Your Paper

You

August 16, 2023

Abstract

1 Introduction

Our situation is that we have obtained measurements, potentially noisy, of both the u and f. However, we do not have knowledge of the ODE parameters ϕ. This is a typical situation in for example lab work, where we have measurements and want to determine an underlying physical parameter or interpolate the data. This is often done with typical linear regression techniques.

2 Fundamentals

2.1 Gaussian Process

As defined in [6] a Gaussian Process is a collection of random variables for which every finite set of variables have a joint multivariate Gaussian distribution. For a Regression model we define a function $f(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathbb{R}$ with the input set $\boldsymbol{X} = \{\boldsymbol{x_n} \in \mathcal{X}\}_{n=1}^{N}$. We now let $f : \mathcal{X} \rightarrow \mathbb{R}$ be our Gaussian Process. To fully define the GP we need to specify a mean function $m(\boldsymbol{x})$ and a covariance function or kernel $k(\boldsymbol{x}, \boldsymbol{x'})$ for $f(\boldsymbol{x})$:

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ k(\boldsymbol{x}, \boldsymbol{x'}) &= \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))] \end{aligned} \tag{1}$$

In this work the mean function $m(\boldsymbol{x})$ of the GP is set to zero. This is quite common, mainly for notational reasons [6].
So what we are left with for $m(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x'})$ is:

$$\begin{aligned} m(\boldsymbol{x}) &= 0 \\ k(\boldsymbol{x}, \boldsymbol{x'}) &= \mathbb{E}[f(\boldsymbol{x}) f(\boldsymbol{x'})] \end{aligned} \tag{2}$$

We now write our Gaussian Process $f(\boldsymbol{x})$ as:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x'})) \tag{3}$$

It is important to note that $k(\boldsymbol{x}, \boldsymbol{x'})$ has to be a positive definite in order to be a valid kernel function. Because of the definition of the GP we can now generate a joint Gaussian distribution for a finite set of points $\boldsymbol{X_*}$:

$$p(\boldsymbol{f_*} | \boldsymbol{X_*}) = \mathcal{N}(\boldsymbol{f_*} | \boldsymbol{0}, \boldsymbol{K_*}) \tag{4}$$

with $[\boldsymbol{K}]_{i,j} = k(\boldsymbol{x_{i*}}, \boldsymbol{x_{j*}})$.

2.2 Regression

The goal of regression is given a training dataset $\mathcal{D} = \{(\boldsymbol{x_i}, y_i) | i = 1, \ldots, N\}$ we want to make predictions for new test points $\boldsymbol{x_*}$ which are not in the training set. There are several methods which can be used to tackle this kind of problem. For example the basic linear regression using least squares and specific basis functions or a neural network that does regression. In this work we make use of the Gaussian process which is a non-parametric model. We will start with a Gaussian Process prior and then use the rules for Gaussian multivariate distributions to define a posterior predictive distribution, which we can then use to make predictions for new test points.

The general setup for regression is that we make observations/measurements y of an underlying function $f(\boldsymbol{x})$ at points $\boldsymbol{x}$ which are corrupted by a zero mean Gaussian noise with variance σ_n^2:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}) + \mathcal{N}(0 | \sigma_n^2 \boldsymbol{I}) \tag{5}$$

When we set the GP Prior onto the function f we have to add the normally distributed noise. To get the prior for the observations $\boldsymbol{y}$ we simply need to change the covariance function to $\operatorname{cov}(\boldsymbol{y_p}, \boldsymbol{y_q}) = k(\boldsymbol{x_p}, \boldsymbol{x_q}) + \sigma_n^2 \delta_{pq}$.
As described before in equation 4 we can generate a Gaussian random vector for new test data $\boldsymbol{X_*}$.

Both $\boldsymbol{y}$ and $\boldsymbol{f_*}$ are now multivariate Gaussians. We can therefore introduce the joint distribution of the observed data and the values of f at the test points $\boldsymbol{X_*}$ as:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f_*} \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2 \boldsymbol{I} & K(\boldsymbol{X},\boldsymbol{X_*}) \\ K(\boldsymbol{X_*},\boldsymbol{X}) & K(\boldsymbol{X_*},\boldsymbol{X_*}) \end{bmatrix}\right) \tag{6}$$

From this joint Gaussian prior we can then calculate the predictive posterior distribution by applying the rules for conditioning Gaussians. A good description for this conditioning can be found in [1] (chapter 2.3.1). The predictive distribution is then given by:

$$\begin{aligned} p(\boldsymbol{f_*}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{X_*}) &= \mathcal{N}(\boldsymbol{f_*}|\boldsymbol{\mu_*},\boldsymbol{\Sigma_*}) \\ \boldsymbol{\mu_*} &= K(\boldsymbol{X_*},\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I}]^{-1}\boldsymbol{y} \\ \boldsymbol{\Sigma_*} &= K(\boldsymbol{X_*},\boldsymbol{X_*}) - K(\boldsymbol{X_*},\boldsymbol{X})[K(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I}]^{-1}K(\boldsymbol{X},\boldsymbol{X_*}) \end{aligned} \tag{7}$$

We now have obtained the predictive mean $\boldsymbol{\mu_*}$ and the predictive covariance for the new test inputs $\boldsymbol{X_*}$.

2.3 Kernels

In [6] a kernel is defined as a function $k(\boldsymbol{x},\boldsymbol{x'})$, which maps two input pairs $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{x'} \in \mathcal{X}'$ into $\mathbb{R}$. To be a valid covariance function the kernel needs to satisfy two main conditions. First it needs to be symmetric so that $k(\boldsymbol{x},\boldsymbol{x'}) = k(\boldsymbol{x},\boldsymbol{x'})$ and second it needs to be positive semidefinite (PSD) The general covariance function used in this work is the squared exponential covariance function (SE) as defined in [6]:

$$k(\boldsymbol{x_p},\boldsymbol{x_q}) = \sigma_f^2 \exp(-0.5(\boldsymbol{x_p} - \boldsymbol{x_q})^\intercal M(\boldsymbol{x_p} - \boldsymbol{x_q})) \tag{8}$$

with the variance hyperparameter σ_f^2 and the matrix $M = \text{diag}(\boldsymbol{l})^{-2}$ containing the vector with the characteristic length scales l_i. The hyperparameters are collected in the vector $\boldsymbol{\theta} = (\boldsymbol{l}, \sigma_f^2)$.
For the one dimensional case the SE covariance function simply becomes:

$$k(x_p,x_q) = \sigma_f^2 \exp\left(-\frac{0.5}{l^2}(x_p - x_q)^2\right) \tag{9}$$

For a two-dimensional case with time being the second dimension, two independent length scale parameters are used for the separate space and time domains $\boldsymbol{l} = (l_x, l_t)$. For this case, we can re-write equation 7 the covariance function with $\boldsymbol{x_p} = (x_p, t_p)^\intercal$ and $\boldsymbol{x_q} = (x_q, t_q)^\intercal$ as the product of two single SE covariance functions:

$$k(\boldsymbol{x_p},\boldsymbol{x_q}) = \sigma_f^2 \exp\left(-\frac{0.5}{l_x^2}(x_p - x_q)^2\right) \exp\left(-\frac{0.5}{l_t^2}(t_p - t_q)^2\right) \tag{10}$$

There are many kernels one can choose from. A good overview over some kernel functions, as well as how they can be modified, can be found in [2] as well as in [4]. Later we will modify the SE kernels with linear operators, more specifically with differential operators.

2.4 Marginal likelihood

For the optimization of the hyperparameters θ and ϕ the negative marginal log likelihood (*nmll*) [6] is minimized. The *nmll*

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^\intercal \boldsymbol{K_y}^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K_y}| - \frac{n}{2}\log(2\pi) \tag{11}$$

with $K_y = K_f + \sigma_n^2\boldsymbol{I}$. As described in [6], the *mll* consists of three parts, which can each be understood separately. The first term $\frac{1}{2}\boldsymbol{y}^\intercal\boldsymbol{K_y}^{-1}\boldsymbol{y}$ is in charge of the data fit, the second term $\log|K_y|$ is the complexity penalty and the last term is a normalization term.

2.5 Linear operators and GPs

A useful property of Gaussian Processes is that their linear transformation is still a Gaussian Process. This main

3 Physics informed framework

We choose our problem setup similar to the one proposed in [5]. We have some sort of linear differential equation or ODE of a physical model:

$$\mathcal{L}_{\boldsymbol{x}}^{\phi} u(\boldsymbol{x}) = f(\boldsymbol{x}) \tag{12}$$

with $\mathcal{L}_{\boldsymbol{x}}^{\phi}$ being the differential operator that resembles the differential equation and ϕ being a list containing the a-priori unknown parameters which also define the differential equation. The function $f(\boldsymbol{x})$ acts as the forcing term and $u(\boldsymbol{x})$ is the solution of the differential equation.

We now start by assuming that $u(\boldsymbol{x})$ is a Gaussian Process of the form:

$$u(\boldsymbol{x}) \sim \mathcal{GP}(\mathbf{0}, k_{uu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta})) \tag{13}$$

with $k_{uu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta})$ being the kernel depending on its hyperparameters θ for the $\mathcal{GP}$ (on) $u(\boldsymbol{x})$. With the known linear relation between u and f from equation 12, we can now say that $f(\boldsymbol{x})$ must also be $\mathcal{GP}$ [6] of form:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(\mathbf{0}, k_{ff}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta})) \tag{14}$$

with now k_{ff} being the kernel defining the $\mathcal{GP}$ (on) $f(\boldsymbol{x})$. As shown in [3], [7] and [5] k_{ff} is defined by:

$$k_{ff}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_{\boldsymbol{x}}^{\phi} \mathcal{L}_{\boldsymbol{x'}}^{\phi} k_{uu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}) \tag{15}$$

We will later need the mixed terms between $u(x)$ and $f(x)$, namely k_{uf} and k_{fu} which are formed by:

$$\begin{aligned} k_{uf}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}, \boldsymbol{\phi}) &= \mathcal{L}_{\boldsymbol{x'}}^{\phi} k_{uu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}) \\ k_{fu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}, \boldsymbol{\phi}) &= \mathcal{L}_{\boldsymbol{x}}^{\phi} k_{uu}(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}) \end{aligned} \tag{16}$$

To simplify the notation, the dependence on the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is not written explicitly anymore. With this construction we now have two Gaussian Processes. As described in [3] we can now create a joint Gaussian process of u and f in the form: (not sure if I understood this correctly)

$$\begin{bmatrix} u \\ f \end{bmatrix} \sim \mathcal{GP}\left(\mathbf{0}, \begin{bmatrix} k_{uu} & k_{uf} \\ k_{fu} & k_{ff} \end{bmatrix}\right) \tag{17}$$

with $k_{uu} : \mathcal{U} \times \mathcal{U}$, $k_{uf} : \mathcal{U} \times \mathcal{F}$, $k_{fu} : \mathcal{F} \times \mathcal{U}$ and $k_{ff} : \mathcal{F} \times \mathcal{F}$ We now look at noisy observations of both u and f,with noise σ_u and σ_f, of the form:

$$\begin{aligned} y_{i,u} &= u_i(\boldsymbol{x_{i,u}}) + \mathcal{N}(0|\sigma_u^2) \\ y_{i,f} &= u_i(\boldsymbol{x_{i,f}}) + \mathcal{N}(0|\sigma_f^2) \end{aligned} \tag{18}$$

with which we create two data sets $\{\boldsymbol{X_u}, \boldsymbol{y_u}\}$ and $\{\boldsymbol{X_f}, \boldsymbol{y_f}\}$. Similarly, as we did before for the test points in equation 6 we can now write (create) a joint Gaussian distribution of the observations $\boldsymbol{y_u}$ and $\boldsymbol{y_f}$. We collect $\boldsymbol{X_u}, \boldsymbol{X_f}$ in $\boldsymbol{X}$ and $\boldsymbol{y_u}, \boldsymbol{y_f}$ in $\boldsymbol{y}$. With the joint GP at equation 17, the joint distribution will take the following form:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \sigma_{n_u}^2, \sigma_{f_u}^2, \boldsymbol{\phi}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k_{uu}(\boldsymbol{X_u}, \boldsymbol{X_u}) + \sigma_{n_u}^2 \boldsymbol{I} & k_{uf}(\boldsymbol{X_u}, \boldsymbol{X_f}) \\ k_{fu}(\boldsymbol{X_f}, \boldsymbol{X_u}) & k_{ff}(\boldsymbol{X_f}, \boldsymbol{X_f}) + \sigma_{n_f}^2 \boldsymbol{I} \end{bmatrix}\right) \tag{19}$$

We can directly use $\boldsymbol{K}$ to compute the mag log likelihood and train the model to obtain the optimized hyperparameters $\boldsymbol{\theta_*}$ and $\boldsymbol{\phi_*}$. More information about the optimization process will be provided in the next chapter. Here we now continue with the optimized hyperparameters.

3.1 Optimization/Finding of the hyperparameters

The optimization of the hyperparameters deemed to be harder than expected. Different optimizers were used.

3.2 Predictions with the Posterior

With the obtained hyperparameters $\boldsymbol{\theta_*}$ and $\boldsymbol{\phi_*}$ we can make new predictions at a test location $\boldsymbol{x_*}$. Similar to the construction in chapter 2.2 we can now formulate the predictive posterior distribution for $u(\boldsymbol{x_*})$ and also for $f(\boldsymbol{x_*})$ as done in [5]:

$$\begin{aligned} p(u(\boldsymbol{x_*})|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta_*}, \boldsymbol{\phi_*}) &= \mathcal{N}(\mu_{u*}, \Sigma_u) \\ p(u(\boldsymbol{x_*})|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta_*}, \boldsymbol{\phi_*}) &= \mathcal{N}(\mu_{f*}, \Sigma_f) \end{aligned} \tag{20}$$

with

$$\begin{aligned} \mu_u = \boldsymbol{\kappa_u} \boldsymbol{K}^{-1} \boldsymbol{y},\ \Sigma_* = k_{uu}(\boldsymbol{x_*}, \boldsymbol{x_*}) - \boldsymbol{\kappa_u} \boldsymbol{K}^{-1} \boldsymbol{\kappa}_{\boldsymbol{u}}^{\intercal} \\ \mu_f = \boldsymbol{\kappa_f} \boldsymbol{K}^{-1} \boldsymbol{y},\ \Sigma_* = k_{ff}(\boldsymbol{x_*}, \boldsymbol{x_*}) - \boldsymbol{\kappa_f} \boldsymbol{K}^{-1} \boldsymbol{\kappa}_{\boldsymbol{f}}^{\intercal} \end{aligned} \tag{21}$$

with $\boldsymbol{\kappa_u}$ and $\boldsymbol{\kappa_f}$ being:

$$\begin{aligned} \boldsymbol{\kappa_u} &= [k_{uu}(\boldsymbol{x_*}, \boldsymbol{X_u}) k_{uf}(\boldsymbol{x_*}, \boldsymbol{X_f})] \\ \boldsymbol{\kappa_f} &= [k_{fu}(\boldsymbol{x_*}, \boldsymbol{X_u}) k_{ff}(\boldsymbol{x_*}, \boldsymbol{X_f})] \end{aligned} \tag{22}$$

4 Results

For testing the framework we will use ODEs and PDEs of physical models. The numerically obtained solutions will be used as the ground truth. For obtaining $u(\boldsymbol{x})$, we will fully define the differential equation with its initial- and boundary conditions and then use mathematicas NDSolve (numerical solution) or NSolve (analytical solution) to compute the solution $\{(\boldsymbol{x}); u(\boldsymbol{x}), f(\boldsymbol{x})\}$. The training and validation points will be sampled from the solution using a sobol sequence[cite]. In addition, we will add some noise to the samples for the training and validation points, to analyse the impact onto the results and to simulate real world data. This will be done by creating two data sets $\{(\boldsymbol{x_u}); \boldsymbol{y_u}\}$ and $\{(\boldsymbol{x_f}); \boldsymbol{y_f}\}$ with $\boldsymbol{y_i} = u(\boldsymbol{x_i}) + \sigma_{ui}^2;\ i = u, f$. The implementation of a sobol sequence helps with evenly spaced training and validation points, which is important for the training process of the model.
Through the optimization of the mll, we will obtain the parameters of the underlying equation ϕ. With these parameters we will then calculate the predictive mean and covariance and then compare the ground truth with the computed mean function. To benchmark the performance a classical GP regression with a general SE kernel, as described in 2.2, will be computed and compared to the PI model.

4.1 Damped oscillator

As a first example we will look at a one dimensional differential equation of a damped oscillator. The differential equation is given by:

$$\frac{\partial^2 u(t)}{\partial t^2} + \gamma \frac{\partial u(t)}{\partial t} + \omega_0^2 u(t) = 0 \tag{23}$$

4.2 Heat Equation

The one dimensional time dependent heat equation with a forcing term is:

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} - \alpha \frac{\partial^2 u(x,t)}{\partial x^2} &= f(x,t) \\ u[0,x] &= \\ u(t,0) =; u(t,L) &= \end{aligned} \tag{24}$$

with α as the thermal diffusivity.

4.3 Wave Equation

$$\begin{aligned} \frac{1}{c^2} \frac{\partial^2 u(x,t)}{\partial t^2} - \frac{\partial^2 u(x,t)}{\partial x^2} &= f(x,t) \\ u[0,x] &= \\ u(t,0) =; u(t,L) &= \end{aligned} \tag{25}$$

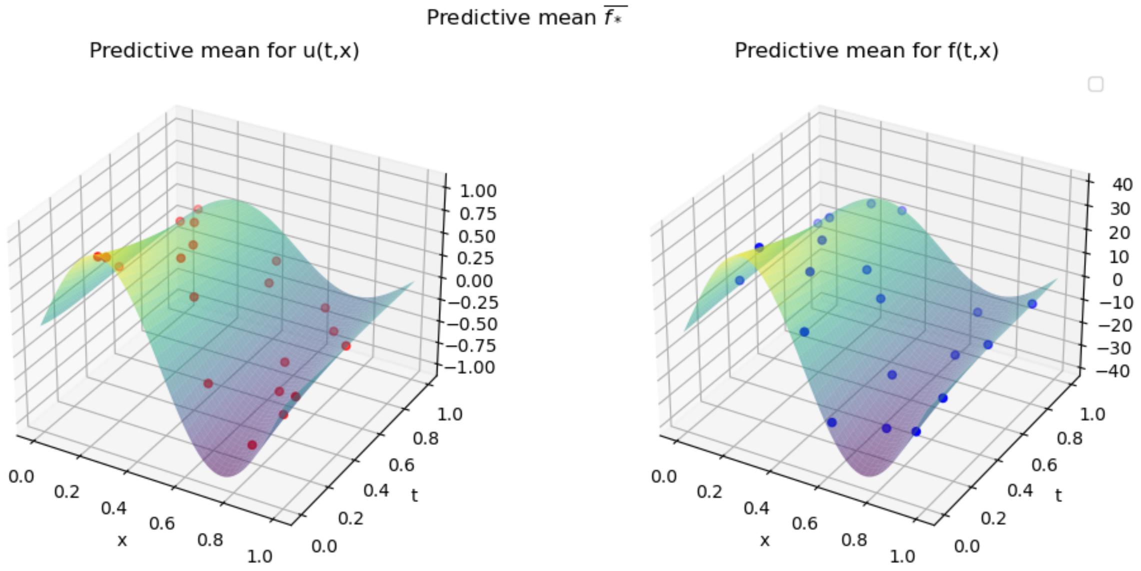

Figure 1

4.4 Damped Oscillator

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. 2006.

[2] D. Duvenaud. Automatic model construction with gaussian processes. 2014.

[3] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[4] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. A visual exploration of gaussian processes. *Distill*, 2019. https://distill.pub/2019/visual-exploration-gaussian-processes.

[5] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.

[6] Carl Edward Rasmussen. *Gaussian processes for machine learning*. Adaptive computation and machine learning. 2006.

[7] Simo Särkkä. Linear operators and stochastic partial differential equations in gaussian process regression. pages 151–158, 06 2011.