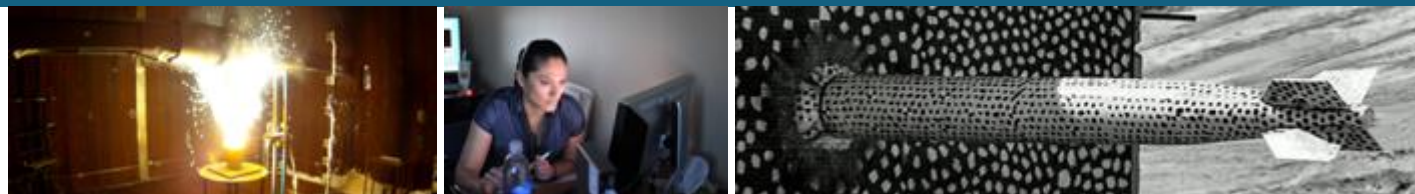


Constrained Gaussian Processes: A Survey



Laura Swiler, Mamikon Gulian, Ari Frankel, Cosmin Safta,
and John Jakeman

*SIAM Computational Science and Engineering
Virtual Conference March 1-5, 2021*



- Tremendous surge in the development and application of machine learning (ML) models in recent years due to their flexibility and capability to represent trends in complex systems.
- In many scientific applications, a large amount of data may not be available for training. Computational and physical experiments are often expensive.
- Even if ample data exists, the ML model may yield behaviors that are inconsistent with what is expected physically when queried in an extrapolatory regime.
- To **improve the process of building ML 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.
- Within the Bayesian regression framework, Gaussian processes (GPs) are popular for constructing *surrogates* or *emulators* of data sources that are very expensive to query.
- An accurate Gaussian process regression (GPR) can often be used constructed using only a relatively small number of training data which consists of pairs of input parameters and corresponding response values.



- We investigated several methods for incorporating constraints within Gaussian processes.
- Constraints surveyed included:
 - positivity or bound constraints,
 - monotonicity and convexity constraints
 - linear differential equation constraints
 - boundary value constraints.
- Constraints can be enforced in a variety of ways:
 - implicitly through data that satisfies the constraint
 - by construction of a tailored sample space
 - by derivation of a constrained covariance kernel
 - by modifying the output of the GP
 - by modifying the likelihood of the GP
- Constraints may be enforced in a global sense or at a finite set of “virtual” or “auxiliary” points or only in an approximate sense. We have pointed to these aspects as key features distinguishing the constraints in the survey.



- This work was completed with funding granted under Sandia's Laboratory Directed Research and Development program.
- Thank you to the organizers.
- See our article for complete discussion for the methods discussed in this presentation, presented in roughly the same order as this presentation.

L.P. Swiler, M. Gulian, A. Frankel, C. Safta, and J.D. Jakeman. A Survey of Constrained Gaussian Process Approaches and Implementation Challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2):119–156 (2020). DOI: [10.1615/JMachLearnModelComput.2020035155](https://doi.org/10.1615/JMachLearnModelComput.2020035155)

- Disclaimer 1: most of the strategies and claims presented here are not original. The survey article provides extensive references for each method discussed in this presentation.
- Disclaimer 2: Some works and types of constraints have been left out to make the survey feasible. Physical constraints are highly varied and may not fit into a taxonomy.
- Thank you for your attention!



- A Gaussian process can be viewed as a distribution over a set of functions. A random draw or sample f from a GP is a realization from the set of admissible functions.
- Specifically, a Gaussian process is a collection of random variables $\{f(x) | x \in X\}$ for which, given any finite set of N inputs $X = \{x_1, x_2, \dots, x_N\}$, the collection $f(x_1), f(x_2), \dots, f(x_N)$ from the random process has a joint multivariate Gaussian distribution.
- A GP is completely defined by its mean and covariance functions which generate the mean vectors and covariances matrices of these finite-dimensional multivariate normals.
- Assumptions such as smoothness of samples f , stationarity, and sparsity are used to construct the mean and covariance of the GP prior and then Bayes' rule is used to constrain the prior with observational/simulation data.

The prediction $f = [f(x_1), f(x_2), \dots, f(x_N)]^T$ of a Gaussian process with mean function $m(x)$ and a covariance function $k(x, x')$ is a random variable such that

$$p(f|X) = \mathcal{N}(f; m(X); k(X, X)) \quad \leftarrow \text{GP Prior}$$



- If the error or noise relating the actual observations $\mathbf{y} = [y(x_1), y(x_2), \dots, y(x_N)]^T$ collected at the set of inputs $X = \{x_i\}_{i=1}^N$ to the GP prediction f is assumed to be Gaussian, then the probability of observing data \mathbf{y} given the GP prior is given by

$$p(\mathbf{y}|X, f) = \mathcal{N}(f, \sigma^2 \mathbf{I}_N) \quad \leftarrow \text{GP Likelihood}$$

- If we assume a squared exponential covariance kernel $k(\mathbf{x}, \mathbf{x}') = \eta^2 \exp \left\{ -\frac{1}{2} \sum_{j=1}^d \left(\frac{x_j - x_{j'}}{\rho_j} \right)^2 \right\}$

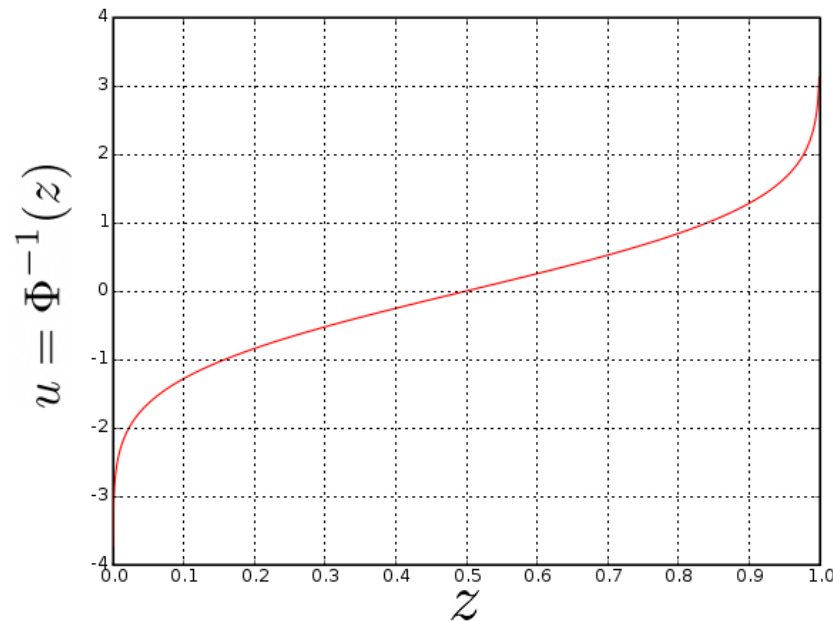
then we obtain the hyperparameters $\theta = [\eta, \rho_1, \rho_2, \dots, \rho_d, \sigma]$ by maximizing the log of the marginal likelihood function, $p(\mathbf{y}|X, \theta) = \int p(\mathbf{y}|X, f, \theta) p(f|X, \theta) df$

- Given the prior $p(f|X, \theta)$ and the Gaussian likelihood $p(\mathbf{y}|X, f, \theta)$, the prediction f of a GPR at a new point \mathbf{x}^* can be calculated as

$$p(f^* | \mathbf{y}, X, \mathbf{x}^*, \theta) = \mathcal{N}(k(\mathbf{x}^*, X) [K(X, X) + \sigma^2 \mathbf{I}_N]^{-1} \mathbf{y},$$

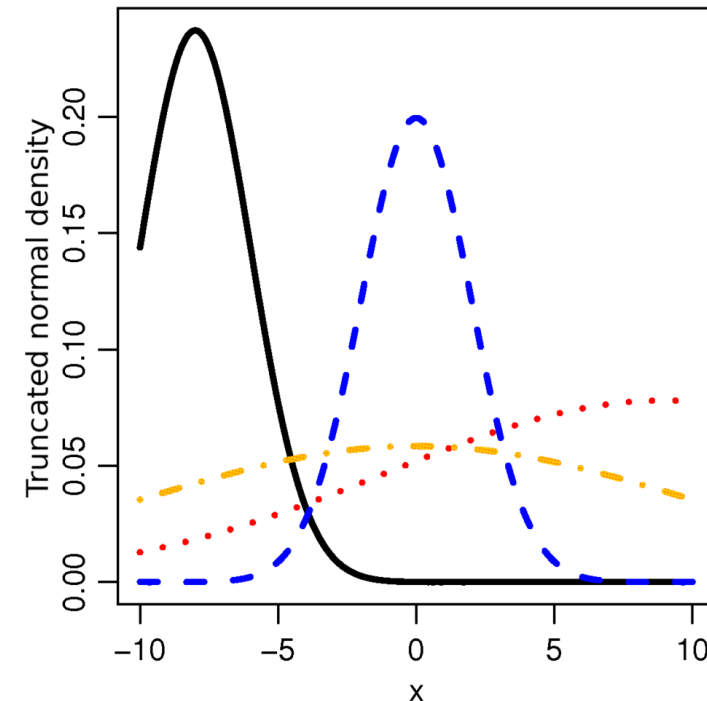
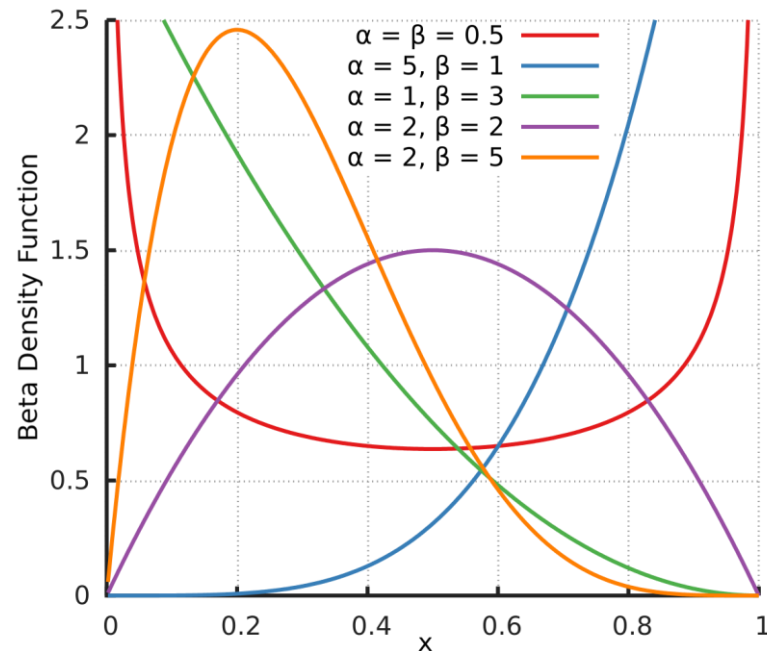
$$k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X) [K(X, X) + \sigma^2 \mathbf{I}_N]^{-1} k(\mathbf{x}^*, X)^T) \quad \leftarrow \text{GP Posterior for zero mean prior case}$$

- Bound constraints of the form $a \leq f \leq b$ over some region of interest arise naturally in many applications, such as chemical concentration data.
- Warping functions are used to transform bounded observations z_i to unbounded observations u_i which can be treated with unconstrained GPR, then transformed back.
- 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 $u_i = \Phi^{-1}(z_i)$.





- In addition to using warping functions, bound constraints can also be enforced using non-Gaussian likelihood functions $p(y|X, f, \theta)$ that are constructed to produce GP observations which satisfy the constraints.
- There are several 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
- Unlike the warping method, the posterior is not analytically tractable; Laplace approximation and expectation propagation can be used for approximate inference with the posterior.



9 Bound Constraints: Truncated Multivariate Normal (MVN)



- Since a Gaussian process is always trained and evaluated at a finite set of points X , a “global” constraint of the form $a \leq f(x) \leq b$ can be approximated by constraints at a finite set of N_c auxiliary or “virtual” points x_1, \dots, x_{N_c} .
- This requires constructing an unconstrained GP and then, over the virtual points, transforming this GP to a truncated multivariate Gaussian distribution

$$\mathcal{TN}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{a}, \mathbf{b}) = \begin{cases} \frac{\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathbb{P}(\mathbf{a} \leq \mathbf{z} \leq \mathbf{b})}, & \text{for } \mathbf{a} \leq \mathbf{z} \leq \mathbf{b} \\ 0, & \text{otherwise} \end{cases}$$

- The constrained mean predictor is conditioned on the original data (X, y) but also augmented by a fixed, finite set of discrete points: $\{x_i\}_{i=1}^{N_c}$

$$\mathbb{E}[f(x^*) | f(X) = y \text{ and } a \leq f(x_i) \leq b \text{ for all } i = 1, 2, \dots, N_c]$$

Bound Constraints: Constrained Max Likelihood to enforce non-negativity constraints



- Constrain the optimization of the log marginal likelihood so that hyperparameters are chosen to enforce bounds. Again, we assume some constraints at a finite set of N_c points $\mathbf{x}_1, \dots, \mathbf{x}_{N_c}$.

$$p\{(f_i^* | \mathbf{y}, X, \mathbf{x}_i^*, \theta) < 0\} \leq \epsilon \quad i = 1, 2, \dots, N_c$$

- For a Gaussian likelihood, the unconstrained posterior f^* follows a Gaussian distribution. We can then write a probabilistic constraint in terms of the unconstrained posterior mean and variance (e.g. enforce the mean minus 2 standard deviations to be positive):

$$\begin{aligned} \text{Seek} \quad & \theta^* = \underset{\theta}{\operatorname{argmax}} \log[p(\mathbf{y}|X, \theta)] \\ \text{subject to} \quad & 0 \leq \hat{m}(\mathbf{x}_i) - 2s(\mathbf{x}_i), \quad i = 1, \dots, N_c \\ \text{and} \quad & 0 \leq v - |y_j - f(\mathbf{x}_j)|, \quad j = 1, \dots, N. \end{aligned}$$



- For any set of spline basis coefficients ξ_i , the function representation is given as a summation over spline knot locations $x_i = \frac{i}{M}$ for $i = 0 \dots M$

$$f(x) = \sum_{i=1}^M \xi_i h(x_i) \quad \text{where } h(x) = \max(1 - |x|, 0)$$
- We have a finite dimensional constraint: $a \leq f(x_i) \leq b$ if $a \leq \xi_i \leq b$
- Suppose we are given a set of N data points at unique locations (x_j, y_j) . Define the matrix A such that $A_{i,j} = h_i(x_j)$. Then any set of spline coefficients that satisfy the equation $A\xi = \mathbf{y}$ 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 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 ξ and \mathbf{y} are governed jointly by a GP prior in the form:

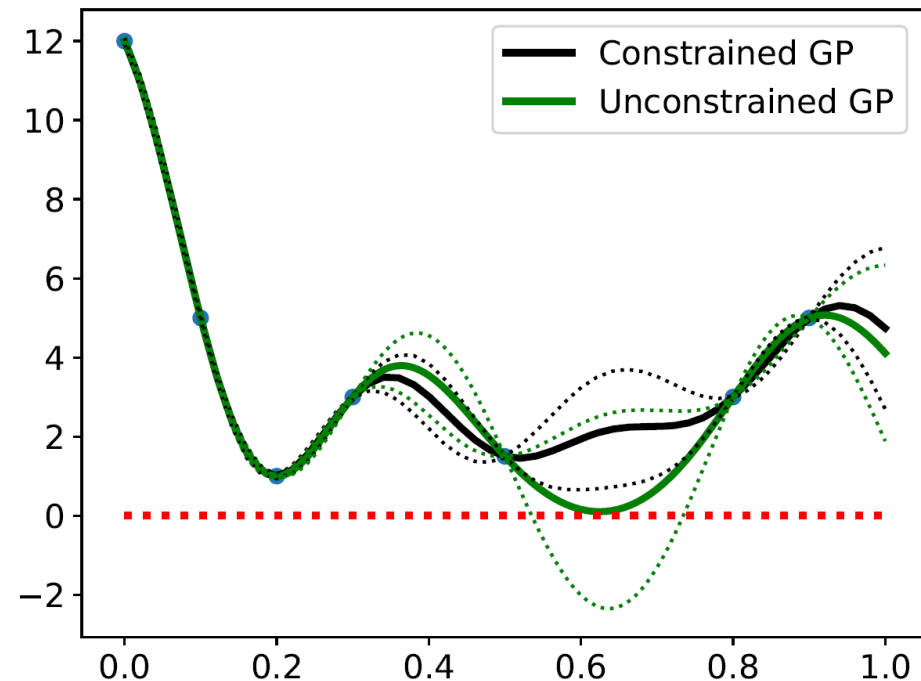
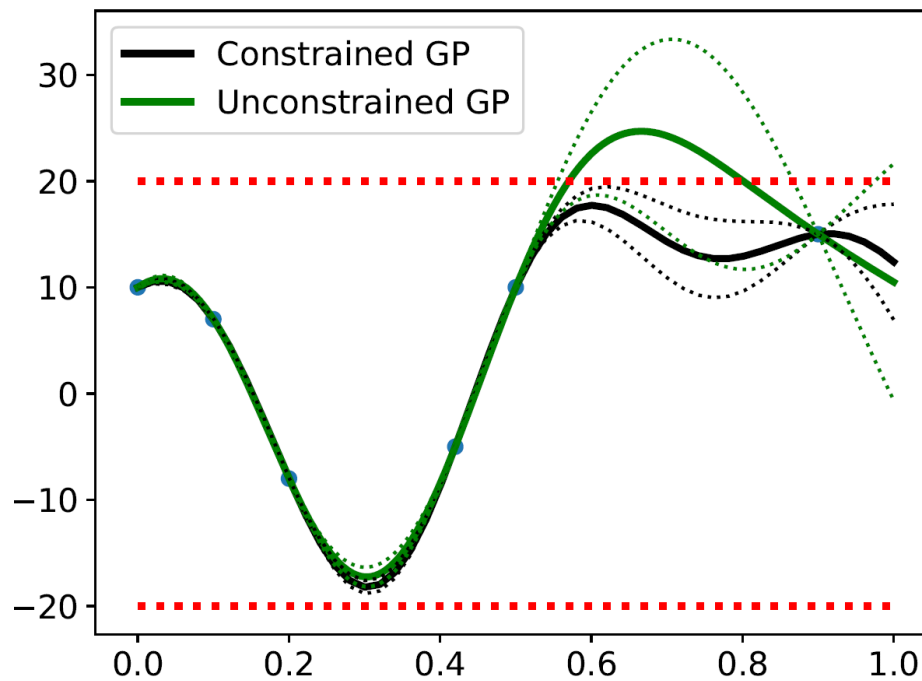
$$\begin{pmatrix} \mathbf{y} \\ \xi \end{pmatrix} = \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \mid \begin{bmatrix} AK A^T & K A^T \\ AK & K \end{bmatrix} \right)$$

- Upon observation of the data y , the conditional distribution of the knot values subject to $y = A\xi$ is given by

$$p(\xi|y = A\xi) = \mathcal{N}(\xi; KA^T(AKA^T)^{-1}y, K - KA^T(AKA^T)^{-1}AK)$$

- We then evaluate the distribution further conditioned on the inequality constraints $\xi \in \mathcal{C}$

$$p(\xi|y = A\xi, \xi \in \mathcal{C}) = \mathcal{TN}(\xi; KA^T(AKA^T)^{-1}y, K - KA^T(AKA^T)^{-1}AK, \mathcal{C})$$





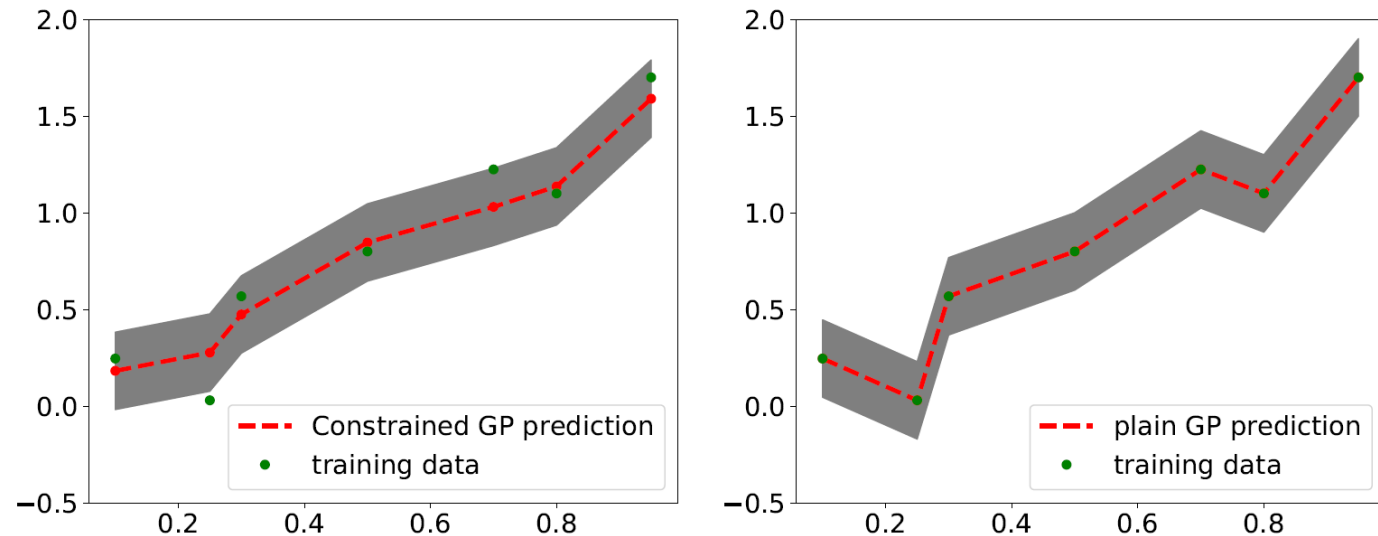
- To enforce $\frac{\partial f}{\partial x_{d_i}}(x_i) \geq 0$ at a set of finite “operating” or virtual points $X_M = \{x_i\}_{i=1}^M$, we use the notation:
- $f'_i = \frac{\partial f}{\partial x_{d_i}}(x_i)$, $f' = [\frac{\partial f}{\partial x_{d_1}}(x_1), \dots, \frac{\partial f}{\partial x_{d_M}}(x_M)]^T = [f'_1, \dots, f'_M]^T$
- An observation of f'_i is denoted by y'_i .
- We use a likelihood function that forces the likelihood to be zero or one in most cases: $p(y'_i | f'_i) = \Phi^{-1}(f'_i \frac{1}{\nu})$
where ν is a small number (10^{-4} to 10^{-6}).
- We now have a joint “four block” prior over the function and derivative observations:

$$p(f, f' | X, X_M) = \mathcal{N}(f_{\text{joint}} | 0, K_{\text{joint}})$$

$$f_{\text{joint}} = \begin{bmatrix} f \\ f' \end{bmatrix}, \quad K_{\text{joint}} = \begin{bmatrix} K_{f,f} & K_{f,f'} \\ K_{f',f} & K_{f',f'} \end{bmatrix}$$



- The covariance terms between the values of the specified partial derivatives, $K_{f',f'}$, or between the function and the derivatives, $K_{f,f'}$, can be shown analytically.
- One can then obtain a posterior $p(f, f' | y, y') = \frac{1}{Z} p(f, f' | X, X_M) p(y|f)p(y'|f')$
- This distribution is analytically intractable because of the non-Gaussian likelihood for the derivative components. MCMC, Laplace approximation, and expectation propagation can be applied.



- Since monotonicity constraints are positivity (bound) constraints on the derivative part of a joint GP, the “co-kriging” setup can be combined with methods for bound constraints to implement monotonicity constraints. This is described for the truncated MVN and spline approach. Convexity can also be handled.



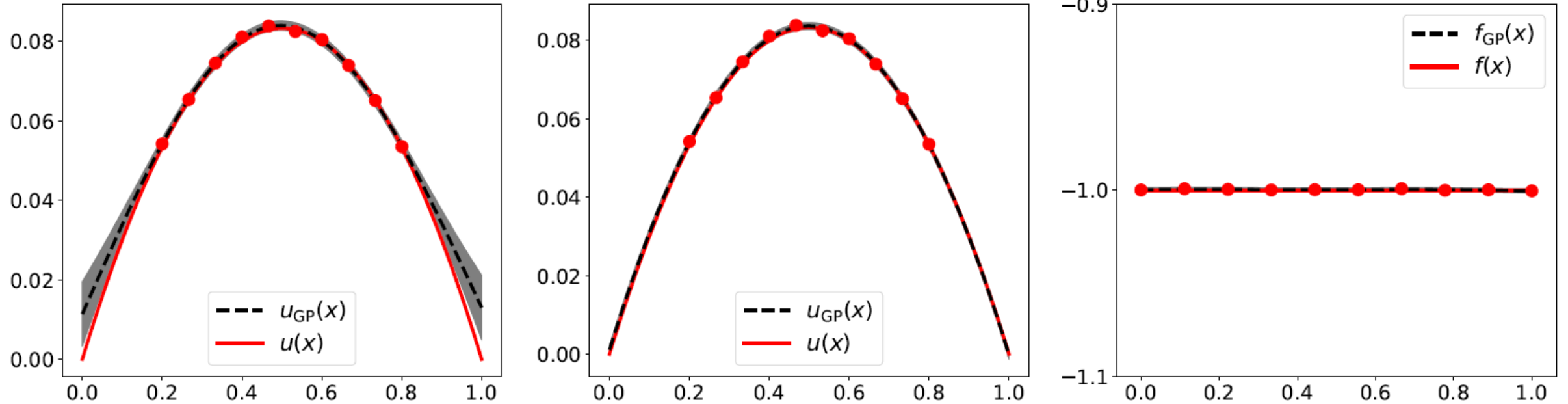
- Gaussian processes may be constrained to satisfy linear operator constraints of the form $\mathcal{L}u = f$ given data on u and f .
- When \mathcal{L} is a linear partial differential equation operator of the form $\mathcal{L} = \sum_{\alpha} C_{\alpha}(x) \frac{\partial^{\alpha}}{\partial x^{\alpha}}$, $\alpha = (\alpha_1 \dots \alpha_d)$, the equation $\mathcal{L}u = f$ can be used to constrain GP predictions to satisfy known physical laws expressed as linear PDEs.
- If $u(x)$ is a GP with mean $m(x)$ and covariance kernel $k(x, x')$, and if $m(x)$ and $k(x, x')$ belong to the domain of \mathcal{L} , then $\mathcal{L}_x \mathcal{L}_{x'} k(x, x')$ defines a valid covariance kernel for a GP with mean function $\mathcal{L}_x m(x)$.
- This Gaussian process is denoted $\mathcal{L}u$: $\mathcal{L}u \sim GP(\mathcal{L}_x m(x), \mathcal{L}_x \mathcal{L}_{x'} k(x, x'))$
- If observations are available for the source term, y_f , along with observations y_u at domain points X_u , a GP co-kriging procedure can be used to form the joint GP $[u; f]$.
- Given the covariance kernel for $u(x)$ is $k(x, x')$, the covariance for the joint GP is:

$$K \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} \right) = \begin{bmatrix} k(x_1, x'_1) & \mathcal{L}_{x'} k(x_1, x'_2) \\ \mathcal{L}_x k(x_2, x'_1) & \mathcal{L}_x \mathcal{L}_{x'} k(x_2, x'_2) \end{bmatrix} = \begin{bmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{bmatrix}$$

Linear PDE Constraint Example



- The PDE is $-1 = \frac{\partial^2 u}{\partial x^2}$ on the interval $[0,1]$. Data is generated from sampling the solution $u = \frac{1}{8} [(2x - 1)^2 - 1]$

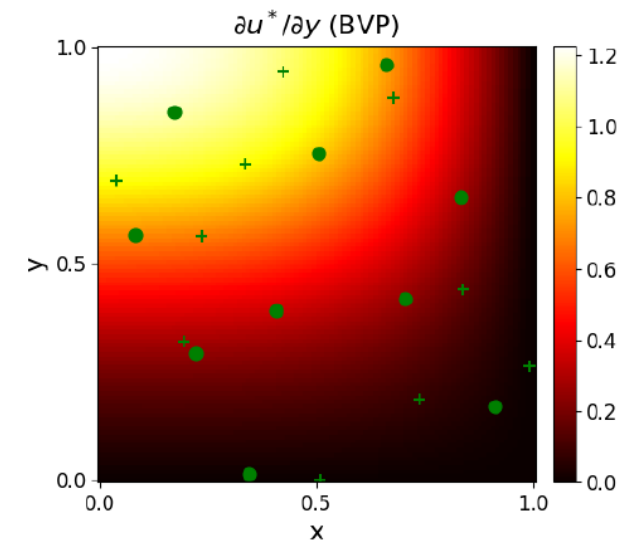
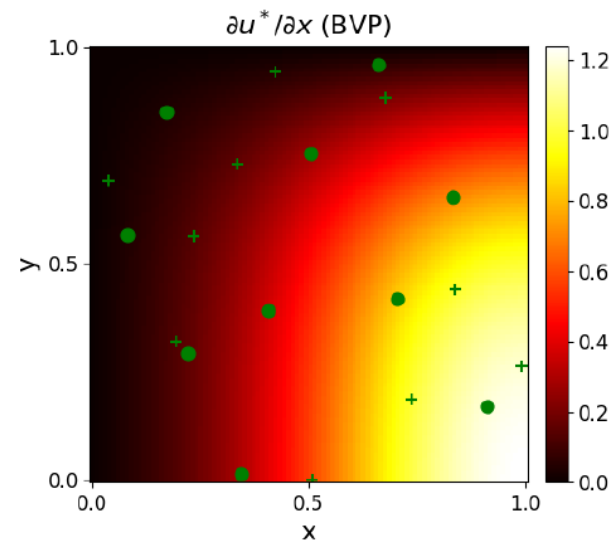
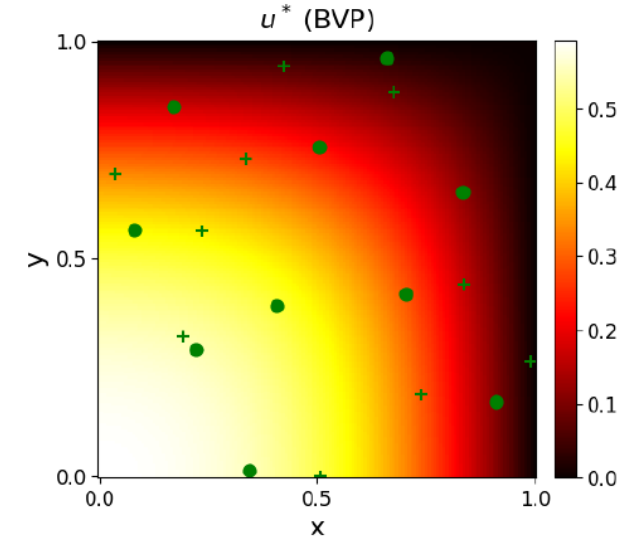
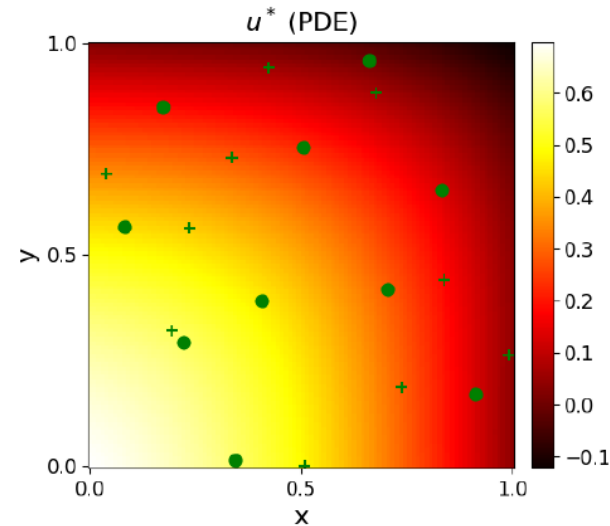


- Left: Reconstruction of u (red line) with an unconstrained GP (black line) using 10 data points (red dots) in $[0.2, 0.8]$.
- Center: Reconstruction of u (red line) with a PDE constrained GP (black line) using the same 10 data points (red dots) in $[0.2, 0.8]$.
- Right: Right-hand side f of the PDE, 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.



- Boundary value constraints can be enforced using GPR over spectral expansions in Dirichlet or Neumann eigenfunctions of the Laplacian.
- The following paper combines the concept of linear PDE constraints using co-kriging and the boundary constraints via covariance kernels approximated by spectral expansions in eigenfunctions of the boundary value problem: arxiv.org/abs/2012.11857. *Gaussian Process Regression constrained by Boundary Value Problems*. M. Gulian, A. Frankel, and L. Swiler.

$$\begin{aligned} \mathcal{L}u(x) &= f(x) \quad x \in \Omega \\ Bu(x) &= g(x) \quad x \in \partial\Omega \end{aligned}$$



Other topics to consider



- The addition of constraints generally increases the size of the covariance matrix and makes the formulation more challenging. There are scalable numerical methods for GPR, including rank-reduced approximations, specialized methods like subset-of-data and inducing point methods, and hierarchical decomposition for non-Gaussian likelihoods (using Kronecker products). These are described in the survey paper.
- State-of-the-art methods for sampling and computing the truncated multivariate normal, which arises in many types of constrained GPR.
- Bayesian approaches for the identification of basis elements in the finite-dimensional approximation of PDE solutions. One such approach is Owhadi's 2015 paper titled "Bayesian Numerical Homogenization." He presents a Bayesian formulation of a numerical homogenization approach in which the optimal bases are shown to be polyharmonic splines and the optimal solution in the case of a white noise source term is a Gaussian field."
- Vector-valued GPs: Curl-free and divergence-free fields have been modeled with constrained GPs.

A curl-free constraint $\mathcal{L}_x f = \nabla \times f = 0$ for a vector field $f: \mathbb{R}^3 \rightarrow \mathbb{R}^3$. A curl free vector field can be written as $f = \nabla g$.

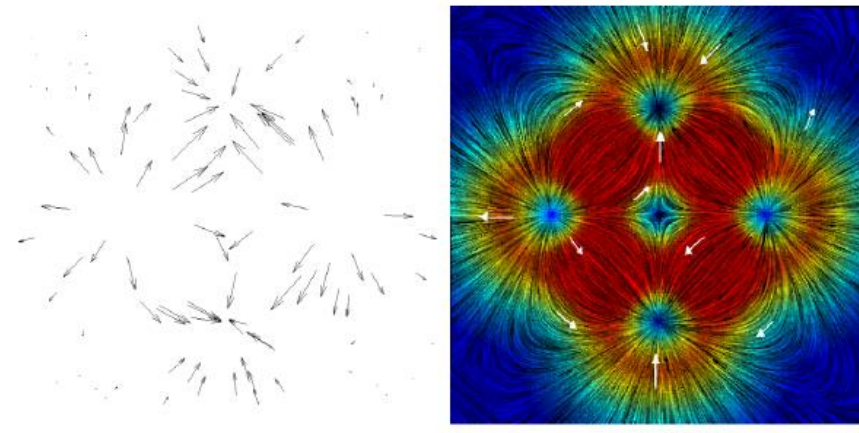


Figure from Macedo and Castro.



- In addition to supplementing limited or expensive scientific data, constraints help improve the generalizability of the model in ways that simply increasing dataset size may not.
- Our survey focused on several important classes of constraints for Gaussian processes. These included positivity or bound constraints, monotonicity and convexity constraints, linear differential equation constraints, and boundary value constraints.
- Constraints can be enforced in an implicit way through data that satisfies the constraint, by construction of a tailored sample space, by derivation of a constrained covariance kernel, or by modifying the output or likelihood of the Gaussian process.
- The constraints may be enforced in a “global sense”, at a finite set of “virtual” or “auxiliary” points, or only in an approximate sense. We have pointed to these aspects as key features distinguishing the constraints in this survey.
- Constraints introduce practical challenges into the construction of GPRs. The adaptation of computational strategies to constrained GPR is a relatively new field, and best practices have not yet been established. Constraints have not made their way into the most widely used production codes for GPR.
- Establishing best practices and furthering these computational aspects of constrained GPR is a promising area.

References (these are some from the 95 references in the Survey.



- SURVEY PAPER: L.P. Swiler, M. Gulian, A. Frankel, C. Safta, and J.D. Jakeman. A Survey of Constrained Gaussian Process Approaches and Implementation Challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2):119–156 (2020). DOI: [10.1615/JMachLearnModelComput.2020035155](https://doi.org/10.1615/JMachLearnModelComput.2020035155)
- Da Veiga, S. and Marrel, A., Gaussian Process Modeling with Inequality Constraints, *Annales Faculte Sci. Toulouse*, vol. **21**, pp. 529–555, 2012.
- Gramacy, R.B., *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Boca Raton, FL: CRC Press, 2020.
- Gulian, M., Frankel, A. and L. Swiler. Gaussian Process Regression constrained by Boundary Value Problems. arxiv.org/abs/2012.11857
- Lopez-Lopera, A.F., Bachoc, F., Durrande, N., and Roustant, O., Finite-Dimensional Gaussian Approximation with Linear Inequality Constraints, *SIAM/ASA J. Uncertainty Quant.*, vol. **6**, no. 3, pp. 1224–1255, 2018.
- Maatouk, H. and Bay, X., Gaussian Process Emulators for Computer Experiments with Inequality Constraints, *Math. Geosci.*, vol. **49**, no. 5, pp. 557–582, 2017.
- Macedo, I. and Castro, R., Learning Divergence-Free and Curl-Free Vector Fields with Matrix-Valued Kernels, Tech. rep., Instituto Nacional de Matematica Pura e Aplicada, 2008.
- Owhadi, H. Bayesian Numerical Homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015.
- Pensoneault, A., Yang, X., and Zhu, X., Nonnegativity-Enforced Gaussian Process Regression, 2020. arXiv: 2004.04632
- Quinonero-Candela, J. and Rasmussen, C.E., Analysis of Some Methods for Reduced Rank Gaussian Process Regression, *Switching and Learning in Feedback Systems*, Berlin: Springer, pp. 98–127, 2005a.
- Raissi, M., Perdikaris, P., and Karniadakis, G.E., Machine Learning of Linear Differential Equations Using Gaussian Processes, *J. Comput. Phys.*, vol. **348**, pp. 683–693, 2017.
- Rasmussen, C.E. and Williams, C.K., *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press, 2006.
- Riihimaki, J. and Vehtari, A., Gaussian Processes with Monotonicity Information, *Proc. of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 645–652, 2010.
- Solin, A. and Sarkka, S., Hilbert Space Methods for Reduced-Rank Gaussian Process Regression, *Stat. Comput.*, 2019.