SAND2020-8541C

A Survey of Constrained Gaussian Process Regression: Approaches and Implementation Challenges

M. Gulian, L. Swiler, A. Frankel, C. Safta, J. Jakeman

SAND XXXX-XXXX X

PhILMs Webinar, August 17 2020.

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

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 Energys National Nuclear Security Administration under contract DE-NA-0003525.

Introduction

Sandia National Laboratories

- Tremendous recent surge in the development and application of machine learning 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.
- 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.

Introduction

Sandia National Laboratories

- 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 (e.g. tens to hundreds), which consists of pairs of input parameters and corresponding response values.
- The GPR can be thought of as a machine-learned metamodel and used to provide fast, cheap function evaluations for the purposes of prediction, sensitivity analysis, uncertainty quantification, calibration, and optimization.

GP Regression: Concept

Sandia National Laboratories

- 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 $\{f(x) \mid x \in X\}$ for which, given any finite set of N inputs $X = \{x_1, x_2, ..., x_N\}$, $x_i \in \mathbb{R}^d$, the collection $f(x_1), f(x_2), ..., f(x_N)$ 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.

GP Regression: Definition

- The prediction $f = [f(x_1), f(x_2), \ldots f(x_N)]^\top$ of a Gaussian process with mean function $m(x)$ and a covariance function $k(x, x')$ is a random variable such that

$$p(f|X) = \mathcal{N}(f; m(X), k(X, X)), \tag{1}$$

  where $m(X)$ denotes the vector $[m(x_1), \ldots, m(x_N)]^\top$ and $k(X, X)$ denotes the matrix with entries $[k(x_i, x_j)]_{1 \leqslant i,j \leqslant N}$.

- The multivariate normal probability distribution $\mathcal{N}(f; m, K)$ with mean vector m and covariance matrix K has the form

$$\mathcal{N}(f; m, K) = \frac{1}{(2\pi)^{N/2}|K^{1/2}|} \exp\left(-\frac{1}{2}(f - m)^\top K^{-1}(f - m)\right).$$

- The covariance kernel function k of a Gaussian process must be symmetric and positive semidefinite, e.g., the squared exponential kernel

$$k(x_i, x_j) = \eta^2 \exp\left[-\frac{1}{2}\sum_{\ell=1}^{d}\left(\frac{x_i{}^\ell - x_j{}^\ell}{\rho_\ell}\right)^2\right]. \tag{2}$$

GP Regression: Likelihood

- The distribution (1) for $p(f|X)$, determined by covariance kernel k and the mean m, is referred to as a *prior* for the GP.
- If the error or noise relating the actual observations $y = [y(x_1), y(x_2), ..., y(x_N)]^\top$ 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 y given the GP prior is given by

$$p(y|X, f) = \mathcal{N}(f, \sigma^2 I_N). \tag{3}$$

- Here, I_N denotes the $N \times N$ identity matrix. The distribution $p(y|X, f)$ is referred to the *likelihood* of the GP, and the Gaussian likelihood (3) is by far the most common. Specific non-Gaussian likelihood functions can be used to enforce certain types of constraints.

GP Regression: MLE

- The parameters in the covariance kernel function of a GP are referred to as *hyperparameters* of the GP. We denote them by θ. For the squared exponential kernel (2), the aggregate vector of hyperparameters is $\theta = [\eta, \rho_1, ..., \rho_d, \sigma]$, where we have included the likelihood/noise parameter σ from (3) as a hyperparameter.
- The marginal likelihood is given by

$$p(y|X, \theta) = \int p(y|X, f, \theta) p(f|X, \theta) df$$

and the log-marginal-likelihood for a GP with a zero-mean prior ($m \equiv 0$) can be written as

$$\log p(y|X, \theta) = -\frac{1}{2} y^\top \left(K(X, X) + \sigma^2 I_N\right)^{-1} y - \frac{1}{2} \log \left|K(X, X) + \sigma^2 I_N\right| - \frac{N}{2} \log 2\pi$$

- This can be optimized to give the most likely values of the hyperparameters given data. This is known as maximum likelihood estimation (MLE) of the hyperparameters.

GP Regression: Posterior

Sandia National Laboratories

- Once the hyperparameters of the GPR have been chosen, the *posterior* of the GP is given by Bayes' rule,

$$\mathsf{p}(\mathsf{f}|\mathsf{X},\mathsf{y},\boldsymbol{\theta}) = \frac{\mathsf{p}(\mathsf{f}|\mathsf{X},\boldsymbol{\theta})\mathsf{p}(\mathsf{y}|\mathsf{X},\mathsf{f},\boldsymbol{\theta})}{\mathsf{p}(\mathsf{y}|\mathsf{X},\boldsymbol{\theta})}. \tag{4}$$

- Given the prior $\mathsf{p}(\mathsf{f}|\mathsf{X},\boldsymbol{\theta})$ (1) and the Gaussian likelihood $\mathsf{p}(\mathsf{y}|\mathsf{X},\mathsf{f},\boldsymbol{\theta})$ (3), the prediction f^* of a GPR at a new point x^* can be calculated as

$$\begin{aligned}\mathsf{p}(\mathsf{f}^*|\mathsf{y},\mathsf{X},\mathsf{x}^*,\boldsymbol{\theta}) = \mathcal{N}\Big(&\mathsf{k}(\mathsf{x}^*,\mathsf{X})(\mathsf{K}(\mathsf{X},\mathsf{X})+\sigma^2\mathsf{I}_\mathsf{N})^{-1}\mathsf{y},\\ &\mathsf{k}(\mathsf{x}^*,\mathsf{x}^*)-\mathsf{k}(\mathsf{x}^*,\mathsf{X})(\mathsf{K}(\mathsf{X},\mathsf{X})+\sigma^2\mathsf{I}_\mathsf{N})^{-1}\left[\mathsf{k}(\mathsf{x}^*,\mathsf{X})\right]^\top\Big)\end{aligned}$$

- Note that the mean of this Gaussian posterior is the mean estimate $\mathbb{E}[\mathsf{f}(\mathsf{x}^*)]$ of the predicted function value f^* at x^* and the variance is the estimated prediction variance of the same quantity.

GP: Complete Example

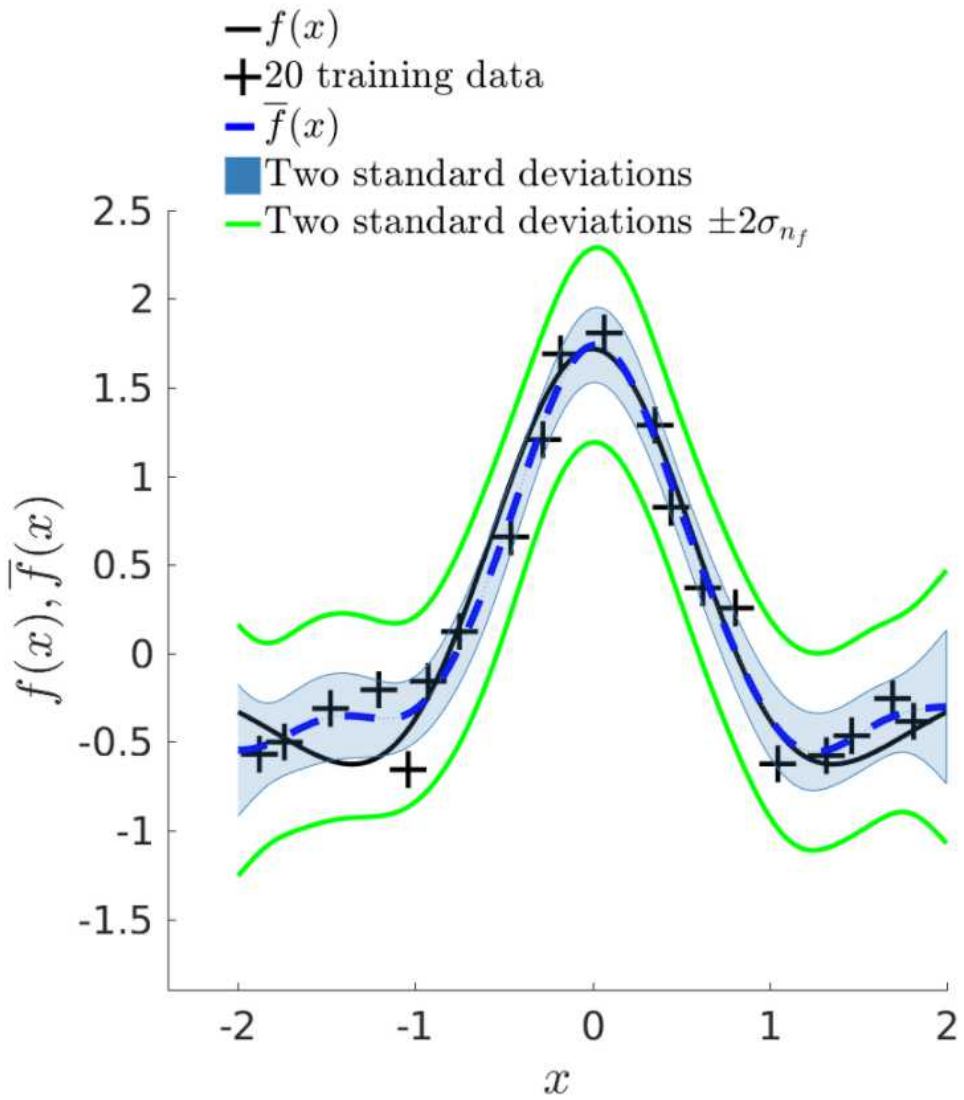


Figure: Noise is added to some locations on the black curve to generate data (black crosses). GPR fits a mean posterior to the data after filtering out some noise with a Gaussian likelihood, with the posterior variance giving an esimate of uncertainty in the prediction. The Gaussian likelihood allows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

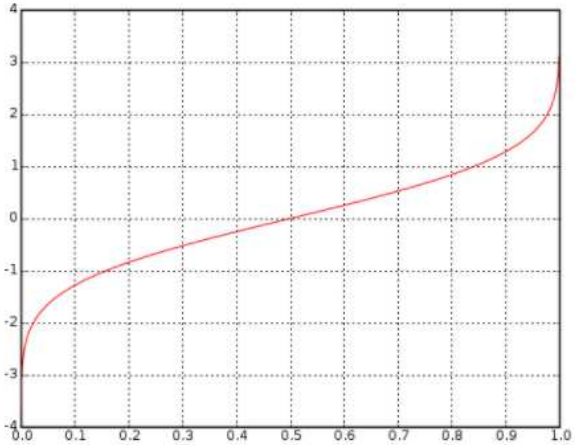
Strategies & Differences to look for

Sandia National Laboratories

- Each step of GPR – sample space/prior, likelihood, posterior – reviewed above 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.
- Some methods relax the global constraints to constraints at a finite set of "virtual" points; others transform the output of the GP to guarantee the predictions satisfy constraints, or construct a sample space of predictions in which every realization satisfies the constraints. This distinction between should be kept in mind when surveying constrained GPs.

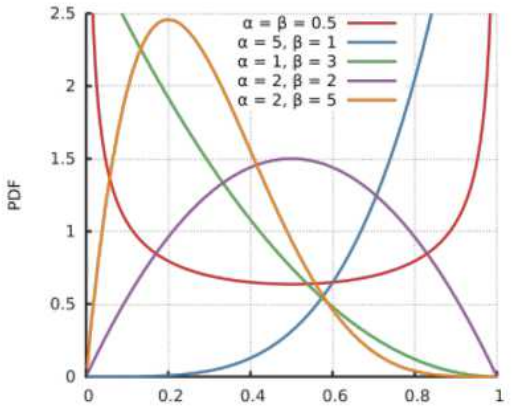
Bounds: Warped Output

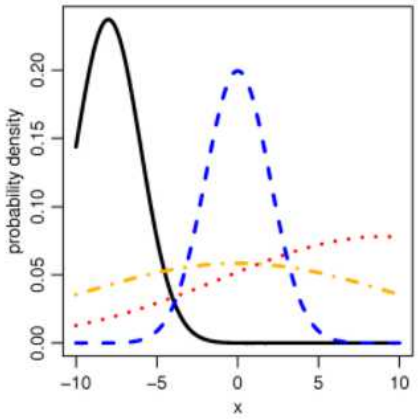
- Bound constraints of the form $a \leqslant f(x) \leqslant b$ over some region of interest arise naturally in many applications, such as chemical concentration data.
- Warping functions are used to transform bounded observations z_i to unbounded observations u_i which can be treated with unconstrained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



Bounds: Transformed Likelihood

- In addition to using warping functions, bound constraints can also be enforced using non-Gaussian likelihood functions $\mathfrak{p}(\mathsf{y}|\mathsf{X},\mathsf{f},\theta)$ that are constructed to produce GP observations which satisfy the constraints.
- 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
- Unlike the warping method, the posterior (4) is not analytically tractable; Laplace approximation and expectation propagation can be used for approximate inference with the posterior.





Bounds: Truncated MVN

- Since a Gaussian process is always trained and evaluated at a finite set of points X, a "global" of the form $a \leqslant f(x) \leqslant b \quad \forall x \in I$ can be approximated by constraints at a finite set of N_c auxiliary or "virtual" points $x_i, \ldots, x_{N_c} \in I$.
- This requires constructing an unconstrained GP and then, over the virtual points, transforming this GP to a *truncated* multivariate Gaussian distribution

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

- The unconstrained mean predictor is conditioned on the data (X, y):

$$\mathbb{E}\left[f(x^*) \mid f(X) = y\right]. \tag{5}$$

This setup is augmented by a fixed, finite set of discrete points $\{x_i\}_{i=1}^{N_c}$, and the predictor (5) is replaced by the predictor

$$\mathbb{E}\left[f(x^*) \mid f(X) = y \text{ and } a \leqslant f(x_i) \leqslant b \text{ for all } i = 1, 2, \ldots N_c\right]. \tag{6}$$

Bounds: Truncated MVN

- In general, sampling and computing the moments of $\mathcal{TN}(\mathbf{z}; \boldsymbol{\mu}, \Sigma, \mathbf{a}, \mathbf{b})$ is computationally demanding – rejection sampling becomes very expensive as the dimension increases. We survey this problem at length in our article.
- In contrast to the warping approaches or the spline approach below, which maintain a global enforcement of the constraints, the bounds in (6) can depend on the location: $a_i \leqslant f(x_i) \leqslant b_i$, representing different bounds in different regions of I.
- A downside of using the approach described here is that it is unclear how many virtual points x_i are needed to approximately constrain the GP globally with a prespecified level of confidence; some studies with increasing N_c are presented by Da Veiga et al. However, if the number of points can be chosen adequately, this approach can be used to enforce not only bound constraints but also monotonicity and convexity constraints.

Bounds: Splines

- 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

$$h(x) = \max(1 - |x|, 0)$$

  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

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

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

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

Bounds: Splines

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

$$A_{ij} = h_i(x_j).$$

  Then any set of spline coefficients ξ that satisfy the equation

$$A\xi = y$$

  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 ξ 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 y are governed jointly by a GP prior in the form

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

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

Bounds: Splines & Example

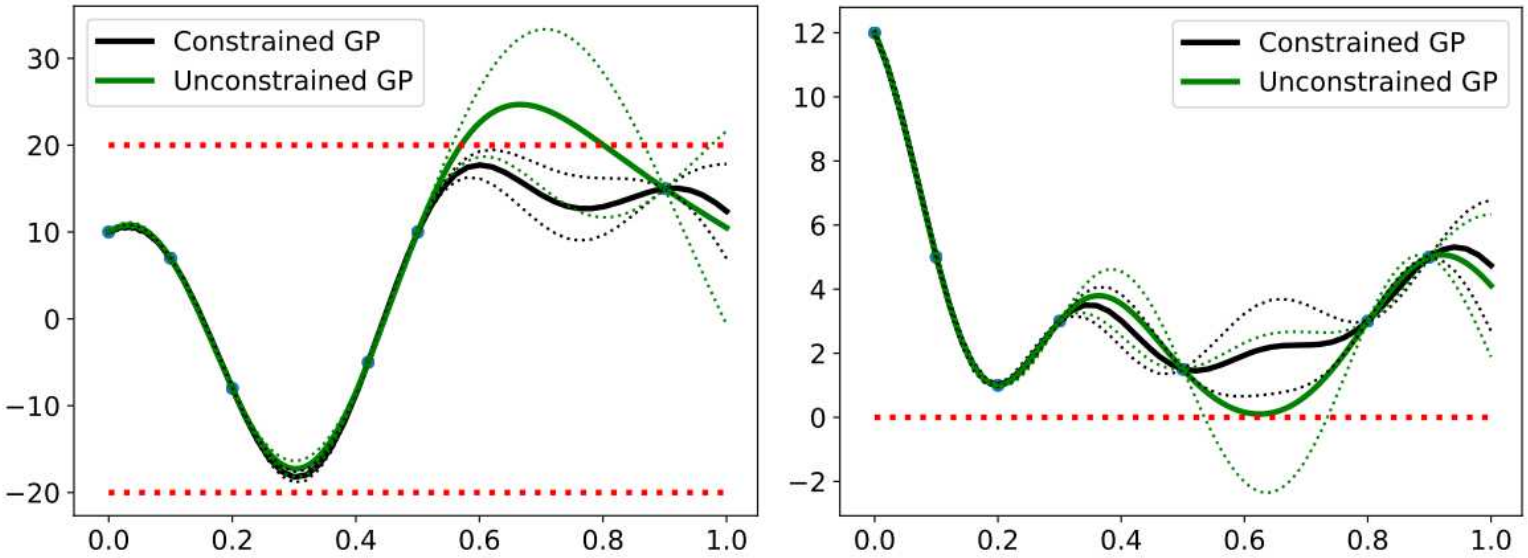
- Upon observation of the data y, the conditional distribution of the knot values subject to $\mathsf{y} = \mathsf{A}\boldsymbol{\xi}$ is given by

$$p(\boldsymbol{\xi} \mid \mathsf{y} = \mathsf{A}\boldsymbol{\xi}) = \mathcal{N}\Big(\boldsymbol{\xi}; \mathsf{K}\mathsf{A}^\top(\mathsf{A}\mathsf{K}\mathsf{A}^\top)^{-1}\mathsf{y}, \mathsf{K} - \mathsf{K}\mathsf{A}^\top(\mathsf{A}\mathsf{K}\mathsf{A}^\top)^{-1}\mathsf{A}\mathsf{K}\Big)$$

- In this case, we are now interested in evaluating the distribution further conditioned on the inequality constraints $\boldsymbol{\xi} \in \mathcal{C}$ given by

$$p(\boldsymbol{\xi} \mid \mathsf{y} = \mathsf{A}\boldsymbol{\xi}, \boldsymbol{\xi} \in \mathcal{C}) = \mathcal{TN}\Big(\boldsymbol{\xi}; \mathsf{K}\mathsf{A}^\top(\mathsf{A}\mathsf{K}\mathsf{A}^\top)^{-1}\mathsf{y}, \mathsf{K} - \mathsf{K}\mathsf{A}^\top(\mathsf{A}\mathsf{K}\mathsf{A}^\top)^{-1}\mathsf{A}\mathsf{K}, \mathcal{C}\Big)$$

- Again, we need to sample from the truncated multinormal distribution.



Monotonicity: Derivative Likelihood

- Monotonicity constraints are an important class of "shape constraints", e.g., the output of the Los Alamos National Laboratory "Lady Godiva" nuclear reactor is known to be monotonic with respect to the density and radius of the spherical uranium core.
- To enforce

$$\frac{\partial f}{\partial x_{d_i}}(x_i) \geqslant 0,$$

at a set of finite "operating" or virtual points $X_m = \{x_i\}_{i=1}^m$, we use the shorthand

$$f'_i = \frac{\partial f}{\partial x_{d_i}}(x_i), \quad \text{and} \quad f' = \left[\frac{\partial f}{\partial x_{d_1}}(x_1) \dots \frac{\partial f}{\partial x_{d_m}}(x_m)\right]^\top = \left[f'_1 \dots f'_m\right]^\top$$

and denote an observation of $f'_i = \partial f / \partial x_{d_i}(x_i)$ by y'_i.

- We use a likelihood

$$p\left(y'_i \middle| f'_i\right) = \Phi\left(f'_i \frac{1}{\nu}\right). \tag{7}$$

Here $\Phi(z)$ is the CDF of the standard normal distribution and approaches a step function as $\nu \to 0$.

- Note that the likelihood function in (7) forces the likelihood to be zero (for non-monotonicity) or one (for monotonicity) in most cases.

Monotonicity: Derivative Likelihood

- The joint prior is now given by:

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

where

$$f_{joint} = \begin{bmatrix} f \\ f' \end{bmatrix} \quad \text{and} \quad K_{joint} = \begin{bmatrix} K_{f,f} & K_{f,f'} \\ K_{f',f} & K_{f',f'} \end{bmatrix}. \tag{8}$$

Here, $K_{f,f} = k(X, X)$ where k denotes the covariance function of f. The $m \times m$ matrix $K_{f',f'}$ in (8) denotes the covariance matrix between the values of the specified partial derivatives of f at the operational points X_m:

$$[K_{f',f'}]_{ij} = \left[\mathrm{cov}\left(f'_i, f'_j\right)\right] = \left[\mathrm{cov}\left(\frac{\partial f}{\partial x_{d_i}}(x_i), \frac{\partial f}{\partial x_{d_j}}(x_j)\right)\right], \quad 1 \leqslant i, j \leqslant m.$$

- By linearity, $\frac{\partial f}{\partial x_{d_i}}$ is a GP with covariance matrix

$$\frac{\partial}{\partial x_{d_i}} \frac{\partial}{\partial x'_{d_j}} k(x, x'),$$

so that

$$[K_{f',f'}]_{ij} = \frac{\partial^2 k}{\partial x_{d_i} \partial x'_{d_j}}(x_i, x'_j), \quad 1 \leqslant i, j \leqslant m.$$

Monotonicity: Derivative Likelihood

- The $n \times m$ matrix $K_{f,f'}$ represents the covariance between f and f', and is given by

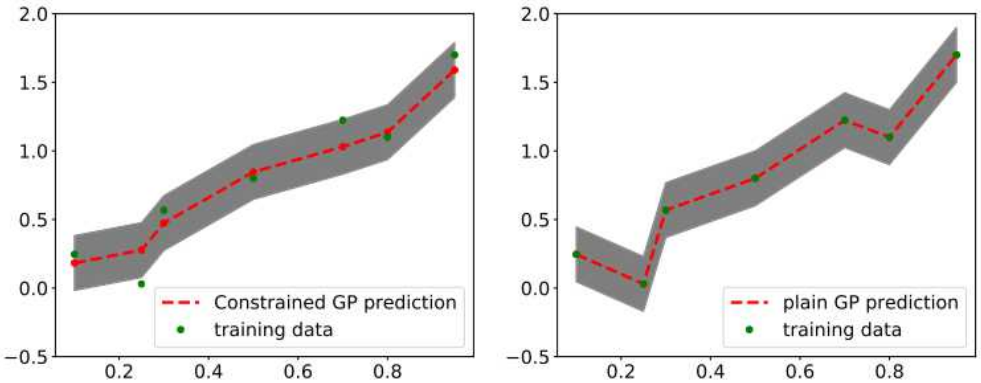
$$[K_{f,f'}]_{ij} = \frac{\partial k}{\partial x'_{d_j}}(x_i, x'_j), \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m,$$

  with $K_{f',f} = K_{f,f'}^\top$ representing the covariance between f' and f.

- The posterior probability of the joint distribution is

$$p(f, f'|y, y') = \frac{1}{Z} p(f, f'|X, X_m) p(y|f) p(y'|f')$$

  where $1/Z$ is a normalizing constant. This distribution is analytically intractable because of the non-Gaussian likelihood for the derivative components; MCMC, Laplace approximation, and expectation propagation can be applied.



Monotonicity: Other Approaches

Sandia National Laboratories

- Roughly speaking, given a method to enforce bound constraints, monotonicity constraints can be enforced by utilizing this method to enforce $\mathsf{f}' \geqslant 0$ on the derivative of the Gaussian process in a "co-kriging" setup for the joint GP $[\mathsf{f}; \mathsf{f}']$.
- 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 $\mathsf{f}'' \geqslant 0$, but more complicated in higher dimensions, where convexity becomes a *nonlinear* constraint between the second partials of a GP.

Linear PDE Constraints

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

$$\mathcal{L}u = f \tag{9}$$

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

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

the equation (9) can be used to constrain GP predictions to satisfy known physical laws expressed as linear partial differential equations.

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

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

and if $m(\cdot)$ and $k(\cdot, x')$ 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 $\mathcal{L}u$:

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

Linear PDE Constraints

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

$$\text{mean}\left(\mathcal{L}[\mathfrak{u}](x)\right) = \mathbb{E}\left[\mathcal{L}[\mathfrak{u}](x)\right] = \mathcal{L}\mathbb{E}\left[\mathfrak{u}(x)\right] = \mathcal{L}\mathfrak{m}(x).$$

- The covariance would be given by

$$\begin{aligned}
\text{cov}\left(\mathcal{L}[\mathfrak{u}](x), \mathcal{L}[\mathfrak{u}](x')\right) &= \mathbb{E}\left[\mathcal{L}_x[\mathfrak{u}(x)]\mathcal{L}_{x'}[\mathfrak{u}(x')]\right] \\
&= \mathbb{E}\left[\mathcal{L}_x\mathcal{L}_{x'}\left[\mathfrak{u}(x)\mathfrak{u}(x')\right]\right] \\
&= \mathcal{L}_x\mathbb{E}\left[\mathcal{L}_{x'}\left[\mathfrak{u}(x)\mathfrak{u}(x')\right]\right] \\
&= \mathcal{L}_x\mathcal{L}_{x'}\mathbb{E}\left[\mathfrak{u}(x)\mathfrak{u}(x')\right] \\
&= \mathcal{L}_x\mathcal{L}_{x'}\left[\text{cov}\left(\mathfrak{u}(x), \mathfrak{u}(x')\right)\right] \\
&= \mathcal{L}_x\mathcal{L}_{x'}\mathfrak{k}(x, x').
\end{aligned}$$

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

Linear PDE Constraints

- If scattered measurements y_f on the source term f in (9) 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 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 $k(x, x')$ for u, the covariance kernel of this joint GP is

$$k\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}\right) = \begin{bmatrix} k(x_1, x_1') & \mathcal{L}_{x'} k(x_1, x_2') \\ \mathcal{L}_x k(x_2, x_1') & \mathcal{L}_x \mathcal{L}_{x'} k(x_2, x_2') \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}.$$

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

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

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 $u = \frac{1}{8}[(2x-1)^2 - 1]$.

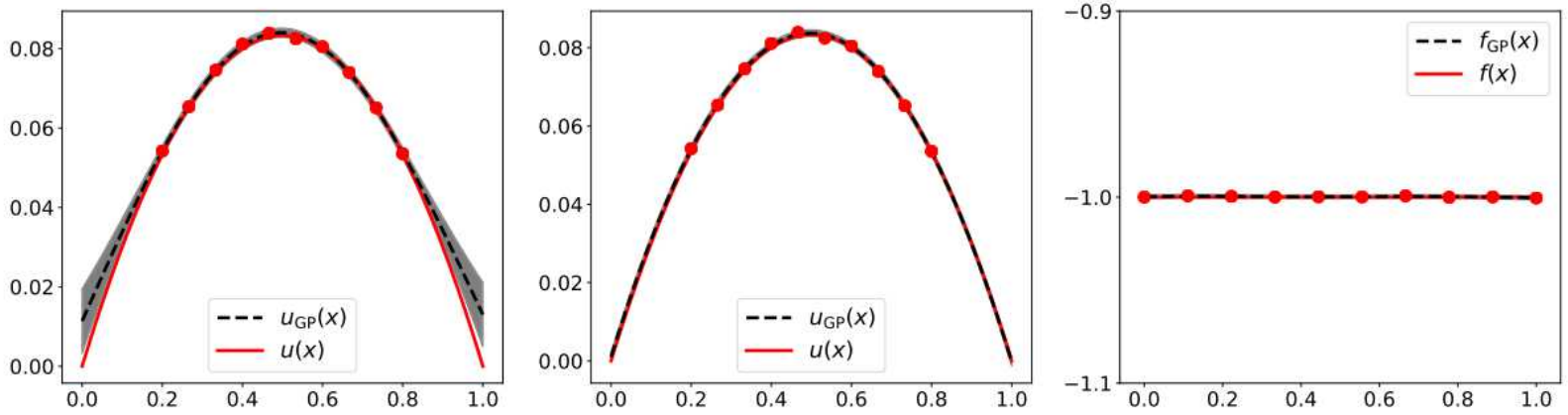


Figure: **Left:** Reconstruction of u (red line) with an unconstrained GP (black line) using 10 data points (red dots) in $[0.2, 0.8]$.
Center: Reconstruction of u (red line) with a PDE constrained GP (black line) using the same 10 data points (red dots) in $[0.2, 0.8]$.
Right: Right-hand side f of the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

PDEs: Transformed covariance

- Given a linear operator $\mathcal{L}_x$ and a vector-valued GP f described using a matrix-valued covariance kernel function that encodes the covariance between the entries of the vector f, the constraint

$$\mathcal{L}_x f = 0$$

is satisfied if f can be represented as

$$f = \mathcal{G}_x g,$$

for a transformation $\mathcal{G}_x$ such that

$$\mathcal{L}_x \mathcal{G}_x = 0.$$

- In other words, the range of the operator $\mathcal{G}_x$ lies in the nullspace of the operator $\mathcal{L}_x$. Further, provided that $\mathcal{G}_x$ is also a linear operator, if g is a GP with covariance kernel k_g, then f is also a GP with covariance kernel

$$k_f = \mathcal{G}_x k_g \mathcal{G}_{x'}^{\top}.$$

Constraints for vector-valued GPs

- Curl-free constraint $\mathcal{L}_x \mathsf{f} = \nabla \times \mathsf{f} = 0$ for a vector field $\mathsf{f} : \mathbb{R}^3 \to \mathbb{R}^3$. A curl-free vector field can be written $\mathsf{f} = \nabla \mathsf{g}$.
- In a similar way, one can enforce a divergence-free condition $\nabla \cdot \mathsf{f} = 0$ for a vector-valued GP f by writing $\mathsf{f} = \nabla \times \mathsf{g}$ and placing a GP prior on a vector field g, as $\nabla \cdot (\nabla \times \mathsf{g}) = 0$.
- When appropriate square-exponential covariance kernel is used for the GP g, curl-free and div-free covariance kernels for the GP f can be derived analytically.

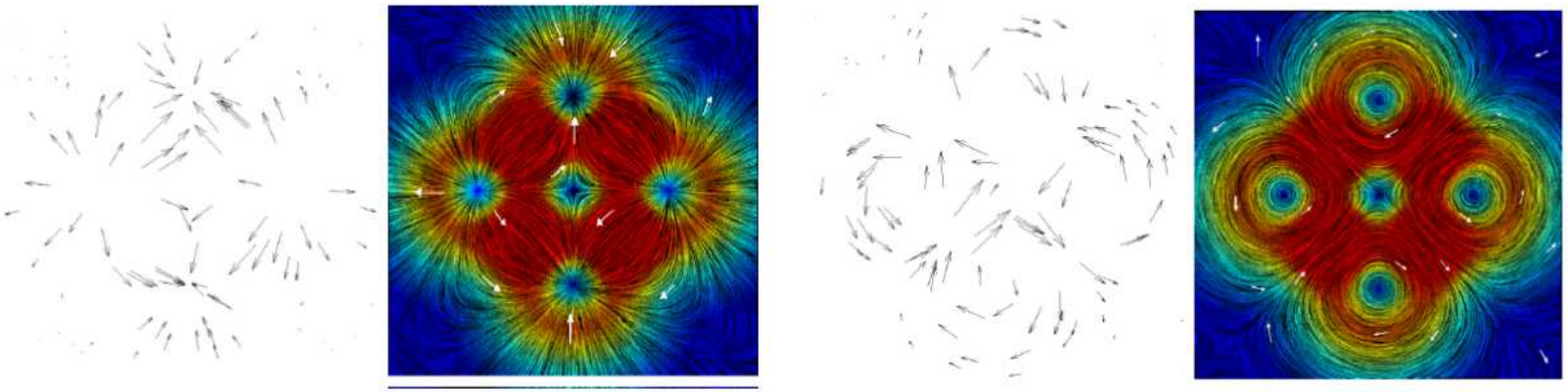


Figure: Curl-free (left) and div-free (right) GP vector field regression, from Macedo and Castro.

PDEs: Empirical Covariance

Sandia National Laboratories

- Given an ensemble of realizations of a random field Y on a set of grid points and a smaller set of high-fidelity data on a subset of the low-fidelity grid points, Yang et al. build a Gaussian process for the unknown field over the unstructured grid which also passes through the high-fidelity data, at the same time ensuring that the GP satisfies the PDE used to generate the low-fidelity ensemble.
- The idea is to compute the mean and covariance of the GP empirically on the grid from these realizations of the random field Y. We assume that we have M realizations $Y^m(x)$ of the output field $Y(x)$ for x in the d-dimensional grid $\{x_i\}_{i=1}^N$ (the low-fidelity data). Then the mean and covariance, respectively, are given by

$$\mu(x) \approx \mu_{MC}(x) = \frac{1}{M}\sum_{m=1}^{M} Y^m(x)$$

$$k(x,x') \approx k_{MC}(x,x') = \frac{1}{M-1}\sum_{m=1}^{M} (Y^m(x) - \mu_{MC}(x))(Y^m(x') - \mu_{MC}(x')).$$

PDEs: Empirical Covariance

Sandia National Laboratories

- Yang et al. have shown that physical constraints in the form of a deterministic linear operator are guaranteed to be satisfied within a certain error in the resulting prediction when using this approach.
- As the method uses an empirical mean and covariance, there is no need to infer the hyperparameters of a covariance function. However, it *cannot* interpolate for the field between the points where the stochastic realizations are available. The step of GPR for prediction at an arbitrary point x^* is not available, as the covariance kernel function is bypassed entirely.
- This is an example of an "implicit constraint" – proving that if the data used in GPR satisfies a PDE, then the GPR must satisfy the PDE as well (within a certain tolerance). Another example is the work of Salzmann and Urtasun, who considered GPR for pose estimation under rigid (constant angle and length) and non-rigid(constant length) constraints between points. They proved that if the data used in the GPR satisfies such constraints, the posterior prediction of the GPR satisfies them as well.

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 $k(x, x') = k(|x - x'|)$ 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,

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

- This enters into the approximation of the kernel:

$$k(x, x') \approx \sum_{\ell=1}^{m} s(\lambda_\ell) \phi_\ell(x) \phi_\ell(x'), \tag{10}$$

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

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 $\{(x_i, y_i)\}_{i=1}^n$, the covariance matrix is approximated using (10) as

$$K_{ij} = k(x_i, x_j) \approx \sum_{\ell=1}^{m} \phi_\ell(x_i) s(\lambda_\ell) \phi_\ell(x_j).$$

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

$$\Phi_{i\ell} = \phi_\ell(x_i), \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant \ell \leqslant m,$$

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

$$K \approx \Phi \Lambda \Phi^\top.$$

Boundary Value Constraints

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

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

  where the $m \times 1$ vector Φ_* is defined by

$$[\Phi_*]_\ell = \phi_\ell(x^*), \quad 1 \leqslant \ell \leqslant m.$$

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

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

Implementation Challenges

Sandia National Laboratories

Constraints introduce new practical challenges into the GPR framework:

- The analytical construction of sample spaces, transformations, or covariance kernels that in-herently provide constraints
- the sampling of truncated multivariate normals or intractable posterior distributions that arise when using non-Gaussian likelihoods;
- increased data and covariance matrix size when enforcing constraints with "virtual" data that leads to expanded"four-block" covariance;
- MLE (hyperparameter optimization) with likelihood functions that implement the constraint;
- calculation of eigenvalues/eigenfunctions in bounded domains withcomplex geometry.

Truncated Multinormal

Sandia National Laboratories

- Given a positive-definite covariance matrix Σ and a set $S \subset \mathbb{R}^d$, the truncated normal distribution $\mathcal{TN}(\mu, \Sigma, S)$ is the conditional distribution of the random variable $x \sim \mathcal{N}(\mu, \Sigma)$ given $x \in S$:

$$\mathcal{TN}(x; \mu, \Sigma, S) = \frac{\mathbb{1}_S(x)}{C} \mathcal{N}(x; \mu, \Sigma).$$

- The normalization constant is given by

$$\begin{aligned} C &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_d}^{b_d} \mathcal{N}(x; \mu, \Sigma)\, dx_1 dx_2 \dots dx_d \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_d}^{b_d} \exp\left(-\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu) \right) dx_1 dx_2 \dots dx_d \end{aligned}$$

- Calculating values of the distribution $\mathcal{TN}$ is called for in constrained maximum likelihood estimation of the GPR hyperparameters.
- Sampling $\mathcal{TN}$ is needed for posterior prediction in several approaches discussed above.

$\mathcal{TN}$ & Constrained MLE

- Naive Monte Carlo methods (like rejection sampling from the mode) scale poorly to higher dimensions.
- Several Markov Chain Monte Carlo (MCMC) methods were studied by Lopez-Lopera et al; comparison of expected sample size metrics suggested that Hamiltonian Monte Carlo (HMC) is the most efficient sampler in the setting of that article.
- For maximum likelihood estimation of hyperparameters within the spline approach, Bayes' rule yields the following constrained log-marginal-likelihood function:

$$\begin{aligned}\mathcal{L}_{\text{cMLE}} &= \log \mathsf{p}_{\boldsymbol{\theta}}(\mathsf{y}|\boldsymbol{\xi} \in \mathcal{C}) \\ &= \log \mathsf{p}_{\boldsymbol{\theta}}(\mathsf{y}) + \log \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}|\Phi\boldsymbol{\xi} = \mathsf{y}) - \log \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}) \quad (11) \\ &= \mathcal{L}_{\text{MLE}} + \log \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}|\Phi\boldsymbol{\xi} = \mathsf{y}) - \log \mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}).\end{aligned}$$

- Unlike the sampling of $\mathcal{TN}$, for which computing such integrals can be avoided with MCMC, calculation of Gaussian orthant probabilities is unavoidable if the user wants to train the hyperparameters with a constrained likelihood function.

Constrained MLE

Sandia National Laboratories

- A thorough discussion of numerical approaches to truncated Gaussian integrals is Genz et al.; Lopez-Lopera utilize the minimax exponential tilting method of Botev, reported to be feasible for quadrature of Gaussian integrals in dimensions as high as 100, to compute the Gaussian orthant probabilities in (11) and compare cMLE with MLE.
- Another current drawback of cMLE is that the gradient of $\mathcal{L}_{\text{cMLE}}$ is not available in closed form, unlike the gradient of $\mathcal{L}_{\text{MLE}}$. Thus, in Lopez-Lopera, MLE was performed using a L-BFGS optimizer, while cMLE was performed using the method of moving asymptotes.
- Lopez-Lopera et al. also studied under which conditions MLE and cMLE yield consistent predictions of certain hyperparameters for fixed-domain asymptotics; they show that MLE and cMLE yield consistent hyperparameters in this limit for the case of boundedness, monotonicity, and convexity constraints, and suggest quantitative tests to determine if the number of data points is sufficient to suggest unconstrained MLE as opposed to the more expensive cMLE.

Scalable Inference

Sandia National Laboratories

- Inference in GPR using the entire training dataset (of size N) scales as N^3 due to covariance matrix inversion.
- This is exacerbated by certain methods to enforce constraints, such as the linear PDE constraints, which require the inclusion of "virtual" constraint points in the training data.
- There have been few studies on improving scalability of constrained GPs. We mention several promising approaches and possible applications to constrained GPs.
- Some strategies, including thesubset of data approach, the inducing point approach, and the spectral expansion approach, are specific to covariance matrices of GPs. Other methods are based on general linear algebra techniques.

Scalable Inference

Sandia National Laboratories

- One notable feature of increasing the density of training data is that the covariance matrixtends to become more ill-conditioned, the result of partially redundant information being added to the matrix. In such situations it is worthwhile to identify a **subset of data** that minimizes prediction error subject to a maximum dataset size constraint.
- Subset-of-data approaches can be based on greedy methods or local approximation.
- Inducing point methods model the data as being conditionally dependent on a few inducing points
- More generic linear algebra methods that can be applied include singular value decompositions, hierarchical matrices, optimization with L^1 regularization, and Gaussian Markov random fields.
- Hierarchical decompositions have been applied for GPR with non-Gaussian likelihoods in tensor-product grids.

Summary

Sandia National Laboratories

- 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.
- Some theoretical properties are not fully understood.

Future Directions

Sandia National Laboratories

- Constraints introduce new practical challenges into GPR.
- Construction of sample spaces, transformations, or covariance kernels that inherently provide constraints; sampling of truncated multivariate normals or intractable posterior distributions from non-Gaussian likelihoods; increased data and covariance matrix size when using "virtual" data that leads to expanded "four-block" covariance; calculation of eigenvalues/eigenfunctions in bounded domains with complex geometry; placement of virtual points or construction of spline grids in higher dimensions; and MLE (optimization) of hyperparameters.
- 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.

Acknowledgements

- This work was completed with funding granted under Sandia's Laboratory Directed Research and Development program.
- Thank you to the organizers.
- Thank you for your attention!

- See our article `https://arxiv.org/abs/2006.09319` for complete discussion for all of the methods discussed in this presentation, presented in roughly the same order as this presentation.
- Disclaimer 1: most of the strategies and claims presented here are not original. The survey article provides 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.