

Gaussian Process Regression constrained by Boundary Value Problems

Mamikon Gulian¹, Ari Frankel², Laura Swiler³

- (1) Quantitative Modeling and Software Engineering, Sandia National Laboratories, Livermore, CA
- (2) Applied Scientist, Uber, San Francisco, CA
- (3) Center for Computing Research, Sandia National Laboratories, Albuquerque, NM



**Sandia
National
Laboratories**

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: SAND2023-03971C.



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

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

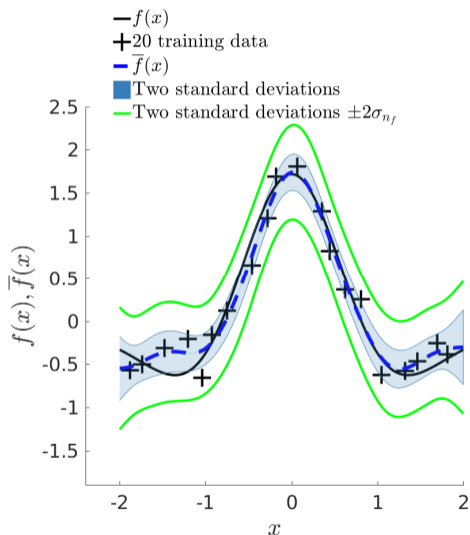
where L denotes a linear partial differential operator, Ω a domain with boundary $\partial\Omega$, and \mathcal{B} a general mixed boundary operator.

- ▶ 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 and/or solution 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.

Background

- ▶ The work of Raissi et al. [RPK17] 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.
- ▶ This is related to the approach of Graepel [Gra03] which considered the case of observations of f only.
- ▶ Solin and Kok [SK19] 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.

Illustration of GPR posterior



- ▶ A prior is specified, encoding some assumptions about smoothness, stationarity (or lack thereof) of function or field via a covariance kernel $K(x, x')$.
- ▶ A noise model – the likelihood function – relating observations y to dependent variable $f(x)$ is specified.
- ▶ Here, white noise from $\mathcal{N}(0, \sigma_{n_f})$ was added to some points on the black curve to generate observations (black crosses).
- ▶ Observations are given, and kernel and likelihood (noise) hyperparameters are tuned by maximizing the log marginal likelihood \mathcal{L} .
- ▶ The mean $\mathbb{E}[f(x^*)]$ of the posterior gives the prediction of GPR, with the posterior variance giving an estimate of uncertainty in the prediction.

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 $K(x, x') = \text{Cov}(u(x), u(x'))$:

$$u \sim \mathcal{GP}(m, K). \quad (2)$$

- ▶ That is, the vector of values $u(X)$ over a finite collection of locations X has a multivariate normal density

$$u(X) \sim \mathcal{N}(m(X), K(X, X)), \quad (3)$$

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

$$K(x, x') = s^2 \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) \quad (4)$$

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

$$p(y|u, X) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - u_i(X_i))^2}{2\sigma^2}\right). \quad (5)$$

Basics of GPR: posterior prediction

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

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

with log-marginal-likelihood

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

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

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

and variance

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

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

PDE-constrained GPR

- ▶ If $u \sim \mathcal{GP}(m(x), k(x, x'))$ and $Lu = f$ for a linear operator L , and if $m(\cdot), k(\cdot, x') \in \text{dom}(L)$ 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 :

$$Lu \sim \mathcal{GP}(L_x m(x), L_x L_{x'} k(x, x')). \quad (10)$$

- ▶ 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

$$k \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} \right) = \begin{bmatrix} k(x_1, x'_1) & L_{x'} k(x_1, x'_2) \\ L_x k(x_2, x'_1) & L_x L_{x'} k(x_2, x'_2) \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}. \quad (11)$$

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

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

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

$$\mathbb{E}[u(x)] = \sum_{i=1}^N c_i k(x, x_i), \quad (13)$$

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

$$\begin{cases} L\phi_n(x) = \lambda_n\phi_n(x), & x \in \Omega, \\ a_i\phi_n(x) + b_i\nabla\phi_n(x) \cdot \hat{n}(x) = 0, & x \in \Gamma_i, \quad i = 1, \dots, n, \end{cases} \quad (14)$$

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

$$k(x, x') = \sum_{n=1}^M S(\sqrt{\lambda_n}) \phi_n(x)\phi_n(x'), \quad (15)$$

where $S(\sqrt{\lambda_n})$ 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

$$S(\omega) = s^2 (2\pi\ell^2)^{d/2} \exp\left(-\frac{1}{2}\ell^2\omega^2\right). \quad (16)$$

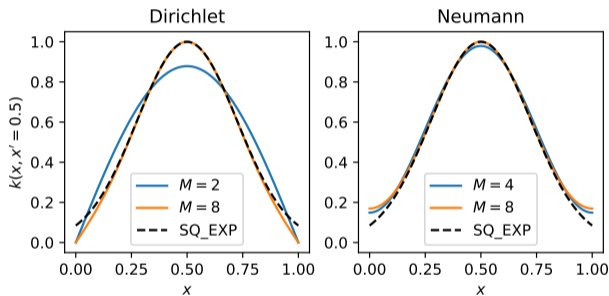


Figure: Comparison of the squared-exponential kernel $k(x, x' = 0.5)$ with the corresponding spectral expansion kernel (15) at $x' = 0.5$ for $x \in \Omega = (0, 1)$, defined using homogeneous Dirichlet (*left*) and Neumann (*right*) spectrum for different M . The squared-exponential kernel satisfies 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

$$\tilde{K} = K + \sigma^2 I_N = \Phi \Lambda \Phi^\top + \sigma^2 I_N, \quad (17)$$

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

$$[\Phi]_{i,j} = \phi_j(x_i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq M, \quad (18)$$

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 ,

$$\Lambda = \text{diag} \left(S \left(\sqrt{[\lambda_1 \ \lambda_2 \ \dots \ \lambda_M]} \right) \right). \quad (19)$$

- ▶ The inverse of the $N \times N$ covariance matrix (17) can be calculated as

$$\tilde{K}^{-1} = \frac{1}{\sigma^2} (I_N - \Phi Z^{-1} \Phi^\top), \quad (20)$$

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

- ▶ Solin and Sarkka [SS19] showed that posterior prediction and likelihood estimation can be expressed in terms of Z^{-1} .
- ▶ The computational complexity now scales as $\mathcal{O}(M^3 + NM^2)$ rather than $\mathcal{O}(N^3)$.

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 (15) 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

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

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

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

$$\begin{bmatrix} u(X_u) \\ f(X_f) \end{bmatrix} \sim \mathcal{GP} \left(\begin{bmatrix} m(X_u) \\ Lm(X_f) \end{bmatrix}, K_{\text{joint}} \right), \quad (21)$$

where

$$K_{\text{joint}} = \begin{bmatrix} \sum_{j=1}^M S(\sqrt{\lambda_j}) \phi_j(X_u) \phi_j(X_u)^\top & \sum_{j=1}^M S(\sqrt{\lambda_j}) \lambda_j \phi_j(X_u) \phi_j(X_f)^\top \\ \sum_{j=1}^M S(\sqrt{\lambda_j}) \lambda_j \phi_j(X_f) \phi_j(X_u)^\top & \sum_{j=1}^M S(\sqrt{\lambda_j}) \lambda_j^2 \phi_j(X_f) \phi_j(X_f)^\top \end{bmatrix}. \quad (22)$$

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

$$[\Phi_u]_{i,j} = \phi_j(x_i), \quad 1 \leq i \leq N_u, \quad x_i \in X_u, \quad 1 \leq j \leq M, \quad (23)$$

$$[\Phi_f]_{i,j} = \lambda_i \phi_j(x_i), \quad 1 \leq i \leq N_f, \quad x_i \in X_f, \quad 1 \leq j \leq M, \quad (24)$$

and the block matrix

$$\Phi_{\text{joint}} = \begin{bmatrix} \Phi_u \\ \Phi_f \end{bmatrix}, \quad (25)$$

the covariance matrix (22) augmented by the Gaussian likelihood can be written as

$$\tilde{K}_{\text{joint}} = K_{\text{joint}} + \sigma^2 I_{N_u+N_f} = \Phi_{\text{joint}} \Lambda \Phi_{\text{joint}}^\top + \sigma^2 I_{N_u+N_f}. \quad (26)$$

- ▶ The form of this kernel mimics that of (17). 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 $-u'' = f$, $u(0) = u(1) = 0$

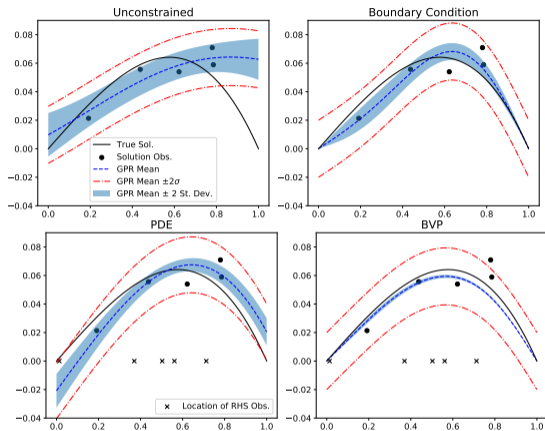


Figure: Top Left: Unconstrained GPR using a standard Sq.Exp. kernel; rel. ℓ^2 error of 42.5%.

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 $-u'' = f$, $u(0) = u(1) = 0$ with BVP data only (no interior observations)

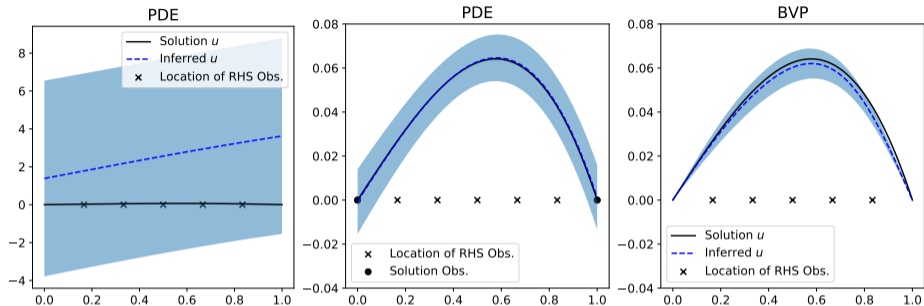


Figure: Effect of enforcing the boundary conditions when inferring u from 5 observations of f .

- ▶ 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

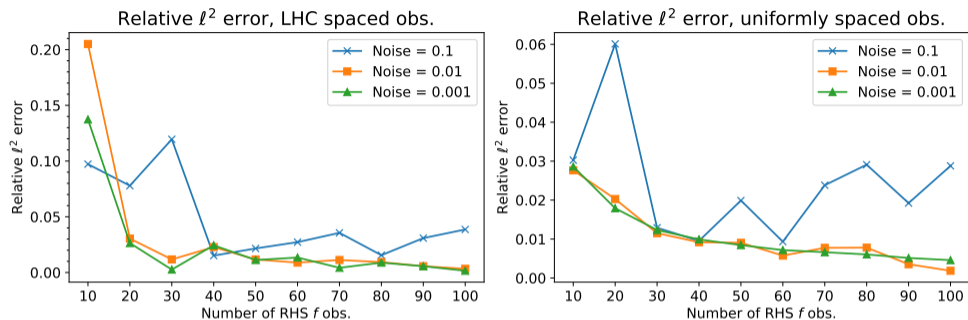
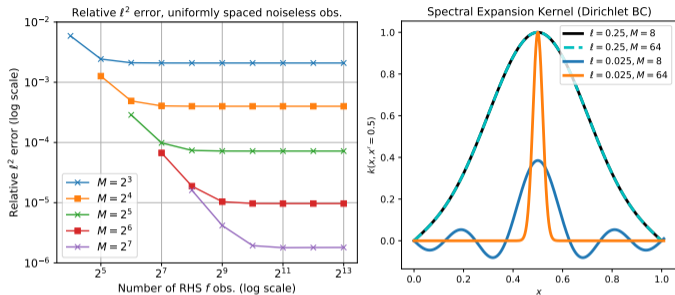


Figure: Plot of the error between the posterior mean prediction u^* and the true solution u , measured in the relative ℓ^2 norm over 100 uniformly spaced test points in $[0, 1]$. For the relatively large value of white noise standard deviation $\sigma = 0.1$ (applied to observations of f), the trend 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.

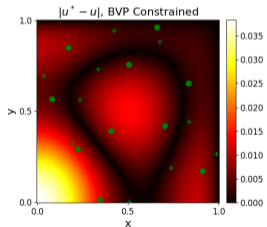
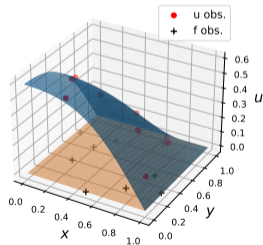
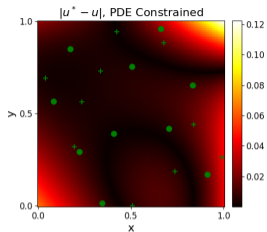
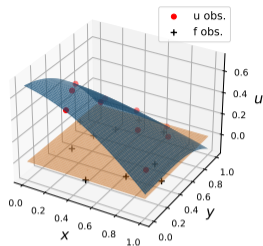








Figure: Comparison of PDE constrained GPR (*top*) and BVP constrained GPR (*bottom*). The left column shows observations of u (red dots) and locations of the observations of the source f (black crosses) and the resulting mean prediction surface u^* (blue). The xy -plane is plotted in orange as a reference 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 [SGF⁺, GFS22] was supported by the LDRD program at SNL, the John von Neumann postdoctoral fellowship at SNL, 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.

- 
- Mamikon Gulian, Ari Frankel, and Laura Swiler, *Gaussian process regression constrained by boundary value problems*, Computer Methods in Applied Mechanics and Engineering **388** (2022), 114117.
- 
- Thore Graepel, *Solving noisy linear operator equations by Gaussian processes: Application to ordinary and partial differential equations*, ICML, 2003, pp. 234–241.
- 
- Maziar Raissi, Paris Perdikaris, and George Em Karniadakis, *Machine learning of linear differential equations using Gaussian processes*, Journal of Computational Physics **348** (2017), 683–693.
- 
- Laura P Swiler, Mamikon Gulian, Ari L Frankel, Cosmin Safta, and John D Jakeman, *A survey of constrained Gaussian process regression: Approaches and implementation challenges*, Journal of Machine Learning for Modeling and Computing, in submission.
- 
- Arno Solin and Manon Kok, *Know your boundaries: Constraining Gaussian processes by variational harmonic features*, Proceedings of Machine Learning Research, vol. 89, PMLR, 16–18 Apr 2019, pp. 2193–2202.
- 
- Arno Solin and Simo Särkkä, *Hilbert space methods for reduced-rank Gaussian process regression*, Statistics and Computing (2019) (English).