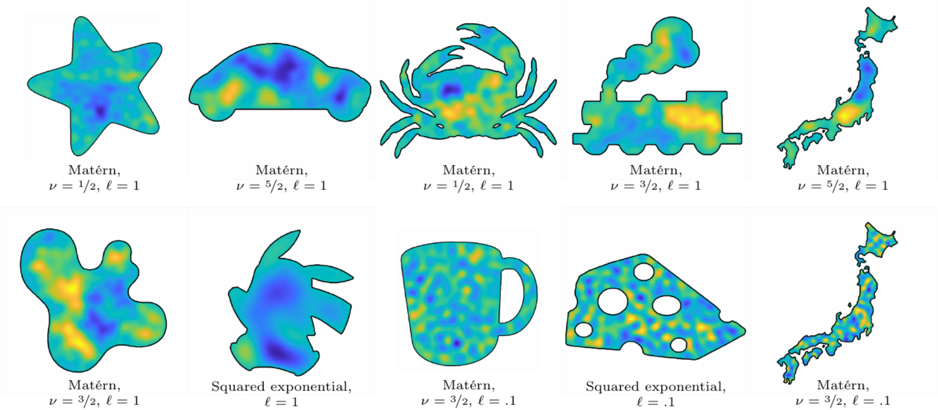
$$s(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}\cdot\mathbf{x}} k(|\mathbf{x}|) d\mathbf{x}. \tag{23}$$

- This enters into the approximation of the kernel:

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{\ell=1}^{m} s(\lambda_\ell) \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}'), \tag{24}$$

where λ_j and ϕ_j are the Dirichlet eigenvalues and eigenfunctions, respectively, of the Laplacian on the domain Ω.

Samples drawn from GPs with zero Dirichelt boundary values based on Matérn kernels



Matérn, $\nu = 1/2$, $\ell = 1$

Matérn, $\nu = 5/2$, $\ell = 1$

Matérn, $\nu = 1/2$, $\ell = 1$

Matérn, $\nu = 3/2$, $\ell = 1$

Matérn, $\nu = 5/2$, $\ell = 1$

Matérn, $\nu = 3/2$, $\ell = 1$

Squared exponential, $\ell = 1$

Matérn, $\nu = 3/2$, $\ell = .1$

Squared exponential, $\ell = .1$

Matérn, $\nu = 3/2$, $\ell = .1$

Figure 1: Random draws from Gaussian process priors constrained to 2D domains of various shapes. The process goes to zero at the boundary (black line). The approach allows for non-convex and disconnected spaces. For each domain, a random draw from a GP is shown and the assigned covariance function is shown next to the domain. The scales are arbitrary and the color map is the same as in Fig. 3.

Figure: From Solin and Kok, “Know your boundaries” (2019).

Boundary Value Constraints

- s is available in closed form for many stationary kernels, such as the squared exponential (SE) and Matérn (M_ν) kernels.
- Given n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the covariance matrix is approximated using (24) as

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \approx \sum_{\ell=1}^{m} \phi_\ell(\mathbf{x}_i) s(\lambda_\ell) \phi_\ell(\mathbf{x}_j). \tag{25}$$

- Introducing the $n \times m$ matrix Φ,

$$\Phi_{i\ell} = \phi_\ell(\mathbf{x}_i), \quad 1 \le i \le n, \quad 1 \le \ell \le m, \tag{26}$$

  and the $m \times m$ matrix $\Lambda = \mathrm{diag}(s(\lambda_\ell)), 1 \le \ell \le m$, this can be written

$$K \approx \Phi \Lambda \Phi^\top. \tag{27}$$

Boundary Value Constraints

- Thus, the covariance matrix K is diagonalized and, for a point $\mathbf{x}^*$, we can write the $n \times 1$ vector

$$\mathbf{k}_* = [k(\mathbf{x}^*, \mathbf{x}_i)]_{i=1}^n \approx \left[\sum_{\ell=1}^m \phi_\ell(\mathbf{x}_i) s(\lambda_\ell) \phi_\ell(\mathbf{x}^*)\right]_{i=1}^n = \Phi\Lambda\mathbf{\Phi}_*, \tag{28}$$

  where the $m \times 1$ vector $\mathbf{\Phi}_*$ is defined by

$$[\mathbf{\Phi}_*]_\ell = \phi_\ell(\mathbf{x}^*), \quad 1 \leq \ell \leq m. \tag{29}$$

- The Woodbury formula can be used to obtain the following expressions for the posterior mean and variance over a point $\mathbf{x}^*$ given a Gaussian likelihood $y_i = f(x_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}^*)] &= \mathbf{k}_*^\top (K + \sigma^2 I)^{-1} \mathbf{y} \\ &= \mathbf{\Phi}_*^\top (\Phi^\top \Phi + \sigma^2 \Lambda^{-1})^{-1} \Phi^\top \mathbf{y}. \\ \mathbb{V}[f(\mathbf{x}^*)] &= k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_*^\top (K + \sigma^2 I)^{-1} \mathbf{k}_* \\ &= \sigma^2 \mathbf{\Phi}_*^\top (\Phi^\top \Phi + \sigma^2 \Lambda^{-1})^{-1} \mathbf{\Phi}_*. \end{aligned} \tag{30}$$

Background

- The work of Raissi et al. studied linear differential equation constraints of the form $Lu(x) = f(x)$ for GPR of a function $u(x)$ through a "co-kriging" setup when scattered observations of $u(x)$ and the forcing term $f(x)$ were available, extending the approach of Graepel which considered the case of observations of f only.
- Solin and Kok demonstrated that zero Dirichlet boundary values can be enforced in GPR by using a covariance kernel expanded in the Dirichlet eigenfunctions of the Laplacian. Rather than merely adding scattered observations of the boundary values, they obtained a noiseless, global enforcement of the boundary condition over $\partial\Omega$.
- We combine such covariance kernels for boundary conditions with the differential equation constraints of Raissi et al. within Ω to obtain a GPR model constrained by a full, well-posed BVP.
- We also considering general mixed boundary conditions, such as Dirichlet conditions in certain regions of $\partial\Omega$ and Neumann conditions in other regions.

PDE-constrained GPR

- If $u \sim \mathcal{GP}(m(x), k(x, x'))$ and $Lu = f$ for a linear operator L, and if $m(\cdot), k(\cdot, x') \in \operatorname{dom}(L)$ then $L_x L_{x'} k(x, x')$ defines a valid covariance kernel for a GP with mean function $L_x m(x)$. This Gaussian process is denoted Lu:

$$Lu \sim \mathcal{GP}(L_x m(x), L_x L_{x'} k(x, x')). \tag{31}$$

- The PDE-constrained co-kriging procedure requires forming the joint Gaussian process $[u(x_1); f(x_2)]$. The covariance kernel of this joint GP is

$$k\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}\right) = \begin{bmatrix} k(x_1, x_1') & L_{x'} k(x_1, x_2') \\ L_x k(x_2, x_1') & L_x L_{x'} k(x_2, x_2') \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}. \tag{32}$$

- The joint Gaussian process for $[u; f]$ is then

$$\begin{bmatrix} u(x_1) \\ f(x_2) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} m(x_1) \\ Lm(x_2) \end{bmatrix}, \begin{bmatrix} K_{11}(x_1, x_1) & K_{12}(x_1, x_2) \\ K_{21}(x_2, x_1) & K_{22}(x_2, x_2) \end{bmatrix}\right), \tag{33}$$

where $K_{12}(x_1, x_2) = [K_{21}(x_2, x_1)]^\top$. Given N_u observations (X_u, y_u) of u and N_f observations (X_f, y_f) of f, GPR for $[u; f]$ can be performed to improve accuracy of predictions for u.

GPR with boundary conditions: spectral expansion covariance kernels

- The posterior mean prediction (8) for u, given data $(X, y) = \{(x_i, y_i)\}_{i=1}^N$, can be written as

$$\mathbb{E}[u(x)] = \sum_{i=1}^{N} c_i k(x, x_i), \tag{34}$$

  for coefficients $c_i \in \mathbb{R}^d$ that depend on k, the hyperparameters, and the data (X, y).

- The spectral theory of elliptic operators provides a variety of conditions under which the solution of an elliptic BVP can be expanded in orthonormal eigenfunctions defined by

$$\begin{cases} L\phi_n(x) = \lambda_n \phi_n(x), & x \in \Omega, \\ a_i \phi_n(x) + b_i \nabla \phi_n(x) \cdot \hat{n}(x) = 0, & x \in \Gamma_i, \quad i = 1, ..., n, \end{cases} \tag{35}$$

  for some eigenvalues λ_n and orthonormal eigenfunctions ϕ_n.

- Any convergent expansion in $\phi_n(x)\phi_{n'}(x')$ will then satisfy the boundary conditions. Solin et. al proposed that the covariance function be given by the specific expansion

$$k(x, x') = \sum_{n=1}^{M} S\left(\sqrt{\lambda_n}\right) \phi_n(x) \phi_n(x'), \tag{36}$$

  where $S\left(\sqrt{\lambda_n}\right)$ is the spectral power density (Fourier transform) of an “original” covariance function of interest.

- Solin et. al also demonstrated a reduced-rank property provided by such kernels.

Illustration of covariance kernels satisfying boundary conditions

- For example, for the squared-exponential covariance kernel (4), the spectral power density is

$$S(\omega) = s^2(2\pi\ell^2)^{d/2}\exp\left(-\frac{1}{2}\ell^2\omega^2\right). \tag{37}$$

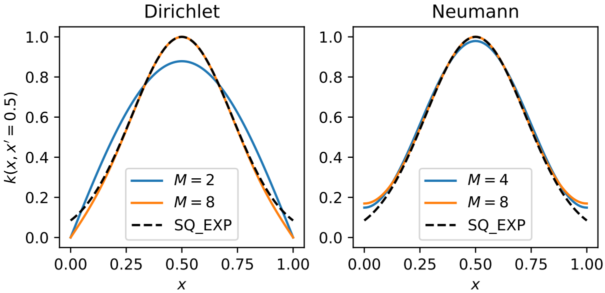


Figure: Comparison of the squared-exponential kernel $k(x, x' = 0.5)$ with the corresponding spectral expansion kernel (36) at $x' = 0.5$ for $x \in \Omega = (0, 1)$, defined using homogeneous Dirichlet (***left***) and Neumann (***right***) spectrum for different M. The squared-exponential kernel satisfies neither zero Dirichlet nor zero Neumann boundary conditions.

The reduced rank advantage to spectral expansion covariance kernels

- Using a spectral expansion covariance kernel with M terms, the covariance matrix augmented with a Gaussian likelihood (white noise) is given by

$$\tilde{K} = K + \sigma^2 I_N = \Phi \Lambda \Phi^\top + \sigma^2 I_N, \tag{38}$$

where Φ is the $N \times M$ matrix of eigenfunctions at the point locations,

$$[\Phi]_{i,j} = \phi_j(x_i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq M, \tag{39}$$

and Λ is the $M \times M$ diagonal matrix of the spectral power density evaluated at the eigenvalues λ_j corresponding to the ϕ_j,

$$\Lambda = \operatorname{diag}\left(S\left(\sqrt{[\lambda_1 \ \lambda_2 \ \ldots \ \lambda_M]}\right)\right). \tag{40}$$

- The inverse of the $N \times N$ covariance matrix (38) can be calculated as

$$\tilde{K}^{-1} = \frac{1}{\sigma^2}(I_N - \Phi Z^{-1} \Phi^\top), \tag{41}$$

where we have defined the $M \times M$ matrix $Z = \sigma^2 \Lambda^{-1} + \Phi^\top \Phi$.

- Solin and Sarkka showed that posterior prediction and likelihood estimation can expressed in terms of Z^{-1}, which no longer scales as N^3.

Combining Boundary Value and Linear PDE Constraints

- Given: observations of both the function u and f at potentially disjoint locations X_u and X_f.
- We also assume that a kernel function of the form (36) is used in which the eigenfunctions and eigenvalues are consistent with the BVP defining the constraint.
- We compute the covariance between the solution u and forcing term f as

$$\mathrm{Cov}(u(x), f(x')) = \mathrm{Cov}(u(x), Lu(x')) = \sum_{j=1}^{M} S\left(\sqrt{\lambda_j}\right)\phi_j(x) L\phi_j(x') = \sum_{j=1}^{M} S\left(\sqrt{\lambda_j}\right)\lambda_j \phi_j(x)\phi_j(x'),$$

$$\mathrm{Cov}(f(x), f(x')) = \mathrm{Cov}(Lu(x), Lu(x')) = \sum_{j=1}^{M} S\left(\sqrt{\lambda_j}\right)\lambda_j^2 \phi_j(x)\phi_j(x').$$

- The covariance matrix between the solution and forcing observations can therefore be constructed in a block-matrix form as

$$\begin{bmatrix} u(X_u) \\ f(X_f) \end{bmatrix} \sim \mathcal{GP}\left(\begin{bmatrix} m(X_u) \\ Lm(X_f) \end{bmatrix}, K_{\text{joint}}\right), \tag{42}$$

where

$$K_{\text{joint}} = \begin{bmatrix} \sum_{j=1}^{M} S(\sqrt{\lambda_j})\phi_j(X_u)\phi_j(X_u)^\top & \sum_{j=1}^{M} S(\sqrt{\lambda_j})\lambda_j\phi_j(X_u)\phi_j(X_f)^\top \\ \sum_{j=1}^{M} S(\sqrt{\lambda_j})\lambda_j\phi_j(X_f)\phi_j(X_u)^\top & \sum_{j=1}^{M} S(\sqrt{\lambda_j})\lambda_j^2\phi_j(X_f)\phi_j(X_f)^\top \end{bmatrix}. \tag{43}$$

Combining Boundary Value and Linear PDE Constraints

- Defining the $N_u \times M$ matrix Φ_u and the $N_f \times M$ matrix Φ_f as

$$[\Phi_u]_{i,j} = \phi_j(x_i), \quad 1 \le i \le N_u, \quad x_i \in X_u, \quad 1 \le j \le M, \tag{44}$$

$$[\Phi_f]_{i,j} = \lambda_i \phi_j(x_i), \quad 1 \le i \le N_f, \quad x_i \in X_f, \quad 1 \le j \le M, \tag{45}$$

  and the block matrix

$$\Phi_{\text{joint}} = \begin{bmatrix} \Phi_u \\ \Phi_f \end{bmatrix}, \tag{46}$$

  the covariance matrix (43) augmented by the Gaussian likelihood can be written as

$$\tilde{K}_{\text{joint}} = K_{\text{joint}} + \sigma^2 I_{N_u+N_f} = \Phi_{\text{joint}} \Lambda \Phi_{\text{joint}}^\top + \sigma^2 I_{N_u+N_f}. \tag{47}$$

- The form of this kernel mimics that of (38). Defining Z with Φ_{joint} in place of Φ allows the entire reduced-rank framework to be utilized, with the matrix Φ_{joint} in place of Φ throughout.
- Allows for reduced-rank GPR with noisy data enhanced by PDE and BC prior knowledge.
- Also allows for a new application: inference of solution u to a BVP with only IC and BC conditions, and scattered observations of f rather than u.

Comparison of unconstrained and constrained GPR for $-u'' = f, \quad u(0) = u(1) = 0$

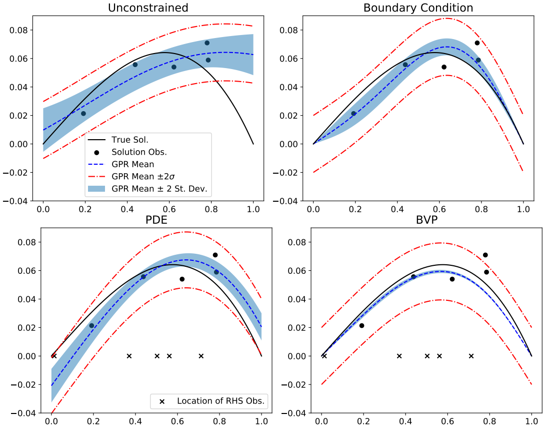


Figure: ***Top Left:*** Unconstrained GPR using a standard Sq.Exp. kernel; rel. ℓ^2 error of 42.5%. ***Top Right:*** BC-constrained GPR using the spectral expansion kernel; rel. ℓ^2 error of 14.6%. ***Bottom Left:*** PDE-constrained GPR using a squared-exponential kernel; rel. ℓ^2 error of 25.9%. ***Bottom Right:*** BVP-constrained GPR; rel. ℓ^2 error of 9.3%.

5 observations (black dots) of the function u at randomly sampled points in $[0, 1]$, obtained by sampling u and adding white noise with $\sigma = 0.01$. PDE and BVP constrained problems use 5 observations of f sampled at the black "x" marks. The relative errors are between the posterior mean of the GPR (dashed blue curve) and the exact solution u (solid black curve).

Inferring the solution to $-u'' = f, \quad u(0) = u(1) = 0$ with BVP data only (no interior observations)

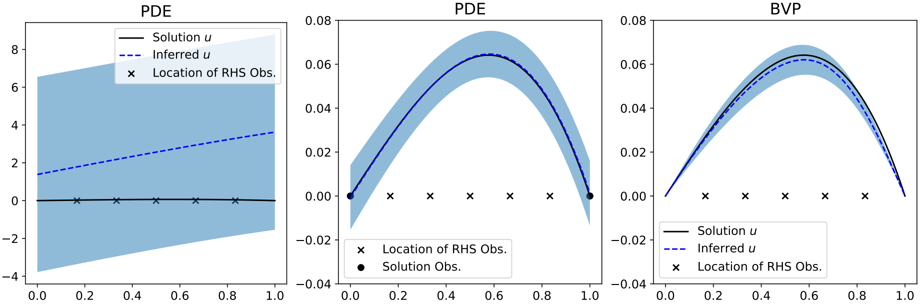


Figure: Effect of enforcing the boundary conditions when inferring u from 5 observations of f.

- When using the PDE-GP method ***(left)***, inference fails without observations of u, as even with complete knowledge of f, u is only determined up to an arbitrary linear function.
- When BCs are treated in the PDE-GP method as point observations of u ***(center)***, accurate inference is possible although uncertainty is nonzero in contrast to the BVP-GP method.
- In the BVP-GP method ***(right)***, the boundary conditions are enforced with certainty via the covariance kernel, not as discrete observations, which is advantageous in higher dimensions.

Error w.r.t. number of observations and noise in observations

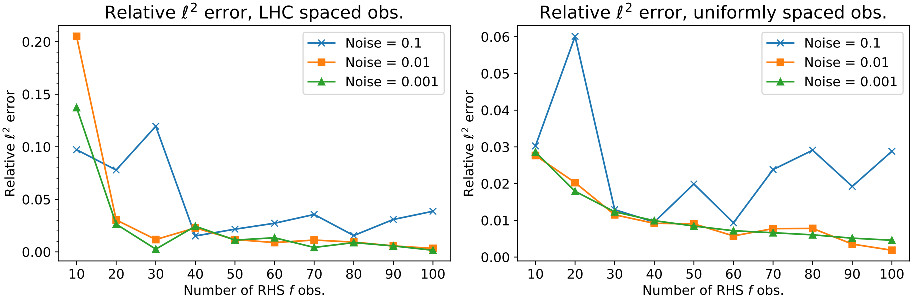
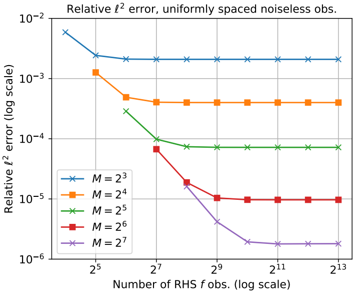
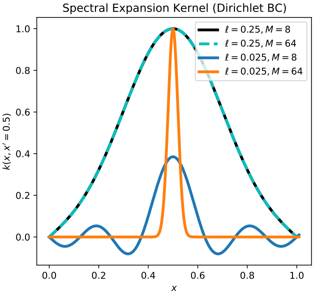


Figure: Plot of the error between the posterior mean prediction u^* and the true solution u, measured in the relative ℓ^2 norm over 100 uniformly spaced test points in $[0, 1]$. For the relatively large value of white noise standard deviation $\sigma = 0.1$ (applied to observations of f), the trend is less consistent, but for $\sigma = 0.01$ and $\sigma = 0.001$ the error trends more consistently and saturates around 1% for both observations at LHC sampled locations and on the uniform grid.

Error w.r.t. number of observations and kernel expansion order





- ***Left:*** Convergence in log-log scale of the error between the posterior mean prediction u^* and the true solution u, trained with noiseless observations, measured in the relative ℓ^2 norm over 100 uniformly spaced test points in $[0, 1]$.
- The noise/likelihood hyperparameter σ is fixed to 10^{-17}. For fixed number M of eigenfunctions defining the covariance kernel, the error decreases with the number n_f of observations. As M increases, the error decreases.
- ***Right:*** Plotting the spectral expansion covariance kernel $k(x, x' = 0.5)$ for various M reveals that artifacts are present when the correlation length hyperparameter ℓ (width of the parent squared exponential kernel) is small, and increasing M reduces these artifacts.

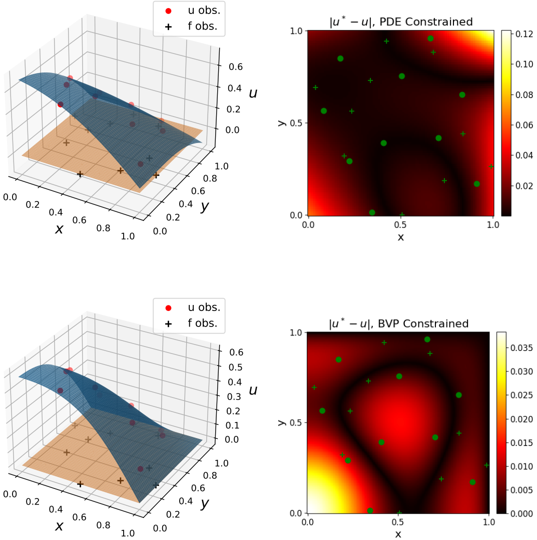


Figure: Comparison of PDE constrained GPR (***top***) and BVP constrained GPR (***bottom***). The left column shows observations of u (red dots) and locations of the observations of the source f (black crosses) and the resulting mean prediction surface u^* (blue). The xy-plane is plotted in orange as a reference for observing the boundary behavior of u^*. The right column plots the absolute error between the mean prediction u^* and the true solution u. The BVP constrained GPR demonstrates a lower relative ℓ^2 error over the uniform 100×100 test grid: 2.88% vs 5.25%.

Conclusion & Acknowledgements

- We have developed a framework that combines the use of spectral decomposition covariance kernels with differential equation constraints in a co-kriging setup to perform Gaussian process regression constrained by boundary value problems.
- Novel application of Gaussian process regression to BVPs with Neumann boundary conditions and to inference of the solution u of BVP from knowledge of the boundary condition and scattered observations of the source term alone.
- The lower-dimensional representation inherent to the spectral covariance kernel yielded an efficient training and inference process.
- The BVP-GP method can be seamlessly used in a spectrum of applications from small datasets with high noise to large, noiseless datasets. In more complex domains, numerically computed eigenfunctions may be substituted.
- This work was supported by the LDRD program at Sandia National Laboratories, and its support is gratefully acknowledged. M. Gulian was also supported by the John von Neumann fellowship at Sandia National Laboratories, and by the U.S. Department of Energy, Office of Advanced Scientific Computing Research under the Collaboratory on Mathematics and Physics-Informed Learning Machines for Multiscale and Multiphysics Problems (PhILMs) project.