Solving Noisy Linear Operator Equations by Gaussian Processes: Application to Ordinary and Partial Differential Equations

Thore Graepel THOREG@MICROSOFT.COM
Microsoft Research, Cambridge
Roger Needham Building
7 J J Thomson Avenue
Cambridge, CB3 0FB
United Kingdom

Abstract

We formulate the problem of solving stochastic linear operator equations in a Bayesian Gaussian process (GP) framework. The solution is obtained in the spirit of a collocation method based on noisy evaluations of the target function at randomly drawn or deliberately chosen points. Prior knowledge about the solution is encoded by the covariance kernel of the GP. As in GP regression, analytical expressions for the mean and variance of the estimated target function are obtained, from which the solution of the operator equation follows by a manipulation of the kernel. Linear initial and boundary value constraints can be enforced by embedding the non-parametric model in a form that automatically satisfies the boundary conditions. The method is illustrated on a noisy linear first-order ordinary differential equation with initial condition and on a noisy second-order partial differential equation with Dirichlet boundary conditions.

1. Introduction

Gaussian processes (GP) have become popular tools for regression (MacKay, 1998) and—more recently—for classification (Williams & Barber, 1998) tasks. Interesting application include the prediction of wind fields (Nabney et al., 1999) and—more in the spirit of this work—inverse quantum theory (Lemm & Uhlig, 1999). GPs have been proposed as an alternative to neural networks, in particular within the Bayesian community, because of their beautiful simplicity, probabilistic interpretability, and analytical tractability—at least in the first stage of Bayesian inference (Gibbs, 1997). Inspired by recent work on solving differential equations by neural networks (Lagaris et al., 1998; van Milligen et al., 1995) on the one hand and by Bayesian methods (Skilling, 1992) on the other, we demonstrate in this paper that GPs are not limited to regression and classification, but can also be applied to the more general problem of solving noisy linear operator equations. This general problem had been discussed earlier in the context of smoothing splines by (Wahba, 1990).

The most widely used operator equations are integral and differential equations. Both types have an extremely wide scope of applications ranging from basic science to engineering. Our method provides a means to solving linear operator equations in stochastic settings where the given data are assumed to be noisy measurements. As an example of a noisy differential equation consider the Poisson equation (cf. Subsection 4.2) for the electric potential given noisy measurements of the charge distribution. Another application of derivative information in Gaussian processes has recently been presented in the context of dynamic systems (Solak et al., 2003). A typical application that requires the inversion of a stochastic integral equation is the deconvolution of a noisy image given the point-spread function of an optical instrument. In this work we focus on noisy differential equations.

The Gaussian process approach we advocate provides solutions in closed analytic form whose degree of differentiability or integrability is determined entirely by the choice of covariance kernel, which may also be used to incorporate prior knowledge into the model. The prior knowledge could include information about expected smoothness or periodicity of the solution. The noise in the data is modelled explicitly and the procedure leads to a distribution over solutions characterised by mean and variance at any given point. Hence, problems of generalisation such as over-fitting, which may result from the noise in the data, can be handled in a principled way. The proposed method is general and can be applied to ODEs and PDEs defined on orthog-

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

onal box boundaries.

Let us consider the solution of operator equations of the form, $\forall x \in \mathcal{X}$

$$\mathcal{A}\psi(x) = y(x) \;, \tag{1}$$

by an unknown function $\psi \in \mathcal{F}$ from a specified class of functions $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, given a function $y : \mathcal{X} \to \mathcal{Y}$ and an operator $\mathcal{A} : \mathcal{F} \to \mathcal{F}$. The existence of a unique solution often requires to impose n initial or boundary conditions on ψ that are conveniently expressed by n functional equations, $\mathcal{B}_i[\psi] := c_i$, with functionals $\mathcal{B}_i : \mathcal{F} \to \mathbb{R}$ and constants $c_i \in \mathbb{R}$. The quality of any given solution $\hat{\psi}(\mathbf{x})$ is determined by defining an integrable loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ and by evaluating the residual integral

$$R[\hat{\psi}] := \int_{\mathcal{X}} l\left(\mathcal{A}\hat{\psi}(x), y(x)\right) dx \,. \tag{2}$$

The problem can also be stated in a more general way by defining a probability measure P_{X} over $\mathcal{X}$ and by replacing the integral in (2) by an expectation. We consider the case when the deterministic function $y(x)$ is replaced by a random variable $\mathsf{Y}_{|\mathsf{X}=x}$ leading to stochastic operator equations (Vapnik, 1998).

Possibly the most straight-forward method to solving operator equations like (1) is by means of so-called collocation methods (Großmann & Roos, 1994): Choose a sample $\boldsymbol{x} := (x_1 \ldots, x_m)$ of elements $x \in \mathcal{X}$ and a class $\hat{\mathcal{F}}$ of functions $\hat{\psi}$. The residual integral $R[\hat{\psi}]$ is approximated by

$$\hat{R}_{\boldsymbol{x}}[\hat{\psi}] := \frac{1}{m} \sum_{i=1}^{m} l\left(\mathcal{A}\hat{\psi}(x_i), y(x_i)\right) \;, \tag{3}$$

and collocation methods aim at minimizing $\hat{R}_{\boldsymbol{x}}[\hat{\psi}]$ with respect to the choice of the function $\hat{\psi} \in \hat{\mathcal{F}}$. In order to make the optimisation tractable, $\hat{\psi}$ is often taken from a parameterised family of function approximators, $\phi_{\mathbf{w}} : \mathcal{X} \times \mathcal{W} \to \mathcal{Y}$. Since the resulting optimization problem is often ill-posed its solution usually requires regularization for stability (Vapnik, 1998).

As an example, the above formulation includes regression at points $\boldsymbol{x}$ with $\mathcal{X} := \mathcal{Y} := \mathbb{R}$ and $l(y, \hat{y}) := (y - \hat{y})^2$ as a special case, where $\mathcal{A}$ is the identity operator, $\mathcal{A} = \mathcal{I}$. Choosing $\mathcal{A}\psi := \operatorname{sign}(\psi)$, that is, a nonlinear operator, leads to the case of binary classification when applied in conjunction with the zero-one loss.

In this paper we focus on the case of linear operators $\mathcal{L}$ (e.g., differential or integral operators), that is, operators that satisfy for all $a_1, a_2 \in \mathbb{R}$ and all $\psi_1, \psi_2 \in \mathcal{F}$ that $\mathcal{L}(a_1\psi_1 + a_2\psi_2) = a_1\mathcal{L}\psi_1 + a_2\mathcal{L}\psi_2$. In Section 2 we show how GPs can be used to approximately solve linear operator equations with certain boundary conditions. In Section 3 we discuss linear differential operators and show how to solve linear differential equations given typical boundary conditions. In Section 4 we illustrate the approach by solving a noisy first-order linear ordinary differential equation (ODE) with initial condition (IC) and a noisy second-order partial differential equation (PDE) with Dirichlet boundary conditions (BCs).

2. Inversion of Linear Operators by Gaussian Processes

2.1. Probabilistic Model

We assume that the measurements of the right-hand side of the operator equation (1) are contaminated by Gaussian noise, that is, we have observations distributed according to

$$\mathsf{T}_{|\mathsf{X}=x} \sim \mathcal{N}\left(\mathcal{L}_x\psi(x), \sigma_t^2\right) \;.$$

As a consequence in our model the conditional distribution of m observations $\boldsymbol{t} := (t_1, \ldots, t_m)$ at the sample points $\boldsymbol{x}$ is given by

$$\mathsf{T}_{|\mathbf{X}=\boldsymbol{x},\mathbf{W}=\mathbf{w}} \sim \mathcal{N}\left(\mathcal{L}_x\phi_{\mathbf{W}}(\boldsymbol{x}), \sigma_t^2\mathbf{I}_m\right) \;,$$

where $\phi_{\mathbf{w}}$ is a function approximator parameterised by $\mathbf{w}$ and

$$\boldsymbol{\phi} := \phi_{\mathbf{w}}(\boldsymbol{x}) := \left(\phi_{\mathbf{w}}(x_1), \phi_{\mathbf{w}}(x_2), \ldots, \phi_{\mathbf{w}}(x_m)\right)^T \;.$$

We assume that $\phi_{\mathbf{w}}(\mathbf{x})$ can be written as a linear combination of basis functions $\phi_j : \mathcal{X} \to \mathcal{Y}$,

$$\phi_{\mathbf{w}}(x) := \sum_{j=1}^{n} w_j\phi_j(x) = \boldsymbol{\phi}^T\mathbf{w} \,.$$

We assume a Gaussian prior over weight vectors, $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, such that the observations are distributed according to

$$\mathsf{T}_{|\mathbf{X}=\boldsymbol{x},\mathbf{W}=\mathbf{w}} \sim \mathcal{N}\left(\mathcal{L}\boldsymbol{\Phi}\mathbf{w}, \sigma_t^2\mathbf{I}_m\right) \;,$$

where with matrix notation $\boldsymbol{\Phi}_{ij} := \phi_j(x_i)$ and $(\mathcal{L}\boldsymbol{\Phi})_{ij} := [\mathcal{L}_x\phi_j(x)]_{x=x_i}$. The posterior density over $\mathbf{W}$ is obtained by Bayes' theorem,

$$\mathsf{f}_{\mathbf{W}|\mathbf{X}=\boldsymbol{x},\mathsf{T}=\boldsymbol{t}}(\mathbf{w}) = \frac{\mathsf{f}_{\mathsf{T}|\mathbf{X}=\boldsymbol{x},\mathbf{W}=\mathbf{w}}(\boldsymbol{t})\,\mathsf{f}_{\mathbf{W}}(\mathbf{w})}{\mathsf{E}_{\mathbf{W}}\left[\mathsf{f}_{\mathsf{T}|\mathbf{X}=\boldsymbol{x},\mathbf{W}=\mathbf{w}}(\boldsymbol{t})\right]} \,,$$

leading to a Gaussian posterior over weight vectors,

$$\mathbf{W}_{|\mathbf{X}=\boldsymbol{x},\mathsf{T}=\boldsymbol{t}} \sim \mathcal{N}\left(\boldsymbol{\mu}\left(\mathcal{L}\boldsymbol{\Phi}, \boldsymbol{t}, \sigma_t^2\right), \boldsymbol{\Sigma}\left(\mathcal{L}\boldsymbol{\Phi}, \boldsymbol{t}, \sigma_t^2\right)\right) \;,$$

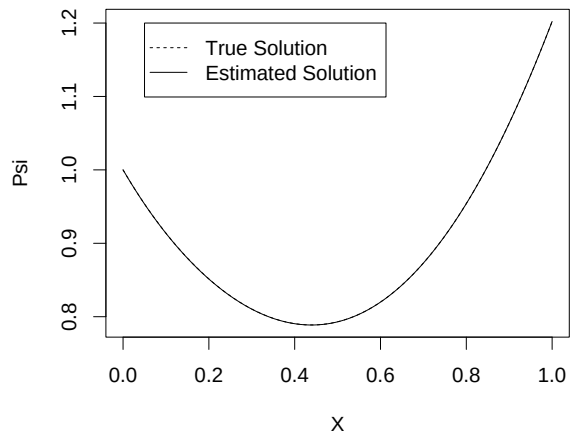

Figure 1. Noise-free ($\sigma_y = 0$) first-order ODE (11) with IC $\psi(0) = 1$ solved at ten points equally spaced in the interval $[0, 1]$ with a Gaussian kernel (10) of $\sigma = 0.3$, and assumed noise level $\sigma_t = \sigma_y = 0$. Shown is the analytical solution $\psi(x)$ and and the (coinciding) estimated solution $\hat{\psi}(x)$.

with expectation

$$\boldsymbol{\mu}\left(\mathcal{L}\boldsymbol{\Phi}, \boldsymbol{t}, \sigma_t^2\right) := \left(\left(\mathcal{L}\boldsymbol{\Phi}\right)^T \mathcal{L}\boldsymbol{\Phi} + \sigma_t^2 \mathbf{I}_n\right)^{-1} \left(\mathcal{L}\boldsymbol{\Phi}\right)^T \boldsymbol{t}\,, \tag{4}$$

and covariance matrix

$$\boldsymbol{\Sigma}\left(\mathcal{L}\boldsymbol{\Phi}, \boldsymbol{t}, \sigma_t^2\right) := \left(\sigma_t^{-2}\left(\mathcal{L}\boldsymbol{\Phi}\right)^T \mathcal{L}\boldsymbol{\Phi} + \mathbf{I}_n\right)^{-1}. \tag{5}$$

Using the Woodbury formula and the fact that the expected squared loss is minimised by the posterior expectation we obtain the prediction $\hat{y}(x) := \mathcal{L}_x \hat{\psi}(x)$ at point x,

$$\begin{aligned} \hat{y}(x) &:= \mathsf{E}_{\mathbf{W}|\mathbf{X}=\boldsymbol{x},\mathbf{T}=\boldsymbol{t}}\left[\mathcal{L}_x \phi_{\mathbf{W}}(x)\right] \qquad (6)\\ &= \boldsymbol{\phi}_{\mathcal{L}}^T(x)\, \boldsymbol{\Phi}_{\mathcal{L}}^T \left(\boldsymbol{\Phi}_{\mathcal{L}} \boldsymbol{\Phi}_{\mathcal{L}}^T + \sigma_t^2 \mathbf{I}_m\right)^{-1} \boldsymbol{t}\,, \end{aligned}$$

where we define $\boldsymbol{\phi}_{\mathcal{L}}(x) := \mathcal{L}\boldsymbol{\phi}(x)$ and $\boldsymbol{\Phi}_{\mathcal{L}} := \mathcal{L}\boldsymbol{\Phi}$. The variance estimate is given by

$$\begin{aligned} \hat{\sigma}_y^2(x) &= \mathsf{Var}\left[\mathcal{L}_x \psi(x)\right] = \mathsf{Var}_{\mathbf{W}|\mathbf{X}=\boldsymbol{x},\mathbf{T}=\boldsymbol{t}}\left[\mathcal{L}_x \phi_{\mathbf{W}}(x)\right]\\ &= \boldsymbol{\phi}_{\mathcal{L}}^T \boldsymbol{\phi}_{\mathcal{L}} - \boldsymbol{\phi}_{\mathcal{L}}^T \boldsymbol{\Phi}_{\mathcal{L}}^T \left(\boldsymbol{\Phi}_{\mathcal{L}} \boldsymbol{\Phi}_{\mathcal{L}}^T + \sigma_t^2 \mathbf{I}_m\right)^{-1} \boldsymbol{\Phi}_{\mathcal{L}} \boldsymbol{\phi}_{\mathcal{L}}\,. \end{aligned}$$

2.2. Operators and Covariance Kernels

Let us define the covariance kernel acted upon by the linear operator $\mathcal{L}$ on both arguments, $k_{\mathcal{L}^2} : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$, by

$$k_{\mathcal{L}^2}(x, \tilde{x}) := \sum_{i=1}^{n} \mathcal{L}_x \phi_i(x)\, \mathcal{L}_{\tilde{x}} \phi_i(\tilde{x}) = \mathcal{L}_x \mathcal{L}_{\tilde{x}} k(x, \tilde{x})\,,$$

and the *inverted* operator kernel, $k_{\mathcal{L}} : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$, by

$$\begin{aligned} k_{\mathcal{L}}(x, \tilde{x}) &:= \mathcal{L}_x^{-1} k_{\mathcal{L}^2}(x, \tilde{x})\\ &= \sum_{i=1}^{n} \phi_i(x)\, \mathcal{L}_{\tilde{x}} \phi_i(\tilde{x}) = \mathcal{L}_{\tilde{x}} k(x, \tilde{x})\,, \end{aligned}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ is the covariance kernel chosen beforehand. By construction $k_{\mathcal{L}^2}$ is positive semi-definite as well. If we define the matrix $\mathbf{K}_{\mathcal{L}^2}$ by $\left(\mathbf{K}_{\mathcal{L}^2}\right)_{ij} := k_{\mathcal{L}^2}(x_i, x_j)$ we can write the predicted left-hand side $\hat{y}(x) = \mathcal{L}_x \hat{\psi}(x)$ of the operator equation from (6) in the form of the well-known kernel expansion,

$$\begin{aligned} \hat{y}(x) &= \mathsf{E}_{\mathbf{W}|\mathbf{X}=\boldsymbol{x},\mathbf{T}=\boldsymbol{t}}\left[\mathcal{L}_x \phi_{\mathbf{W}}(x)\right]\\ &= \sum_{j=1}^{m} \hat{\alpha}_j k_{\mathcal{L}^2}(x, x_j)\\ \hat{\boldsymbol{\alpha}} &:= \left(\mathbf{K}_{\mathcal{L}^2} + \sigma_t^2 \mathbf{I}_m\right)^{-1} \boldsymbol{t}\,. \end{aligned}$$

Defining $\left(\mathbf{k}_{\mathcal{L}^2}\right)_i := k_{\mathcal{L}^2}(x, x_i)$ the estimated variance is given by

$$\begin{aligned} \hat{\sigma}_y^2(x) &:= \mathsf{Var}_{\mathbf{W}|\mathbf{X}=\boldsymbol{x},\mathbf{T}=\boldsymbol{t}}\left[\mathcal{L}_x \phi_{\mathbf{W}}(x)\right]\\ &= k_{\mathcal{L}^2}(x, x) - \mathbf{k}_{\mathcal{L}^2}^T \left(\mathbf{K}_{\mathcal{L}^2} + \sigma_t^2 \mathbf{I}_m\right)^{-1} \mathbf{k}_{\mathcal{L}^2} \qquad (7) \end{aligned}$$

Using the relation $\mathcal{L}_x \hat{\psi}(x) = \hat{y}(x)$ and the fact that by construction $\mathcal{L}_x k_{\mathcal{L}}(x, \tilde{x}) = k_{\mathcal{L}^2}(x, \tilde{x})$ and we obtain an estimator $\hat{\psi}(x)$ for the solution ψ of the stochastic linear operator equation,

$$\hat{\psi}(x) = \sum_{j=1}^{m} \hat{\alpha}_j k_{\mathcal{L}}(x, x_j)\,.$$

We can also calculate the variance from (7),

$$\hat{\sigma}_{\psi}^2(x) = k(x, x) - \mathbf{k}_{\mathcal{L}}^T \left(\mathbf{K}_{\mathcal{L}^2} + \sigma_t^2 \mathbf{I}_m\right)^{-1} \mathbf{k}_{\mathcal{L}}\,.$$

Examples of both $\hat{\psi}(x)$ and $\hat{\sigma}_{\psi}^2(x)$ are shown in Figure 3 for a simple ODE with initial condition.

2.3. Initial and Boundary Conditions

We must ensure that given BCs are satisfied by the solutions. We restrict ourselves to linear ICs/BCs, that is, we require for all $a_1, a_2 \in \mathbb{R}$ and for all $\psi_1, \psi_2 \in \mathcal{F}$ that

$$\mathcal{B}_i\left[a_1 \psi_1 + a_2 \psi_2\right] = a_1 \mathcal{B}_i\left[\psi_1\right] + a_2 \mathcal{B}_i\left[\psi_2\right]\,.$$

Following (Lagaris et al., 1998) we choose candidate solutions $\theta_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$ that are related to the function approximator $\phi_{\mathbf{w}}$ by

$$\theta_{\mathbf{w}}(x) := b(x) + \mathcal{C}_x \phi_{\mathbf{w}}(x)\,,$$

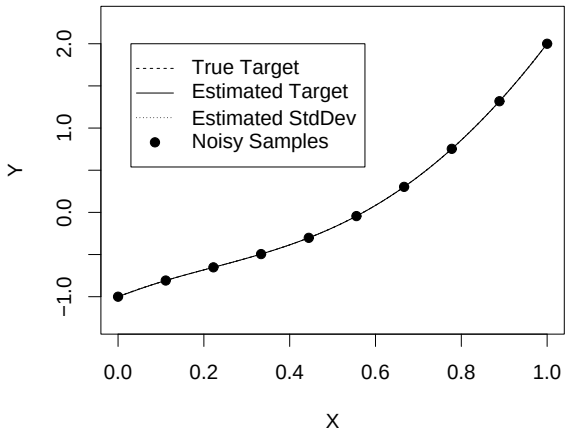

Figure 2. Scenario as in Figure 1. Shown is the given target function $y(x)$ and the (coinciding) estimated target function $\hat{y}(x)$, and sample points (x_i, t_i)

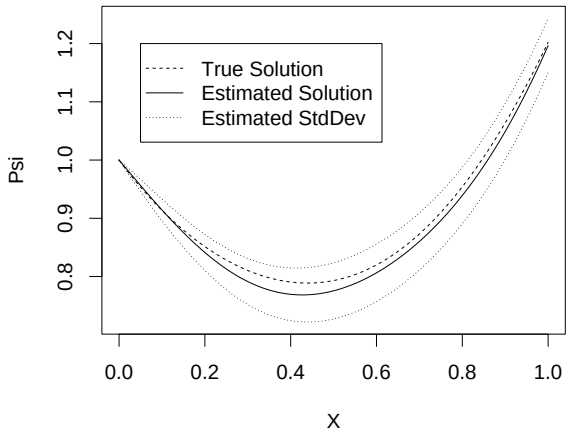

Figure 3. Noisy ($\sigma_y = 0.3$) ODE (11) with IC $\psi(0) = 1$ solved at ten points equally spaced in the interval $[0,1]$ with a Gaussian kernel (10) of $\sigma = 0.5$ and assumed noise level $\sigma_t = \sigma_y = 0.3$. Shown is the analytical solution $\psi(x)$ and the estimated solution $\hat{\psi}(x) \pm \hat{\sigma}_\psi(x)$.

with a function $b : \mathcal{X} \to \mathcal{Y}$ that satisfies the BCs/ICs, $\mathcal{B}_i[b] = c_i$, and another linear operator $\mathcal{C}_x : \mathcal{F} \to \mathcal{F}$ such that $\mathcal{B}_i[\mathcal{C}_x \phi_{\mathbf{w}}] = 0$. The corresponding unconstrained operator equation (1) becomes

$$\mathcal{L}_x \mathcal{C}_x \phi_{\mathbf{w}}(x) \quad = \quad y(x) - \mathcal{L}_x b(x) \ ,$$

and the effective kernel function is given by

$$k_{\mathcal{L}^2}(x, \tilde{x}) := \mathcal{L}_x \mathcal{C}_x \mathcal{L}_{\tilde{x}} \mathcal{C}_{\tilde{x}} k(x, \tilde{x}) \ .$$

How b and $\mathcal{C}_x$ are chosen will be discussed for specific boundary conditions in Section 3.

3. Differential Equations and Boundary Conditions

We now focus on the case of differential equations involving a linear differential operator $\mathcal{D}_x$. For the case of ODEs we take $\mathcal{X} \subset \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$. The general linear differential operator of order $N \geq 0$ then reads

$$\mathcal{D}_x := \frac{d^N}{dx^N} + \sum_{i=0}^{N-1} f_i(x) \frac{d^i}{dx^i} ,$$

with $f_i : \mathcal{X} \to \mathcal{Y}$ arbitrary functions and $\frac{d^0}{dx^0} := 1$.

In order to obtain unique solutions we need additional constraints in the form of ICs/BCs. Let us consider an Nth order linear differential equation with $n = N$ *initial conditions (ICs)* $\psi^{(i-1)}(0) = c_i$ on the derivatives of ψ. We can write the candidate solution in the form $\theta_{\mathbf{w}}(x) := b(x) + \mathcal{C}_x \phi_{\mathbf{w}}(x)$ as

$$\theta_{\mathbf{w}}(x) := \sum_{i=1}^{N} x^{i-1} c_i + x^N \phi_{\mathbf{w}}(x) \ , \qquad (8)$$

which is constructed so as to satisfy the ICs defined above. For a first order linear ODE with IC $\psi(0) = c$ this corresponds to the candidate solution

$$\theta_{\mathbf{w}}(x) := c + x \phi_{\mathbf{w}}(x) \ .$$

Let us discuss the case of PDEs, with $\mathcal{X} = [0,1] \times [0,1] \subset \mathbb{R}^2$ for illustration. Typical *boundary conditions (BCs)* are of the Dirichlet type: $\psi(0, x_2) = c_{10}(x_2)$, $\psi(1, x_2) = c_{11}(x_2)$, $\psi(x_1, 0) = c_{20}(x_1)$, and $\psi(x_1, 1) = c_{21}(x_1)$ with $c_{ij} : [0,1] \to \mathbb{R}$. The candidate solution is written as

$$\theta_{\mathbf{w}}(\mathbf{x}) := b(\mathbf{x}) + x_1(1 - x_1) x_2(1 - x_2) \phi_{\mathbf{w}}(\mathbf{x}) \ , \quad (9)$$

where $b(\mathbf{x})$ is given by

$$b(\mathbf{x}) := b_1(\mathbf{x}) + b_2(\mathbf{x}) \ ,$$

with

$$b_1(\mathbf{x}) := (1 - x_1) c_{10}(x_2) + x_1 c_{11}(x_2) \ ,$$

and

$$\begin{aligned} & b_2(\mathbf{x}) := \\ & (1 - x_2)(c_{20}(x_1) - ((1 - x_1) c_{20}(0) + x_1 c_{20}(1))) \\ & + x_2(c_{21}(x_1) - ((1 - x_1) c_{21}(0) + x_1 c_{21}(1))) \ . \end{aligned}$$

Note that von Neumann BCs could be treated in a similar way.

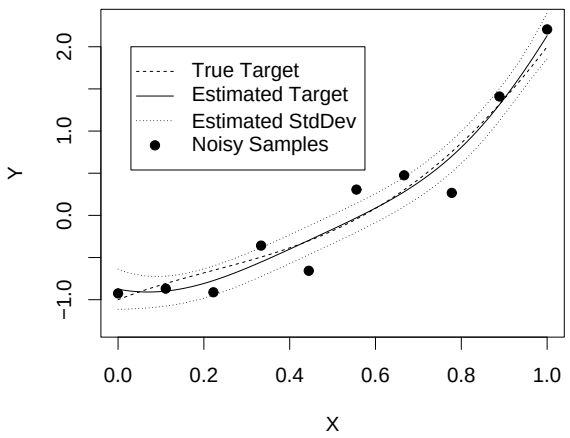

Figure 4. Scenario of Figure 3. Shown are the true target function $y(x)$, sample points (x_i, t_i), and estimated target function $\hat{y}(x) \pm \hat{\sigma}_y(x)$.

4. Experimental Results

We illustrate our approach by solving two differential equations with given ICs/BCs, whose analytical solutions for the noise-free case are known (Lagaris et al., 1998). For our experiments we used the Gaussian covariance kernel (see Appendix 6 for derivatives)

$$k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) := \exp\left(-\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|^2}{2\sigma^2}\right) . \qquad (10)$$

Other, problem-specific choices are desirable in practice. In order to make the examples easy to reproduce we provide some detailed results about the evaluation of differential operators applied to kernels.

4.1. First-Order Linear Ordinary Differential Equation

Let us consider the first ODE toy problem from (Lagaris et al., 1998)

$$\frac{d\psi(x)}{dx} + f(x)\,\psi(x) = g(x) , \qquad (11)$$

with

$$f(x) := \left(x + \frac{1+3x^2}{1+x+x^3}\right) ,$$

and

$$g(x) := x^3 + 2x + x^2\frac{1+3x^2}{1+x+x^3} .$$

The initial conditions are $\psi(0) = 1$ and $\mathcal{X} = [0, 1]$. The analytic solution is

$$\psi(x) = \frac{\exp\left(-x^2/2\right)}{(1+x+x^3)+x^2} .$$

According to (8) the candidate solution of this first-order ODE is written as

$$\theta_{\mathbf{w}}(x) := b(x) + \mathcal{C}_x\phi_{\mathbf{w}}(x) = 1 + x\phi_{\mathbf{w}}(x) .$$

The complete operator $\mathcal{L}_x$ acting upon $\phi_{\mathbf{w}}$ is

$$\mathcal{L}_x = \mathcal{D}_x\mathcal{C}_x := \left(\frac{d}{dx} + f(x)\right) x ,$$

and the target function $y(x)$ of the operator equation becomes

$$y(x) := g(x) - \mathcal{D}_x 1 = g(x) - f(x) .$$

Using the results in Appendix 6 and the shorthand notation $\partial_x := \frac{\partial}{\partial x}$ and $f_x := f(x)$ we obtain for the inverted operator kernel

$$\begin{aligned} k_{\mathcal{L}}(x, \tilde{x}) &= (1 + \tilde{x}\partial_{\tilde{x}} + f_{\tilde{x}}\tilde{x})\,k(x, \tilde{x}) \\ &= \left(1 + \tilde{x}\frac{\Delta x}{\sigma^2} + \tilde{x}f_{\tilde{x}}\right) k(x, \tilde{x}) . \end{aligned}$$

Writing $\Delta x := x_i - x_j$ and $\partial^2_{x\tilde{x}} = \partial_x\partial_{\tilde{x}}$ we obtain for $k_{\mathcal{L}^2}(x, \tilde{x})$:

$$\begin{aligned} &k_{\mathcal{L}^2}(x, \tilde{x}) = \\ &\Big[\, x\tilde{x}\left(\partial^2_{x\tilde{x}} + f_{\tilde{x}}\partial_x + f_x\partial_{\tilde{x}} + f_x f_{\tilde{x}}\right) \\ &+ \left(x\partial_x + \tilde{x}\partial_{\tilde{x}} + f_x x + f_{\tilde{x}}\tilde{x} + 1\right) \\ &+ x\tilde{x}\left(\frac{1}{\sigma^2} + f_x f_{\tilde{x}} + \left(f_x - f_{\tilde{x}} - \frac{\Delta x}{\sigma^2}\right)\frac{\Delta x}{\sigma^2}\right) \\ &+ \left(1 - \frac{\Delta x^2}{\sigma^2} + \tilde{x}f_{\tilde{x}} + x f_x\right)\Big]\, k_\sigma(x, \tilde{x}) . \end{aligned}$$

We generated a sample of 10 equally spaced points in $[0, 1]$ from $g(x)$ in a noise-free version and contaminated with Gaussian noise of variance $\sigma_y = 0.3$. We applied the method from Section 2 using the Gaussian RBF kernel (10) with $\sigma = 0.5$ and an assumed noise level σ_t matching the actual noise used, i.e., $\sigma_t = 0$ and $\sigma_t = 0.3$ respectively.

Figure 1 shows the true and the estimated solution for the ODE problem (11) without noise, resulting in an exact match. Figure 2 shows the corresponding estimated target function which essentially coincides with the true target function. The accuracy is very high and appears to correspond to that of the neural network methods (Lagaris et al., 1998), who do not provide exact figures. Their neural network, however, uses 30 free parameters (weights) as opposed to 10 expansion coefficients for the Gaussian process approach.

Figure 4 shows the result of the estimation with noise ($\sigma_y = 0.3$) in the target values: Although slightly off, the solution is rather robust w.r.t. the introduced

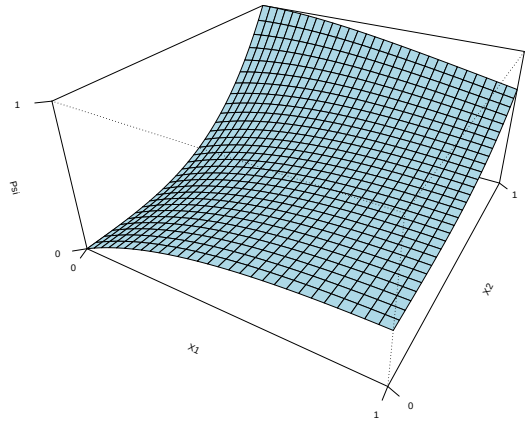

Figure 5. Analytical solution $\psi(\mathbf{x})$ of PDE (12) with Dirichlet BCs (13). The BCs can be seen as functions on the boundary of the plot.

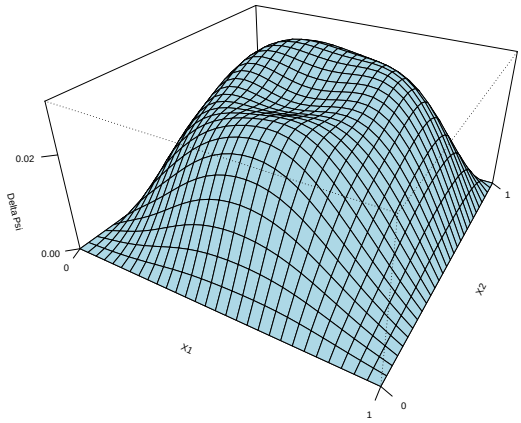

Figure 6. Noisy ($\sigma_y = 0.01$) PDE (12) with Dirichlet BCs solved at 100 points equally spaced in $[0,1] \times [0,1]$ with a Gaussian kernel (10) and $\sigma = 0.2$, assumed noise level $\sigma_t = \sigma_y = 0.01$. Shown is the deviation $\psi(\mathbf{x}) - \hat{\psi}(\mathbf{x})$.

noise. No comparison was made with the neural network methods (Lagaris et al., 1998) because they do not provide a way to deal with noise in the right-hand side of the differential equation.

4.2. Second-Order Linear Partial Differential Equation

Let us consider the following Poisson equation (Lagaris et al., 1998, Problem 1),

$$\nabla^2_{\mathbf{x}} \psi(\mathbf{x}) = \underbrace{\exp(-x_1)\left(x_1 - 2 + x_2^3 + 6x_2\right)}_{g(\mathbf{x})}, \qquad (12)$$

with Dirichlet BCs (see boundaries of Figure 5)

$$\begin{aligned}
\psi(0, x_2) &= x_2^3 =: c_{10}(x_2)\,, \\
\psi(1, x_2) &= \left(1 + x_2^3\right)\exp(-1) =: c_{11}(x_2)\,, \\
\psi(x_1, 0) &= x_1 \exp(-x_1) =: c_{20}(x_1)\,, \\
\psi(x_1, 1) &= (x_1 + 1)\exp(-x_1) =: c_{21}(x_1)\,, \qquad (13)
\end{aligned}$$

and $\mathcal{X} = [0,1] \times [0,1]$. The analytic solution is

$$\psi(\mathbf{x}) = \exp(-x_1)\left(x_1 + x_2^3\right)\,.$$

Poisson equations describe the spatial variation of a potential function for given source terms and have important applications in electrostatics and fluid dynamics. The candidate solution is given by (9) with functions c_{ij} as defined above. The complete operator $\mathcal{L}_{\mathbf{x}}$ acting upon $\phi_{\mathbf{w}}$ is

$$\mathcal{L}_{\mathbf{x}} = \mathcal{D}_{\mathbf{x}}\mathcal{C}_{\mathbf{x}} := \nabla^2_{\mathbf{x}} x_1 (1 - x_1)\, x_2 (1 - x_2)\,,$$

and the target function of the operator equation is

$$y(\mathbf{x}) := g(\mathbf{x}) - \nabla^2_{\mathbf{x}} b(\mathbf{x})\,,$$

with

$$\begin{aligned}
\nabla^2_{\mathbf{x}} b(\mathbf{x}) &= (1 - x_1)\, 6x_2 + x_1 6x_2 e^{-1} \\
&+ (1 - x_2)(x_1 - 2)\, e^{-x_1} + x_2 (x_1 - 1)\, e^{-x_1}\,,
\end{aligned}$$

as can be shown using the identity

$$\nabla^2 fg = f\nabla^2 g + g\nabla^2 f + 2\nabla f \cdot \nabla g\,.$$

Applying the rules of differential calculus we obtain for the inverted operator kernel

$$\begin{aligned}
k_{\mathcal{L}}(\mathbf{x}, \tilde{\mathbf{x}}) &= \mathcal{D}_{\tilde{\mathbf{x}}}\mathcal{C}_{\tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}}) \\
&= \Big[\Big(\tilde{H}_0 \nabla^2_{\tilde{\mathbf{x}}} + + \tilde{H}_3\Big) \\
&\quad + 2\left(\tilde{H}_1 \partial_{\tilde{x}_1} + \tilde{H}_2 \partial_{\tilde{x}_2}\right)\Big]\, k(\mathbf{x}, \tilde{\mathbf{x}})\,,
\end{aligned}$$

where we define $\tilde{H}_0 := h_{\tilde{x}_1} h_{\tilde{x}_2}$, $\tilde{H}_1 := h'_{\tilde{x}_1} h_{\tilde{x}_2}$, $\tilde{H}_2 := h_{\tilde{x}_1} h'_{\tilde{x}_2}$, and $\tilde{H}_3 := -2\left(h_{\tilde{x}_1} + h_{\tilde{x}_2}\right) + h'_{\tilde{x}_1} h'_{\tilde{x}_2}$ and $h_x := x(1 - x)$ and $h'_x := 1 - 2x$. For the double operator kernel $k_{\mathcal{L}^2}(\mathbf{x}, \tilde{\mathbf{x}})$ we obtain

$$\begin{aligned}
k_{\mathcal{L}^2}(\mathbf{x}, \tilde{\mathbf{x}}) = \Big[& H_0 \tilde{H}_0 \nabla^2_{\mathbf{x}} \nabla^2_{\tilde{\mathbf{x}}} + \\
& + 2H_0 \left(\tilde{H}_1 \nabla^2_{\mathbf{x}} \partial_{\tilde{x}_1} + \tilde{H}_2 \nabla^2_{\mathbf{x}} \partial_{\tilde{x}_2}\right) \\
& + 2\tilde{H}_0 \left(H_1 \partial_{x_1} + H_2 \partial_{x_2}\right) \nabla^2_{\tilde{\mathbf{x}}} \\
& + 4H_1 \left(\tilde{H}_1 \partial_{x_1} \partial_{\tilde{x}_1} + \tilde{H}_2 \partial_{x_1} \partial_{\tilde{x}_2}\right) \\
& + 4H_2 \left(\tilde{H}_1 \partial_{x_2} \partial_{\tilde{x}_1} + \tilde{H}_2 \partial_{x_2} \partial_{\tilde{x}_2}\right) \\
& + H_3 \left(\tilde{H}_0 \nabla^2_{\tilde{\mathbf{x}}} + 2\left(\tilde{H}_1 \partial_{\tilde{x}_1} + \tilde{H}_2 \partial_{\tilde{x}_2}\right)\right) \\
& + \tilde{H}_3 \left(H_0 \nabla^2_{\mathbf{x}} + 2\left(H_1 \partial_{x_1} + H_2 \partial_{x_2}\right)\right)\Big]\, k(\mathbf{x}, \tilde{\mathbf{x}})\,.
\end{aligned}$$

Both expressions can be further evaluated given the results of Appendix 6 for a Gaussian covariance kernel, or for any other component-wise twice differentiable kernel.

We generated a noisy sample ($\sigma_y = 0.01$) from the right-hand side $g(\mathbf{x})$ of the PDE (12) with Dirichlet BCs at 100 points equally spaced in $[0,1] \times [0,1]$. We solved the problem using a Gaussian kernel (10) and $\sigma = 0.2$ and an assumed noise level $\sigma_t = \sigma_y = 0.01$.

Figure 5 shows the analytical solution to the PDE problem (12) and Figure 6 shows the deviation of the estimated solution. Clearly, the strong boundary conditions enforce equality at the boundary, and the deviation is stronger at the centre of $[0,1] \times [0,1]$. Figure 7 shows the given target function $y(\mathbf{x})$ and Figure 8 the deviation of the estimated target function, which—in contrast to the estimated solution $\hat{\psi}$—is uniformly noisy over $[0,1] \times [0,1]$.

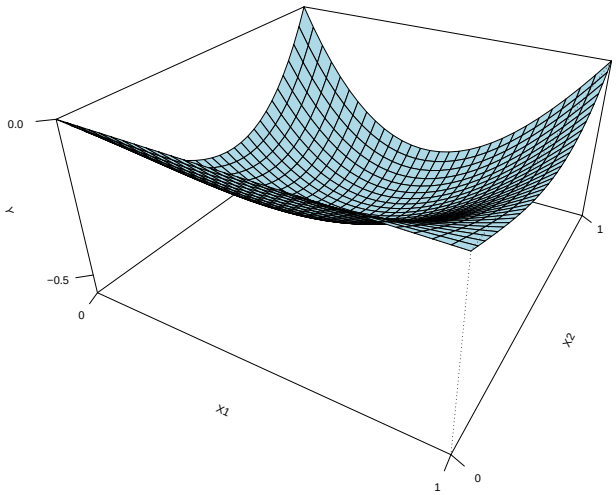

Figure 7. PDE (12) as in Figure 5. Shown is the target function $y(\mathbf{x})$.

Again, the results cannot be compared directly to other methods, because neither the neural network methods (Lagaris et al., 1998) nor standard methods like finite-elements provide a straight-forward way of dealing with noise. It could be argued that the squared-loss used corresponds to a maximum-likelihood method under a Gaussian noise assumption, but the neural network is really only used as a function approximator without any statistical interpretation. For the noise-less case, the neural network method is reported to give comparable results to the finite elements methods on the training data and much better results on the test data. In comparison, the deviations of the GP method on the noisy PDE example with $\sigma_t = 0.01$ are of the order of 0.02 as can be seen from Figure 6. This represents a respectable degree of stability and generalisation considering the ill-posed nature of operator inversion tasks that tend to make the solution critically dependent on fluctuation on the right-hand side of the equation.

5. Conclusion and Outlook

We proposed to use GPs to solve noisy differential equations with orthogonal box boundaries in particular. Our approach is related to collocation methods in that the operator inversion is reduced to an optimisation problem on a grid. However, our approach is probabilistic and allows for incorporating prior knowledge about the nature of the solution into the covariance kernel. Also, the noise in the data is taken into account explicitly and contributes to a distribution over solutions characterised by expectation and variance. The solution obtained is as differentiable as the covariance kernel used and is given as a linear expansion of the covariance kernel evaluated between grid points and test points. The computational costs of the algorithm are $\mathcal{O}(m^3)$ for the inversion of the matrix in (4). While these costs may exceed those of training a neural network, they can be reduced to $\mathcal{O}(m^2)$ or in the case of approximate inversion to $\mathcal{O}(m \log m)$ by working on a grid with a covariance kernel of finite support (Storkey, 1999).

The close relationship between GPs and other kernel methods suggests the development of related algorithms, e.g., by replacing the squared loss by the ε-insensitive loss (Smola & Schölkopf, 1998). Also the whole machinery of Bayesian inference (e.g., evidence maximisation) can be applied to operator inversion based on this work. Future work may aim at the treatment of more complex boundaries (Lagaris et al., 2000), on approximate GP models for the inversion of non-linear operators (similar to GP classification (Williams & Barber, 1998)), or the application of GP techniques to operator eigenvalue problems, e.g., in the context of quantum mechanics (Lagaris et al., 1997).

6. Appendix: Derivatives of Gaussian Covariance Kernel

We define $\boldsymbol{\Delta} := \mathbf{x} - \tilde{\mathbf{x}}$ and $\Delta^2 := \|\boldsymbol{\Delta}\|^2$ for shorter notation.

$$\nabla_{\tilde{\mathbf{x}}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) = -\nabla_{\mathbf{x}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\boldsymbol{\Delta}}{\sigma^2} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}})$$

$$\nabla_{\mathbf{x}} \nabla_{\tilde{\mathbf{x}}}^T k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{k_\sigma(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma^2} \left(\mathbf{I}_d - \frac{1}{\sigma^2} \boldsymbol{\Delta}\boldsymbol{\Delta}^T \right)$$

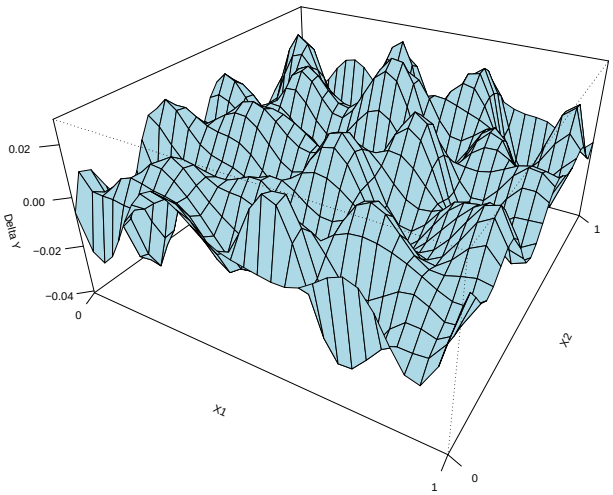

Figure 8. Noisy PDE (12) as in Figure 6. Shown is the deviation $y(\mathbf{x}) - \mathcal{L}_x \hat{\psi}(x)$ of estimated from given target function.

$$\begin{aligned} \nabla^2_{\mathbf{x}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) &= \nabla^2_{\tilde{\mathbf{x}}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) = \nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) \\ &= \frac{1}{\sigma^2}\left(\frac{\Delta^2}{\sigma^2} - d\right) k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

$$\begin{aligned} \nabla_{\mathbf{x}} \nabla^2_{\tilde{\mathbf{x}}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) &= -\nabla_{\tilde{\mathbf{x}}} \nabla^2_{\mathbf{x}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) \\ &= \frac{k_\sigma(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma^4}\left(d + 2 - \frac{\Delta^2}{\sigma^2}\right) \boldsymbol{\Delta} \end{aligned}$$

$$\nabla^2_{\mathbf{x}} \nabla^2_{\tilde{\mathbf{x}}} k_\sigma(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{k_\sigma(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma^4}\left(\left(d + 2 - \frac{\Delta^2}{\sigma^2}\right)^2 - 2d - 4\right)$$

Acknowledgements

I would like to thank Michele Milano and Petros Koumoutsakos for inspiring this work. I would also like to thank Manfred Opper, Matthias Seeger, Nici Schraudolph, and John Shawe-Taylor for valuable discussions. This project was started at the Institute for Computational Science at the Swiss Federal Institute of Technology, Zurich, and it was completed at Royal Holloway College, University of London, under EPSRC grant GR/R55948/01.

References

Gibbs, M. N. (1997). *Bayesian Gaussian methods for regression and classification.* Doctoral dissertation, University of Cambridge.

Großmann, C., & Roos, H.-G. (1994). *Numerik partieller differentialgleichungen.* Teubner Studienbücher. Stuttgart: Teubner.

Lagaris, I. E., Likas, A., & Fotiadis, D. I. (1997). Artificial neural network methods in quantum mechanics. *Computer Physics Communications, 104*, 1–14.

Lagaris, I. E., Likas, A., & Fotiadis, D. I. (1998). Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks, 9.*

Lagaris, I. E., Likas, A., & Papageougiou, D. G. (2000). Neural network methods for boundary value problems with irregular boundaries. *IEEE Transactions on Neural Networks, 11.*

Lemm, J. C., & Uhlig, C. (1999). Bayesian inverse quantum theory. *Few Body Systems, 29*, 25–52.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. In C. M. Bishop (Ed.), *Neural networks and machine learning*, 133–165. Berlin: Springer.

Nabney, I. T., Cornford, D., & Williams, C. K. I. (1999). Bayesian inference for wind field retrieval. *Neurocomputing Letters, 26–27*, 1013–1018.

Skilling, J. (1992). Bayesian solution of ordinary differential equations. In G. J. Erickson and C. R. Smith (Eds.), *Maximum entropy and bayesian methods.* Berlin: Kluwer.

Smola, A. J., & Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica, 22*, 211–231.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., & Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. *Advances in Neural Information Processing Systems 15.* Cambridge, MA: MIT Press.

Storkey, A. J. (1999). Truncated covariance matrices and Toeplitz methods in Gaussian processes. *Proceedings of the Ninth International Conference on Artificial Neural Networks, ICANN99* (pp. 55–60). Cambridge, MA.

van Milligen, B. P., Tribaldos, V., & Jimënes, J. A. (1995). Neural network differential equation and plasma equilibrium solver. *Physical Review Letters, 75*, 3594–3597.

Vapnik, V. (1998). *Statistical learning theory.* New York: John Wiley and Sons.

Wahba, G. (1990). *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Philadelphia: SIAM.

Williams, C. K. I., & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI, 20*, 1342–1351.