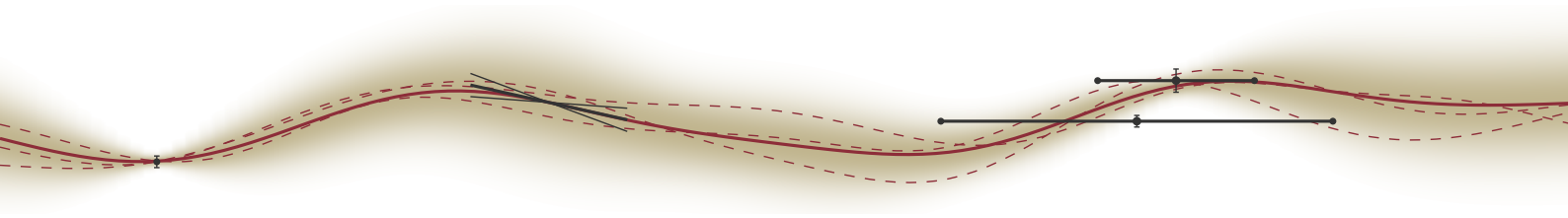
Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Gaussian Process Inference in Mechanistic Models based on Linear Partial Differential Equations

by

Marvin Pförtner



A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

in

Machine Learning

July 2022

First Examiner	Prof. Dr. Philipp Hennig
Second Examiner	Prof. Dr. Ingo Steinwart
Supervisor	Jonathan Wenger

Abstract

Linear partial differential equations (PDEs) are an important, widely applied class of PDEs, describing physical processes such as heat transfer, electromagnetism, and wave mechanics. In practice, virtually all PDEs are solved using specialized numerical methods with discretization at their core. However, these algorithms largely fail to return useful estimates of the inherent approximation error. Moreover, classical PDE solvers generally assume all model parameters, such as initial and boundary conditions, and the right-hand side of the PDE, to be known exactly. In this thesis, we develop a general mathematical framework for incorporating mechanistic knowledge in the form of linear PDEs into Gaussian process (GP) models. Our approach frames solving a linear PDE as Bayesian inference given affine observations. Crucially, this allows us to (1) quantify discretization error; and (2) propagate uncertainty in the model parameters to the solution. En route, we generalize a widely used theorem for conditioning GPs on finite-dimensional linear observations to observations made via a bounded linear operator. Demonstrating the applicability of our framework, we show how it recovers existing algorithms and how it can be used in practice to solve stochastic boundary value problems involving linear PDEs.In summary, our results enable the seamless integration of mechanistic models as modular building blocks into probabilistic models by blurring the boundaries between numerical analysis and Bayesian inference.

Zusammenfassung

Lineare partielle Differentialgleichungen (PDEs) sind eine wichtige, weit verbreitete Klasse von PDEs, die genutzt wird, um physikalische Prozesse wie zum Beispiel Wärmetransport, Elektromagnetismus und Wellenmechanik zu beschreiben. In der praxis werden PDEs fast immer mit speziellen numerischen Methoden gelöst, die die Gleichung diskretisieren. Allerdings, stellen diese Algorithmen fast nie Schätzer des inhärenten Approximationsfehlers in der numerischen Lösung zur Verfügung. Zudem nehmen klassische numerische Löser für PDEs im Allgemeinen alle Modellparameter wie etwa Anfangs- und Randwerte und die rechte Seite der PDE als gegeben an. Diese Arbeit entwickelt einen allgemeines mathematischen Gerüst für Methoden, die mechanistisches Wissen in Form einer linearen PDE in probabilistische Modelle basierend auf Gaußprozessen (GPs) integriert. Hierfür wird das numerische Lösen einer PDE als Bayessche Inferenz mit affinen Observationen formuliert. Dieser Ansatz erlaubt es Diskretisierungsfehler zu quantifizieren und Unsicherheit in den Modellparametern auf die Lösung zu propagieren. Als theoretische Grundlage wird ein Satz formuliert und bewiesen, der es erlaubt Gaußprozesse auf Observationen durch einen beschränkten linearen Operator zu konditionieren. Es wird gezeigt, dass existierende klassische und probabilistische Methoden für lineare PDEs Spezialfälle der hier entwickelten Methode sind. Die Anwendbarkeit des Ansatzes wird durch ein praktisches Beispiel demonstriert, in dem ein physikalisches System durch Lösung einer linearen PDE mit stochastischen Randwerten und unsicherer rechter Seite simuliert wird. Zusammenfassend ermöglichen unsere Resultate die nahtlose Integration von mechanistischen Modellen als modulare Blöcke in probabilistische Modelle, indem die Grenzen zwischen numerischer Analysis und Bayesscher Inferenz aufgeweicht werden.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to Philipp Hennig and, especially, Jonathan Wenger. Without your feedback, encouragement, patience, and support, this work would not have been possible. I would also like to extend thanks to Filip Tronarp and Ingo Steinwart for invaluable technical discussions. Moreover, I thank the members of the Methods of Machine Learning group, particularly Jonathan Schmidt, Felix Dangel, Nicholas Krämer, Katharina Ott, Nathanael Bosch, Lukas Tatzel, Julia Grosse, Emilia Magnani, and Franziska Weiler, for an inspiring work atmosphere and intriguing conversations. I am especially grateful for Franziska Weiler's invaluable administrative support. I thank Holger Heidrich, Sarah Müller, Nina Effenberger, Christian Fröhlich, and Samuel Wunderlich, particularly for their open ears in (probably way too) long conversations about this work. I would like to express my thanks to Inge Sörensen for proof-reading parts of this thesis. Finally, I want to express my thanks to Nicole Kämmel, Jörg Pförtner, Bernd Kämmel, and, last but definitely not least, Ann-Kathrin Schäfer, for constant and unwavering emotional support.

Contents

1. Introduction

In the natural sciences, particularly physics, but also in applied fields such as engineering and medicine, partial differential equations (PDEs) are powerful mechanistic models for the behavior of static and dynamic systems with continuous spatial interactions [Borthwick, 2018]. Examples of physical phenomena that can be simulated by such models include heat diffusion, electromagnetism, quantum mechanics, and various branches of continuum mechanics, including fluid and solid mechanics. More specifically, the solutions of PDEs are physically accurate descriptions of e.g. the temperature distribution in a piece of metal, the concentration of chemicals in a liquid during a reaction, the electric potential due to a given distribution of charges, or the magnetic field in the core of a transformer caused by an alternating current in one of its coils, all as functions of time and space [Demtröder, 2015, 2013]. In medical applications, PDE-based fluid dynamics models can be used to analyze the hemodynamics (blood flow) around a mechanical heart valve, aiding engineers in minimizing its thrombogenic potential, i.e. the chance of the device causing blood clots [Dumont et al., 2007]. Moreover, in mechanical engineering, PDEs provide powerful continuous models of material behavior, e.g. its deformation, under the influence of heat and/or strain, for instance when simulating car crashes. However, PDEs can not only be used to simulate physical phenomena, but they find a wide range of applications in pure and applied mathematics (see e.g. [Särkkä and Solin, 2019, Chapter 5], [Müller, 1966], or [Evans, 2010, Chapter 8]). An example in the field of mathematical finance, is the famous Black-Scholes equation which predicts price dynamics of certain types of so-called options [Øksendal, 2003, Section 12.3]. In practice, mechanistic models based on linear PDEs are mainly used for two purposes:

1. *Simulation*: Given a model of a physical system in the form of a partial differential equation, we can simulate the system's state or evolution by finding the solution of this equation subject to a given set of initial and/or boundary conditions. This is useful for predicting the behavior of the system under known conditions. For instance, given all relevant material parameters and all forces involved, a PDE can predict the deformation of and the stress in a metal beam under load. This task is also often referred to as the solution of the forward problem.

2. *Inverse Problem*: Often, the parameters of the PDE are not known, which renders simulation impossible. In this case, we can usually gather a finite set of measurements of the modeled phenomenon in an experiment. Subsequently, the parameters of the PDE can be estimated by identifying parameter values which produce simulations consistent with the measured data. Identifying the parameters of a PDE from measurements of its solution is commonly referred to as an inverse problem.

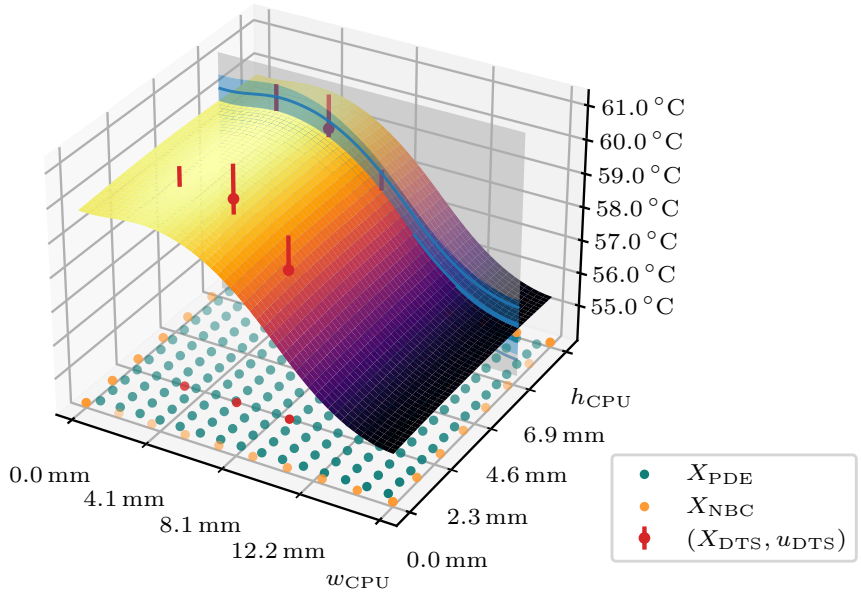


Figure 1.1.: Gaussian process posterior modeling the stationary temperature distribution in an idealized hexa-core CPU die subjected to sustained computational load. The Gaussian process integrates *prior knowledge* about the temperature distribution, *mechanistic knowledge* about heat conduction in the form of a linear PDE, and *empirical knowledge* in the form of noisy temperature measurements $(X_{\mathrm{DTS}}, u_{\mathrm{DTS}})$ taken by so-called *digital thermal sensors* (DTSs) on the CPU die into a common probabilistic model. The surface plot shows the mean function of the Gaussian process, while the 1D slice shows its 95% marginal credible interval along with a few samples. See section 2.4 and chapter 3 for more details about this model.

In the previous example, measurements of the beam's deformation provide information about the material parameters. Typically, solving the inverse problem involves simulating the system repeatedly, which makes inverse problems more difficult to solve than simulation problems.

Unfortunately, hardly any practically relevant PDE can be solved analytically [Borthwick, 2018]. By extension, the same holds true for the solutions to inverse problems. This means that one must generally resort to using numerical solution methods. These algorithms come with several severe downsides.

Problems of Numerical Algorithms Virtually all methods for the numerical solution of PDEs have discretization at their core. This means that the solution estimates produced by these algorithms are inherently subject to approximation error. Generally speaking,

the approximation error can be decreased by choosing a finer discretization, resulting in a trade-off between accuracy of the simulation and computational cost.

Typically, numerical PDE solvers require vast amounts of computational resources to produce accurate results. This is especially problematic in classical approaches to solving inverse problems, since multiple forward solves are typically required, which multiplies the high demand for computational resources. As a result, parameter estimates obtained from these methods might have low accuracy given a fixed computational budget and a PDE which is difficult to solve.

A vast number of solution methods and preconditioners have been proposed in the literature, each featuring different properties and capabilities while making different assumptions about the problem to be solved. To combat both approximation error and computational resource requirements, these methods often exploit problem structure and are hence often highly specialized to a particular type of PDE or even one specific equation. In larger computational pipelines, these solvers are, more often than not, used as monolithic black boxes, which is due to their considerable complexity. While there is often an abundance of knowledge about the problem structure in applied scenarios, it is hence very difficult to tailor the PDE solver to the specific use-case at hand.

While there are error estimation techniques for PDE solvers, the solution estimates are often used as is, especially if the solver is embedded in a larger computational pipeline and the PDE solution is further processed downstream. On the one hand, this is due to the fact that PDE solvers are, as stated above, often implemented as black boxes, while simultaneously failing to provide interpretable and calibrated error estimates by default. However, on the other hand, even when estimates of the approximation error are available, it is not straightforward to find uses for them in downstream computations.

In simulation, the mathematical formulation of the model typically requires its parameters to be known exactly. Examples of such parameters include the position and strength of heat sources or the local charge density in a given region of space, material parameters such as thermal conductivity or dielectric permittivity at every point within some object, or how well a thermal insulator blocks heat flow in and out of a device or experimental setup. This is highly unrealistic in practice, since most of these parameters are only available through noisy experimental measurements or the previous solution of an inverse problem, in which noisy observational data and lacking identifiability lead to errors in the parameter estimation. In larger computational pipelines, these estimation errors propagate and might amplify downstream.

Why Uncertainty? The presence of discretization and estimation errors is particularly severe if decisions are made on the basis of simulation results. To illustrate this, we consider the case of *electrical impedance tomography* (EIT), a non-invasive medical imaging method, in which local electrical properties such as impedance of tissue inside a patient's body are inferred from voltage and current measurements from electrodes placed on the patient's skin [Holder, 2005]. This technique has for instance been applied to detect breast cancer and to provide brain imaging for treating epilepsy and strokes [Holder, 2005]. Essentially, EIT is an inverse problem, in which local material parameters of

Maxwell's equations are estimated from the measurements. Due to measurement noise and the fact that measurements can only be made on the surface of the body, i.e. usually far away from the region of interest, the image predicted by EIT will always be a highly uncertain estimate, possibly with large estimation errors. Calibrated estimates of the error/uncertainty in the prediction is crucial in clinical applications, since decisions about invasive diagnostic or treatment options may need to be made on the basis of such imaging. For instance, in breast cancer screening, a false positive tumor detection could lead to unnecessary biopsy or curative surgery, while failure to detect a tumor is potentially lethal.

Bayesian Methods as Remedies A promising remedy for all of these shortcomings is to equip the solver with the capability to quantify its own uncertainty about the solution of the PDE, both arising from discretization error and from poorly identified model parameters. A very natural way to achieve this is by means of Bayesian statistics, since this framework is designed to estimate unknown quantities by integrating noisy data and uncertain prior knowledge into a common model, while providing consistent uncertainty estimation for all of its predictions. Specifically, we phrase the problem as statistical inference of the unknown solution function, where the data is the observation that the PDE holds. By placing a prior probability measure over the PDE's solution and integrating the observed data by computing the corresponding conditional measures, we obtain a posterior probability measure quantifying the algorithm's uncertainty within a whole set of solution candidates. This is in contrast to the single point estimate returned by classical PDE solvers. The approach outlined above is pursued in the field of *probabilistic numerics* [Hennig et al., 2015, Cockayne et al., 2019, Oates and Sullivan, 2019, Owhadi et al., 2019].

Typically, the aforementioned prior and conditional probability measures take the form of random processes. These in- and output objects turn out to be a very powerful encoding of the solution estimates when integrating Bayesian PDE solvers in larger computational pipelines. This is due to the fact that the associated probability measures virtually never collapse down to a point estimate, which means that the structured uncertainty they provide can be used to aggregate diverse sources of information from all steps in a computational pipeline in a modular fashion. In other words, using the language of random processes to communicate intermediate results in a computational pipeline essentially alleviates the need for the individual components of the pipeline to be aware of one another, e.g. for the purposes of uncertainty propagation.

The assumption of prior knowledge is often criticized in the context of Bayesian machine learning. In the context of PDEs however, there is often an abundance of prior knowledge about the solution. At the very least, this includes mathematical information such as the space of functions in which the solution must lie, because the equation is only well-defined if certain regularity properties such as (weak) differentiability are fulfilled. Moreover, since PDEs often model concrete physical systems, we usually have access to expert knowledge a-priori, including well-studied general physical properties of the system as well as more subjective estimates from previous experience with similar systems.

Additionally, solutions of related but simpler PDEs or, in the context of inverse problems, a previous solve of a PDE with slightly altered parameters might also provide useful prior information, especially if they come with calibrated uncertainty quantification.

Linear PDEs and Gaussian Processes Linear PDEs are an important subclass of partial differential equations, encompassing many widely-used models of the aforementioned phenomena. Moreover, linearization is a common way to approximate more complex nonlinear phenomena. More precisely, a linear PDE is an equation of the form

$$\mathcal{D}[u] = f, \tag{1.1}$$

where $\mathcal{D}$ is a linear differential operator (see definition B.2) [Evans, 2010]. Essentially, this means that $\mathcal{D}$ is a linear map between vector spaces of functions, where $\mathcal{D}[u](x)$ is a linear combination of partial derivatives and the function value of u at x. Note the similarity to a linear system $Ax = b$. A prototypical example of a linear PDE used in thermodynamics, electrostatics and Newtonian gravity is given by the *Poisson equation* [Evans, 2010, Demtröder, 2015]

$$-\Delta u = f, \tag{1.2}$$

where $\Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$ is the so-called Laplacian.

It is a well-known fact that *finite-dimensional* Gaussian distributions are closed under linear maps and conditioning on linear observations [Bishop, 2006]. It has been also been observed that the same holds for Gaussian processes and linear maps that apply to the sample paths such as derivatives or integrals [Rasmussen and Williams, 2006, Särkkä, 2011]. However, even though the latter is a nontrivial statement, to the best of our knowledge, no proof has been provided in the literature. Hence, at this point, this is rather a conjecture than a fact and crucial information such as assumptions under which the statement holds is unavailable.

Bayesian inference in linear-Gaussian models is tractable in closed-form, which makes Gaussian processes a natural choice for the prior when solving linear PDEs. This is why, in this article, we will focus on this type of equation. In essence, when choosing a Gaussian process prior u over the solution, a Bayesian PDE solver for linear PDEs aims at computing $u \,|\, \mathcal{D}[u] - f = 0$. Since this is usually intractable, we will, in the following, develop a general framework for principled approximation of this conditional process, which aims at propagating approximation error to its uncertainty estimate.

2. Inferring the Solutions of Linear PDEs from Gaussian Process Priors

Consider the linear PDE

$$\mathcal{D}[u] = f, \tag{2.1}$$

where $\mathcal{D}\colon U \to V$ is a linear differential operator between spaces U, V of real-valued functions defined on some domain $D \subset \mathbb{R}^d$, and $f \in V$ is the known *right-hand side* function. Our goal is to find $u \in U$ which satisfies equation (2.1) for given $\mathcal{D}$ and f. Any such u is called a *solution* to the PDE. Since there is generally no closed-form expression for the solution u [Borthwick, 2018], we need to estimate it. For illustrative and motivational purposes, we will strongly focus on PDEs which describe physical phenomena such as thermal conduction. In this case, u usually describes the value of some (measurable) physical quantity as a function of space and time, which describes the dynamics of the system exhaustively. Hence, we often have $D = [0, T] \times D'$, where $D' \subset \mathbb{R}^3$ describes the spatial extent of a physical body.

Example 2.1 (Thermal Conduction and the Heat Equation)**.** *Assume that we want to simulate heat conduction in a solid physical body, e.g. a piece of metal. In other words, we want to find a function $u\colon [0, T] \times D \to \mathbb{R}$ of time and space, describing the temperature distribution in a rigid body, whose spatial extent is given by $D \subset \mathbb{R}^3$. Neglecting heat transport due to radiation and convection, we can describe this phenomenon by means of the* heat equation *[Lienhard and Lienhard, 2020], a second-order linear PDE. Its most general form is given in equation* (B.3)*, but assuming spatially and temporally uniform, isotropic material parameters $c_p, \rho, \kappa \in \mathbb{R}$, a simpler equivalent version is given by*

$$c_p \rho \frac{\partial u}{\partial t} - \kappa \Delta u = \dot{q}_V. \tag{2.2}$$

Here, the right-hand side function $f = \dot{q}_V$ is the so-called volumetric heat source *in units of $\mathrm{W\,m^{-3}}$ describing the density of incoming heat energy due to heat sources like electric currents, entering the system at each point of the physical body. This is indeed a linear PDE whose linear differential operator is given by*

$$\mathcal{D} = c_p \rho \frac{\partial}{\partial t} - \kappa \Delta, \tag{2.3}$$

where $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x^2}$ is the Laplacian.

Note however that our focus on PDEs from physics is merely used to provide intuition. The methodology presented in this article applies to any linear PDE.

2.1. PDEs are Indirect Observations of Their Solution

Typically, we think of observations of a physical system as a finite number of direct measurements of the value of the function u, corrupted by noise due to imperfect, finite-precision sensors. It is important to note that u is not uniquely defined by finitely many measurements unless additional assumptions about the function are made. This is due to the fact that, when unconstrained, u could behave arbitrarily complex in between measurements. Contrasting the aforementioned notion of an observation, it is very common to formulate the laws of physics as so-called *conservation laws*, i.e. observations that physical quantities like mass, momentum, charge or energy are conserved over time Borthwick [2018]. These observations have been made in numerous representative experiments and are generally assumed to transfer to any comparable system. Conservation laws from physics are usually formalized in the language of PDEs, which are linear in many practically important cases.

Example 2.1 (continuing from p. 7)**.** *The assumption that the heat equation* (2.2) *holds in a physical body expresses that energy is conserved over time. In particular, it states that the local change of thermal energy at every point in the body is the difference of the thermal energy flowing into the point and the thermal energy flowing out of the point due to thermal conduction and independent local heat sources Lienhard and Lienhard [2020]. It turns out very useful to rearrange the heat equation into*

$$\underbrace{c_p \rho \frac{\partial u}{\partial t}}_{\Delta E_{therm}} - \underbrace{\kappa \Delta u}_{\substack{conduction/\\ diffusion}} - \underbrace{\dot{q}_V}_{\substack{heat\\ sources}} = 0. \tag{2.4}$$

The net-zero balance shows that no energy is lost or gained. Any energy flowing into a region due to diffusion or local heat sources must be accounted for by an increase in internal energy of the material.

Note that this is, generally speaking, an infinite set of observations, since we usually assume that the conservation law holds at each point in time and every point of the domain. In conclusion, using a mechanistic model based on a PDE to simulate a physical system u often amounts to an indirect observation of u, i.e. that a quantity derived from u is conserved in the system.

We can generalize this notion to more abstract PDEs. For instance, as in equation (2.4), we can rearrange the general form of a linear PDE from equation (2.1) into

$$\mathcal{D}[u] - f = 0. \tag{2.5}$$

We can interpret this as an indirect mathematical observation of the function u, where we observe some local mathematical property of u. To be precise, we observe that the image of u under the affine map

$$\mathcal{I}[u] := \mathcal{D}[u] - f \tag{2.6}$$

has a known value 0 at every point of the domain, which provides information about the unknown function u. In the probabilistic numerics literature, the affine map $\mathcal{I}$ is known as an *information operator* [Cockayne et al., 2019]. A concrete example of a (potentially non-affine) information operator which is the analogue of $\mathcal{I}$ in the context of ordinary differential equations is given in Tronarp et al. [2019]. Note that the measurement process defined by the information operator assumes a noise-free observation.

2.2. Solving PDEs as Statistical Estimation

Interpreting a PDE as an indirect local observation of its own solution as above directly entails a fresh perspective on the process of solving these equations. Having established that the information operator (2.6) defined by the PDE provides information about the unknown function u, we can heavily draw on the statistical estimation toolbox. More precisely, we can phrase the numerical problem of finding the solution to a partial differential equation as a statistical estimation problem of an unknown function u from noise-free but indirect observations or measurements u made through $\mathcal{I}$.

Fortunately, in practice, the solution u is not hopelessly unconstrained, but we usually have some a-priori information about it at our disposal. At the very least, we know the space of functions U in which we search for u, but such information might also include noisy measurements of function values $u(x_i)$ at a finite set of points $\{x_i\}_{i=1}^n$ from an experiment, expert knowledge about the rough shape and value range of u, and solutions to related, but different (e.g. simpler) PDEs, just to name a few. This makes the problem particularly amenable to the framework of Bayesian statistics, in which we can incorporate this prior knowledge, be it imprecise or even vague, by means of a prior probability measure over u.

Given such a prior probability measure over u, Bayesian statistics provides us with a principled way to update our knowledge given the information from the measurements induced by the PDE. Namely, in the language of probability theory, enforcing the PDE equation (2.1) means to posit that the event $\mathcal{I}[u] = \mathcal{D}[u] - f = 0$ has occurred. Hence, the appropriate way to incorporate this information into our belief about u is by means of conditioning, i.e. computing the random variable $u \,|\, \mathcal{D}[u] - f = 0$ or, equivalently, $u \,|\, \mathcal{I}[u] = 0$.

2.3. Solving PDEs by Gaussian Process Inference

In this article, the prior over u will always be a Gaussian process

$$u \sim \mathcal{GP}(m, k) \tag{2.7}$$

with mean function $m\colon D \to \mathbb{R}$ and covariance function $k\colon D \times D \to \mathbb{R}$. Gaussian processes are well-suited to represent uncertainty over the solution of a linear PDE, since

1. for certain choices of the covariance function, Gaussian processes define probability measures over the function spaces, in which the PDE's solution is sought.

2. measurement noise often follows a Gaussian distribution.

3. the language of positive kernels, which are used as covariance functions, makes GPs a powerful modeling toolkit for incorporating prior information.

However, the most important reasons for our choice of a Gaussian process prior are its favorable closure properties under linear operations. Among those is the fact that, under certain assumptions, $\begin{pmatrix} u & \mathcal{D}[u] - f \end{pmatrix}^\top$ is a multi-output Gaussian process

$$\begin{pmatrix} u \\ \mathcal{D}[u] - f \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m \\ \mathcal{D}[m] - f \end{pmatrix}, \begin{pmatrix} k & k\mathcal{D}^* \\ \mathcal{D}k & \mathcal{D}k\mathcal{D}^* \end{pmatrix}\right), \tag{2.8}$$

and that the conditional process $u \mid \mathcal{D}[u] - f = 0$ is again a Gaussian process with closed-form expressions for its mean and covariance functions available (see theorem 1). Unfortunately, the closed-form expressions for these mean and covariance functions involve computing the pseudoinverse of the operator $\mathcal{D}k\mathcal{D}^*$, which is at least as difficult as solving the PDE directly.

For this reason, we have to approximate the conditional process by a tractable alternative. Recalling our characterization of the PDE as an observation of a local property of the function u at every point in the domain, we can immediately give a somewhat canonical example of such an approximation. Concretely, we can condition u on the fact that the PDE holds at a finite sequence of well-chosen domain points, i.e. we compute $u \mid \mathcal{D}[u](X) - f(X) = 0$, where $X = \{x_i\}_{i=1}^n \subset \operatorname{int}(D)$. Intuitively speaking, if the set X of domain points is dense enough, we obtain a reasonable approximation to the exact conditional process. This approach, the *probabilistic meshfree method* [Cockayne et al., 2017] is analogous to existing non-probabilistic approaches to solving PDEs, commonly referred to as *collocation methods*, wherein the points X are called *collocation points*. In this case, we can apply corollary 2 to see that the conditional process is again a Gaussian process

$$u \mid \mathcal{D}[u](X) - f(X) = 0 \sim \mathcal{GP}\left(m_{u \mid \mathcal{D}[u](X) - f(X) = 0}, k_{u \mid \mathcal{D}[u](X) - f(X) = 0}\right), \tag{2.9}$$

whose moments

$$m_{u \mid \mathcal{D}[u](X) - f(X) = 0}(x) \coloneqq m(x) + (k\mathcal{D}^*)(x, X)^\top (\mathcal{D}k\mathcal{D}^*)(X, X)^\dagger \left(f(X) - \mathcal{D}[m](X)\right), \tag{2.10}$$

and

$$k_{u \mid \mathcal{D}[u](X) - f(X) = 0}(x_1, x_2) \coloneqq k(x_1, x_2) - (k\mathcal{D}^*)(x_1, X)^\top (\mathcal{D}k\mathcal{D}^*)(X, X)^\dagger (\mathcal{D}k)(X, x_2), \tag{2.11}$$

with

$$(k\mathcal{D}^*)(x_1, x_2) \coloneqq \mathcal{D}k(x_1, t_2)\big|_{t_2 = x_2} \tag{2.12}$$

$$(\mathcal{D}k)(x_1, x_2) \coloneqq (k\mathcal{D}^*)(x_2, x_1) \tag{2.13}$$

$$(\mathcal{D}k\mathcal{D}^*)(x_1, x_2) \coloneqq \mathcal{D}\left(\mathcal{D}k(t_1, t_2)\big|_{t_1 = x_1}\right)\Big|_{t_2 = x_2} \tag{2.14}$$

are now tractable. We will now give an illustrative example on how to apply this method to a concrete problem.

2.4. Example: Modeling the Temperature Distribution in a CPU

Modern central processing units (CPUs) are pieces of computing hardware that are constrained by the vast amounts of heat they dissipate under load. Surpassing the maximum temperature threshold a CPU is rated for a prolonged period of time will likely result in permanent hardware damage or considerable decrease of its longevity [Michaud, 2019]. To counteract overheating, air or water cooling systems are attached to the CPU when deployed. These cooling systems are controlled by temperature sensors on the CPU die, such that, the more heat is dissipated by the die, the more heat is extracted from it by the cooler. Unfortunately, these sensors only provide local measurements of the die's temperature and it is still possible that the temperature surpasses a critical threshold on unmonitored areas of the chip.

In the following, we will develop a probabilistic model for the temperature distribution in an idealized hexa-core CPU die to showcase the capabilities of the methodology developed in this article. For illustrative purposes, we will keep this model deliberately simple. However, there are several possible practical use cases for more realistic and hence more involved versions of this model. For instance, a simulation of the temperature distribution in the whole CPU die based on the temperature measurements mentioned above might aid in monitoring the global maximum of the die temperature while keeping the amount of monitoring hardware at bay. Moreover, should one of the temperature sensors fail, its reading could be replaced by a simulated value.

For simplicity, we assume that the CPU is subjected to sustained computational load and that the cooler is controlled in such a way that the die reaches thermal equilibrium. In thermal equilibrium, the temperature distribution becomes *stationary*, i.e. it does not change over time. Temperature distributions of systems in thermal equilibrium are modeled by the *stationary heat equation* (2.15).

Example 2.1 (continuing from p. 7)**.** *Often, we only care about the temperature distribution after it has reached a steady state, i.e. once it is at thermal equilibrium. Assuming that we ever reach such a state, at thermal equilibrium, it must hold that* $\frac{\partial u}{\partial t} = 0$ *[Lienhard and Lienhard, 2020]. Enforcing this constraint turns the heat equation into the* stationary *or* steady-state heat equation

$$-\kappa \Delta u - \dot{q}_V = 0. \tag{2.15}$$

Due to our choice of material parameters, the steady-state heat equation is, in this case, equivalent to the Poisson equation (1.2) *with* $f = \dot{q}_V/\kappa$*. We will use this equation as a recurring illustrative PDE-based model throughout the first few chapters of this article.*

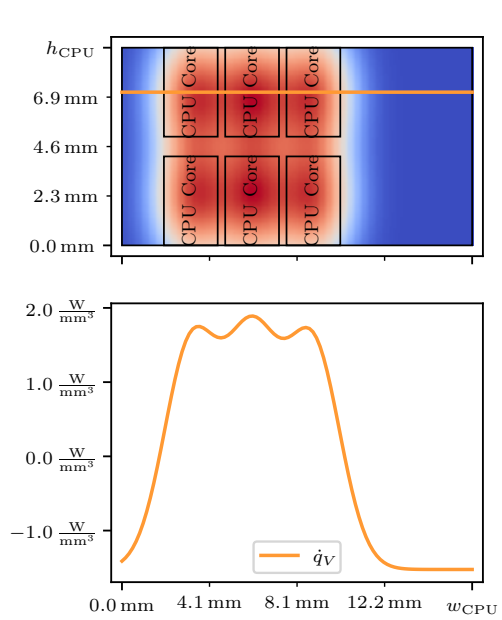
In order for the temperature distribution to remain stationary, the total amount of heat flowing into the CPU must be equal to the amount of heat being drawn from it. In this simulation, we assume that the CPU cores are the only heat sources, while the cooler extracts all heat produced by the cores uniformly over the whole surface of the CPU. These assumptions are expressed by the volumetric heat source $\dot{q}_V$ visualized as a heat map in the top part of figure 2.1(a). This function acts as right-hand side function

in the stationary heat equation. Note that $\dot{q}_V$ integrates to 0 over the whole die, which means that our assumptions about the heat flow are consistent with the assumption of stationarity.

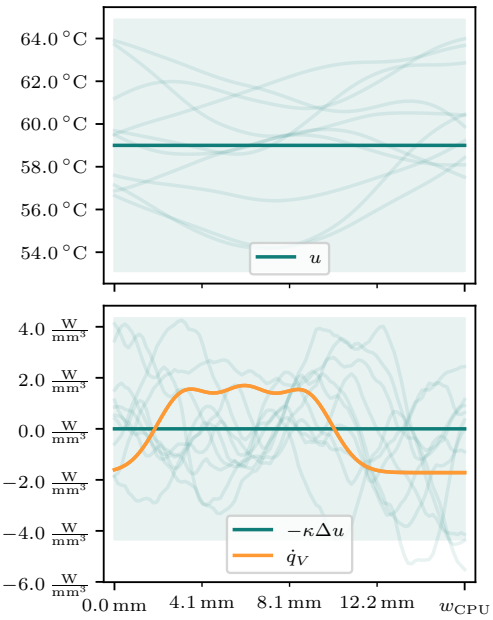
We mostly restrict ourselves to simulating the temperature distribution in a 1D slice across the surface of the CPU, which is visualized in the top part of figure 2.1(a). This is due to the fact that conceptual visualizations of the methodology for two- or three-dimensional models are much harder to read and understand than those of their one-dimensional counterparts. However, this does not mean that our approach is limited to one-dimensional models. In fact, figure 1.1 shows how our approach can be used to simulate the temperature distribution across the full surface of the CPU die. Nevertheless, one might justify the one-dimensional version of the simulation as an approximation to the temperature distribution, which assumes uniformity along its other two dimensions. Indeed, figure 1.1 indicates that this assumption is reasonable, since the temperature distribution is fairly uniform in x_2-direction. The bottom part of figure 2.1(a) shows the values of the volumetric heat source in the slice.

We posit a Gaussian process prior $u \sim \mathcal{GP}\left(m, \sigma^2 k\right)$ with a Matérn-$\frac{7}{2}$ covariance function k over the temperature distribution in the slice whose mean function m and output scale σ are chosen in a realistic range. Figure 2.1(b) shows the prior process u on the top, along with its image $\mathcal{D}[u] \sim \mathcal{GP}\left(\mathcal{D}[m], \sigma^2 \mathcal{D} k \mathcal{D}^*\right)$ under the PDE's differential operator $\mathcal{D} = -\kappa\Delta$ on the bottom. A draw from $\mathcal{D}[u]$ can be interpreted as the distribution of heat sources (and sinks) that must have generated the temperature distribution given by the corresponding draw from u, assuming that the PDE holds. If u solves the PDE, then the uncertainty in $\mathcal{D}[u]$ collapses to 0 and $\mathcal{D}[u]$ and $\dot{q}_V$ coincide.

We can now inform the Gaussian process prior about the mechanistic information encoded in the PDE by choosing a set of collocation points $X_{\text{PDE}} = \{x_{\text{PDE},i}\}_{i=1}^n$ and then conditioning on the observation that the PDE holds (exactly) at these collocation points, i.e. the event $-\kappa\Delta u\left(x_{\text{PDE},i}\right) - \dot{q}_V(x_{\text{PDE},i}) = 0$ for all $i = 1, \ldots, n$. In other words, we update our belief about the temperature distribution in the slice by computing the physically-informed random process $u \mid \text{PDE} \coloneqq u \mid -\kappa\Delta u\left(x_{\text{PDE},i}\right) - \dot{q}_V(x_{\text{PDE},i}) = 0$. The result of this operation is visualized in figure 2.2(a). First of all, we can see that the resulting conditional process indeed solves the PDE exactly at the collocation points (in the lower part of the figure). While the samples exhibit much more similarity to the mean function and as well as less spatial variation, the marginal uncertainty hardly decreases. The latter is mostly due to the fact that the PDE does not identify a unique solution. Indeed, adding any affine function to u does not alter its image under the differential operator, since $\Delta(a^\top x + b) = 0$. This suggests that there is an at least two-dimensional subspace of functions which can not be observed. Classical approaches to formulating PDE-based models resolve this ambiguity by introducing *boundary conditions*. For simplicity, we will impose so-called *Dirichlet boundary conditions* [Evans, 2010], which fix the values of the solution at the boundary, i.e. $u|_{\partial D} = g$ for some known function $g\colon \partial\Omega \to \mathbb{R}$. In our particular setup, the values of the function g can be thought of as measurements of the CPU's temperature in a lab setup. Interpreting the Dirichlet boundary conditions as measurements strongly hints at the canonical way of enforcing



(a) The CPU cores are assumed to be the only heat sources, while the CPU cooler extracts heat uniformly over the whole surface of the CPU. The lower subplot shows the magnitude of heat sources (and sinks) in the 1D slice indicated by the orange line in the upper subplot.



(b) Gaussian process prior over the temperature distribution in the CPU die. If the prior were an exact model of the temperature distribution, then the marginal credible interval of $-\kappa\Delta u$ would collapse to zero and its mean and all samples would coincide with $\dot{q}_V$.

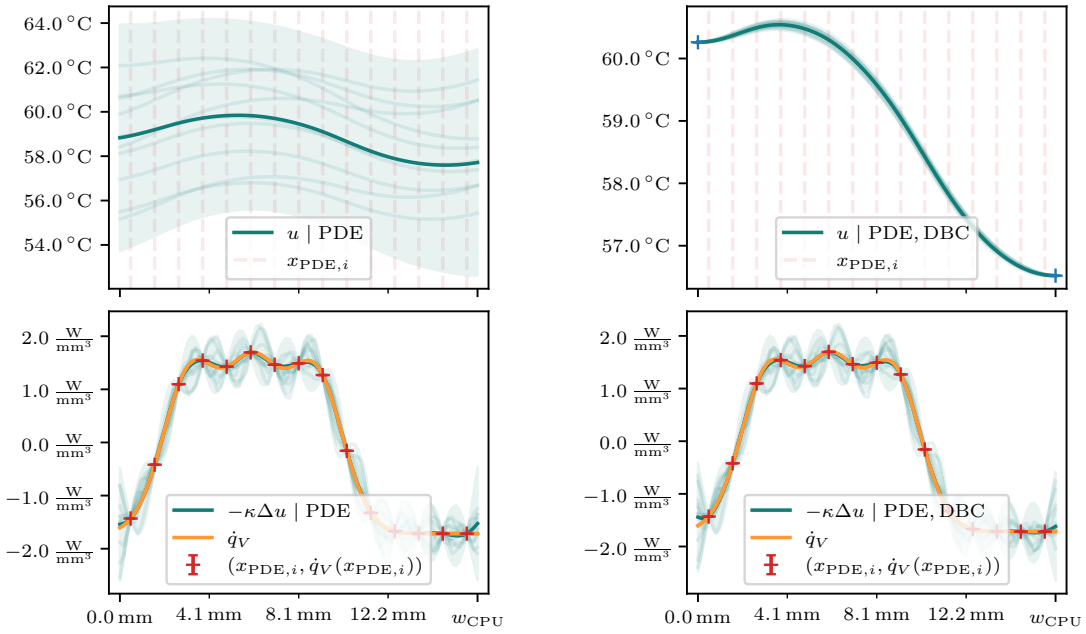
Figure 2.1.: We model the stationary temperature distribution in a CPU die under a sustained computational load. For illustrative purposes, we limit ourselves to the 1D slice along the surface of the die visualized in the top part of figure 2.1(a), assuming a spatially uniform temperature distribution along its other extents. The lower part of figure 2.1(a) shows the assumed volumetric heat source arising from the CPU's thermal dissipation and the cooler acting as a heat sink. To infer the solution of the corresponding PDE in our Bayesian framework, we posit a Gaussian process prior with a Matérn-$\frac{7}{2}$ covariance function over the unknown temperature distribution (see figure 2.1(b)).

$u|_{\partial D} = g$ in our belief. Namely, we can simply interpret the Gaussian process $u \mid \text{PDE}$ as a physics-informed prior and condition on the values of g as in standard GP regression. The resulting conditional process

$$u \mid \text{PDE}, \text{BC} \coloneqq (u \mid \text{PDE}) \,\big|\, u|_{\partial D} = g \tag{2.16}$$

$$= u \,\big|\, -\kappa \Delta u\,(x_{\text{PDE},i}) - \dot{q}_V(x_{\text{PDE},i}) = 0, u|_{\partial D} = g \tag{2.17}$$

is visualized in figure 2.2(b). The remaining uncertainty in u is due to the approximation error introduced by only conditioning on a finite number of collocation points. This can also be seen in the lower part of the figure, since the image of $u \mid \text{PDE}, \text{BC}$ under the differential operator induces a probability measure over possible right-hand sides which are consistent with $u \mid \text{PDE}, \text{BC}$ under the PDE. Intuitively speaking, this also shows that a larger set of collocation points would reduce the uncertainty in $u \mid \text{PDE}, \text{BC}$.



(a) Belief after conditioning on the PDE.

(b) Belief after conditioning on the PDE and the boundary values.

Figure 2.2.: We integrate mechanistic knowledge about the system by conditioning our prior belief u about the temperature distribution in the CPU (see figure 2.1(b)) on PDE observations $-\kappa\Delta u\,(x_{\text{PDE},i}) - \dot{q}_V(x_{\text{PDE},i}) = 0$ at the collocation points $x_{\text{PDE},i}$ (see figure 2.2(a)), resulting in the conditional process $u \mid \text{PDE} \coloneqq u \mid -\kappa\Delta u\,(x_{\text{PDE},i}) - \dot{q}_V(x_{\text{PDE},i}) = 0$. The large remaining uncertainty illustrates that the PDE does not identify a unique solution. By conditioning the physics-informed process $u \mid \text{PDE}$ on boundary values $g\colon \partial D \to \mathbb{R}$ (see figure 2.2(b)), we obtain a posterior process $u \mid \text{PDE}, \text{BC} \coloneqq u \mid -\kappa\Delta u\,(x_{\text{PDE},i}) - \dot{q}_V(x_{\text{PDE},i}) = 0, u|_{\partial D} = g$ whose uncertainty is solely due to discretization error.

3. Propagating Uncertainty in the Input Data to the Solution

In the previous chapter, we inferred the solution of a linear PDE subject to Dirichlet boundary conditions by conditioning a Gaussian process on the observation that the PDE holds at finite number of domain points and on observations of the boundary values. However, in practice, the values of the PDE's right-hand side as well as the initial and boundary values are usually not known exactly. For instance, they might be derived from noisy measurements of some physical quantity or they might be rough estimates arising from experience with related systems. This means that simply conditioning on a point estimate of these values will likely propagate errors to the solution estimate without increasing its uncertainty estimate. Hence, the uncertainty in the resulting belief is not a calibrated estimate of the error in the solution, compared to its true value in an experiment. In our example, underestimating uncertainty in the temperature distribution might lead to overheating of parts of the CPU, which can in turn cause permanent hardware damage, especially if this happens over a longer period of time. Boundary conditions, initial conditions, the right-hand side of the PDE and the coefficients α of the differential operator (see definition B.2) are sometimes collectively referred to as the *input data* [Borthwick, 2018]. We argue that the ability to handle uncertainty in the input data is vital in practical applications, since the assumption that they are known exactly is hardly ever fulfilled. In this chapter, we will show that our GP-based approach admits a natural solution of PDE boundary value problems with uncertain input data by consistently propagating the uncertainty to the solution estimate.

3.1. Uncertain Boundary Conditions

Returning to our sample model of the temperature distribution in a CPU from section 2.4, we note that the boundary conditions constitute somewhat unrealistic assumptions about the typical deployment of such devices. We restricted ourselves to scenarios in which the values $u|_{\partial D}$ of the unknown function on the boundary ∂D of the domain D are given as a boundary function $g\colon \partial D \to \mathbb{R}$, i.e. we imposed Dirichlet boundary conditions. This was due to the fact that, in classical approaches to PDE-based modeling, boundary conditions are required to render the problem well-posed by ensuring that a unique solution exists. In practice, these values could for instance be obtained through experimental measurements. Unfortunately, the absence of a suitable sensor at the boundaries of a deployed CPU entails that the temperature in these locations is generally unavailable in our example model.

We might however assume that the CPU cooler extracts heat (approximately) uniformly from all exposed parts of the CPU's surface, i.e. also from the sides, rather than just from its top surface. Instead of directly specifying the value of the temperature distribution, this assumption provides access to the density $\dot{q}_A$ of heat flowing out of each point on the CPU's boundary. We can use another thermodynamical law to turn this assumption into information about the temperature distribution:

Example 2.1 (continuing from p. 7)**.** Fourier's law *states that the local density of heat* $\dot{q}_A$ *flowing through a surface with normal vector* ν *is proportional to the inner product of the negative temperature gradient and the surface normal* ν*, i.e.*

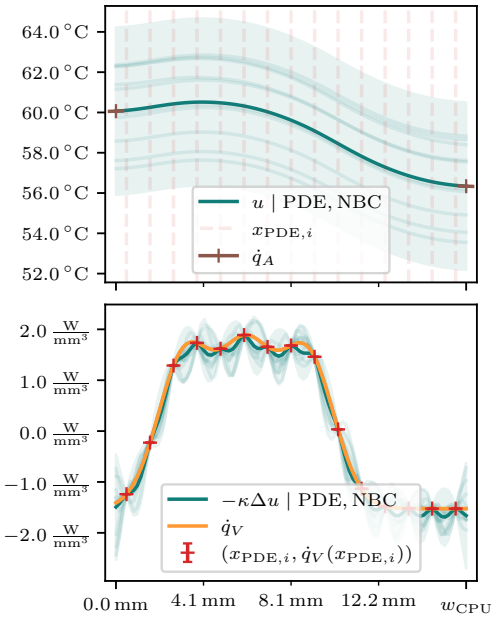
$$\dot{q}_A = -\kappa \langle \nu, \nabla u \rangle , \tag{3.1}$$

where k *is the material's thermal conductivity in* $\mathrm{W\,m^{-1}\,K}$ *[Lienhard and Lienhard, 2020].*

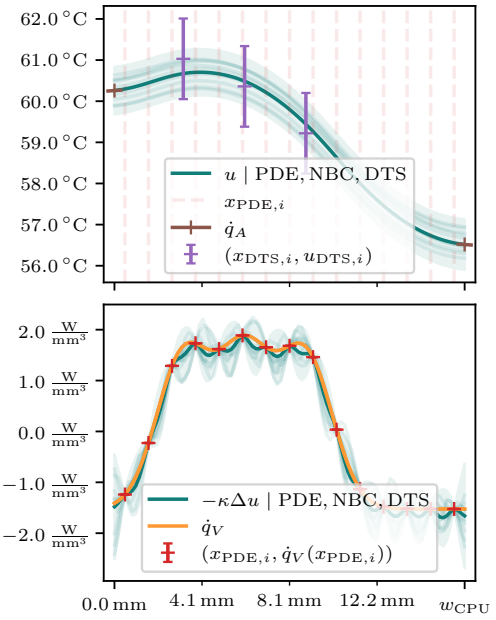
Assuming sufficient differentiability of u, the inner product above is equal to the directional derivative $\partial_\nu u$ of u in direction ν. We can assign an outward-pointing or *exterior normal vector* $\nu(x)$ to (almost) every point $x \in \partial D$ on the boundary of the domain. Since the boundary of the CPU domain is exactly its surface, we can summarize all the above in a boundary condition $-\kappa \partial_{\nu(x)} u\,(x) = \dot{q}_A(x)$ for $x \in \partial D$. This is an example of a so-called *Neumann boundary condition* [Evans, 2010], which specifies the value of the exterior normal derivative at every point of the boundary ∂D of the domain.

Both types of boundary conditions presented above are instances of a much larger class of linear boundary conditions $\mathcal{B}[u] - g = 0$, where $\mathcal{B} \colon U \to W$ is the linear *boundary operator* mapping functions $u \in U$ onto functions $\mathcal{B}[u] : \partial D \to \mathbb{R}^d$ defined on the boundary. Choosing $\mathcal{B}$ such that $\mathcal{B}[u] = u|_{\partial D}$, i.e. the *restriction of* u *onto* ∂D, recovers Dirichlet boundary conditions as a special case, while we can realize Neumann boundary conditions with $\mathcal{B}[u](x) \coloneqq \partial_{\nu(x)} u\,(x)$. Under mild assumptions, we can again use corollary 2 to condition our GP belief about u on linear boundary conditions.

While being more realistic than Dirichlet boundary conditions, our assumptions about the value of $\dot{q}_A$ in the CPU example above are a crude approximation to the true value (for multiple reasons). Consequently, enforcing the resulting Neumann boundary conditions would likely introduce estimation error. Fortunately, our probabilistic framework allows us to turn a point estimate of $\dot{q}_A$ into a distributional estimate, quantifying our uncertainty about the true value of $\dot{q}_A$. Specifically, for the 1D case, we assume a bivariate normal distribution over $\dot{q}_A \coloneqq (\dot{q}_A(0), \dot{q}_A(w_{\mathrm{CPU}}))$. We can then use corollary 2 once more to condition $u \,|\, \mathrm{PDE}$ on the Neumann boundary condition, using the distributional estimate of $\dot{q}_A$. The resulting belief is visualized in figure 3.1(a). The structure of the samples in figure 3.1(a) illustrates that most of the remaining uncertainty about the solution lies in a one-dimensional subspace of U corresponding to constant functions. This is due to the fact that two Neumann boundary conditions on both sides of the domain only determine the solution of the PDE up to an additive constant. Hence, we need an additional data source to address the remaining degree of freedom.



(a) Belief after conditioning $u \mid \text{PDE}$ on uncertain Neumann boundary conditions.



(b) Belief after conditioning $u \mid \text{PDE}, \text{NBC}$ on noisy temperature measurements.

Figure 3.1.: Updated belief about the temperature distribution in the CPU obtained by successively conditioning the physics-informed process $u \mid \text{PDE}$ on different sources of noisy observational data. We first impose uncertain Neumann boundary conditions, approximating the outgoing heat flux at the respective boundary surface elements of the CPU (see figure 3.1(a)). Moreover, since the Neumann boundary conditions only identify the solution of the PDE up to an additive constant, we condition on noisy measurements of the CPU's temperature obtained from digital thermal sensors (DTSs) located at $x_{\text{DTS},i}$, i.e. inside the CPU cores (see figure 3.1(b)).

3.2. Direct Measurements of the Solution

Fortunately, CPUs are equipped with digital thermal sensors (DTSs) located close to each of the cores [Michaud, 2019], which provide (noisy) local measurements of the core temperatures. These measurements can be straightforwardly accounted for in our model by performing standard GP regression using $u \mid \text{PDE}, \text{NBC}$ from figure 3.1(a) as prior. The resulting belief about the temperature distribution is visualized in figure 3.1(b). We can see that integrating the interior measurements effectively reduces the uncertainty due to the remaining degree of freedom, but the conditional measure does not contract completely. The remaining uncertainty is mostly due to the model's consistent accounting for noise in the DTS readings, which is visualized by the fact that the 95% credible interval agrees perfectly with the error bars of the measurements. However, the uncertainty in the Neumann boundary conditions and the discretization error incurred by only choosing a relatively small set of collocation points are also accounted for.

The conditional Gaussian process in figure 1.1 is the two-dimensional analogue of $u \mid \text{PDE}, \text{NBC}, \text{DTS}$ from figure 3.1(b). For this model, we use a tensor product of Matérn-$7/2$ kernels as prior covariance function, a constant prior mean, and the volumetric heat source visualized as a heat map in the top part of figure 2.1(a). The markers in the (x_1, x_2)-plane show the locations X_{PDE} of the PDE collocation points, the locations X_{DTS} of the thermal sensors, and the locations X_{NBC} of the collocation points for the Neumann boundary conditions. The latter are now necessary, since the boundary is now the union of four line segments, i.e. an infinite set of points, and hence, enforcing exact boundary conditions is no longer possible. Note that there is hardly any variation in the temperature distribution along the x_2 axis. This can be seen as justification that the one-dimensional model is a reasonable approximation to the temperature distribution, since it implicitly assumes uniformity along the x_2 axis.

3.3. Uncertainty in the Right-hand Side

Above, we always assumed the right-hand side of the PDE to be known exactly. However, in practice, this assumption is just as unrealistic as the assumption that the values of the heat flux $\dot{q}_A$ across the boundary surface are known exactly. The function $\dot{q}_V$ used up until now is an idealized, crude approximation of the true heat source term. A straightforward attempt at relaxing this assumption is to replace $\dot{q}_V$ by a Gaussian process prior whose mean is given by the estimate of $\dot{q}_V$ used above. Note that, technically speaking, replacing the right-hand side function by a Gaussian process turns the PDE into a stochastic partial differential equation (SPDE).

Physical Consistency Recall that, in order to get rid of the time-dependence, we assumed in section 2.4 that the temperature distribution in the CPU is stationary, i.e. not changing over time. Unfortunately, a naive prior over $\dot{q}_A$ will break the model assumption. This is due to the fact that the temperature distribution can only be stationary if the amount of heat entering the CPU is equal to the amount of heat leaving the CPU via

its surface. If this were not the case, then the CPU would either heat up or cool down. Mathematically, we can express this constraint as

$$\dot{Q}\left[\dot{q}_V, \dot{q}_A\right] \coloneqq \int_{D_{2\mathrm{D}}} \dot{q}_V(x)\,\mathrm{d}x - \int_{\partial D_{2\mathrm{D}}} \dot{q}_A(x)\,\mathrm{d}A = 0, \tag{3.2}$$

where $D_{2\mathrm{D}} \coloneqq [0, w_{\mathrm{CPU}}] \times [0, h_{\mathrm{CPU}}] \subset \mathbb{R}^2$ (see top part of figure 2.1(a)). The (jointly) linear functional $\dot{Q}$ computes the net amount of thermal energy that the CPU gains per unit time. We can use this fact and corollary 1 to construct a joint GP prior for $\dot{q}_V$ and $\dot{q}_A$, which is consistent with the assumption of stationarity. Namely, we can posit a multi-output Gaussian process prior over $\dot{q}_V$ and $\dot{q}_A$, and condition both on the fact that equation (3.2) holds. This is possible because both integrals are linear functionals acting on the paths of the Gaussian processes. In this section, we choose $\dot{q}_V \perp\!\!\!\perp \dot{q}_A$.

In the one-dimensional model developed above, we can simplify equation (3.2) by assuming that heat is drawn uniformly from the sides of the CPU. In this case, the GP prior over $\dot{q}_V$ turns into a four-dimensional Gaussian random vector

$$\begin{pmatrix}\dot{q}_{A,N} & \dot{q}_{A,E} & \dot{q}_{A,S} & \dot{q}_{A,W}\end{pmatrix}^\top \sim \mathcal{N}\left(m_{\dot{q}_A}, \Sigma_{\dot{q}_A}\right)$$

and the stationarity constraint is equivalent to

$$h_{\mathrm{CPU}} \int_{D_{1\mathrm{D}}} \dot{q}_V(x)\,\mathrm{d}x - h_{\mathrm{CPU}}\left(\dot{q}_{A,E} + \dot{q}_{A,W}\right) - w_{\mathrm{CPU}}\left(\dot{q}_{A,N} + \dot{q}_{A,S}\right) = 0, \tag{3.3}$$

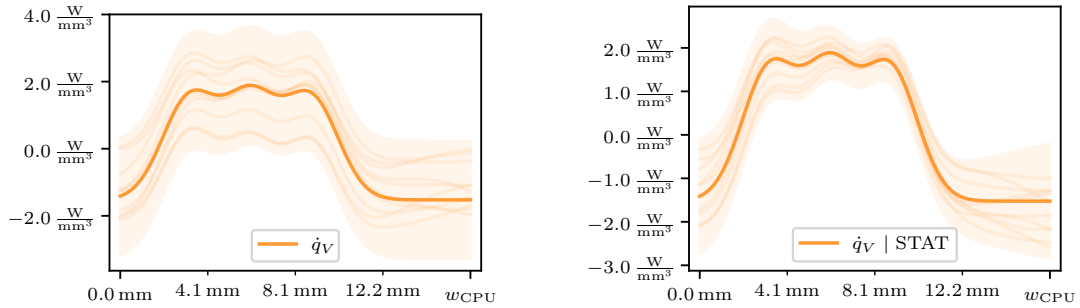
where $D_{1\mathrm{D}} = [0, w_{\mathrm{CPU}}]$. The effect of this constraint on the prior over $\dot{q}_V$ is visualized in figure 3.2(b). The conditional mean is the same as the prior mean, since the prior mean is explicitly constructed to fulfill equation (3.3). However, note that the samples and the marginal credible interval change substantially. Prior samples seem to lie consistently above or below the mean, indicating that there is a net increase or decrease in thermal energy. In contrast, each sample from the conditional process balances values above and below the mean function. This shows that samples from the conditional process $\dot{q}_V, \dot{q}_A \mid$ STAT conserve the amount of thermal energy in the system.

Note that the prior over $\dot{q}_A$ in section 3.1 can also be inconsistent with the assumption of a stationary temperature distribution, which is why we constructed $\dot{q}_V$ and $\dot{q}_A$ from section 3.1 such that equation (3.3) is fulfilled. This also means that $\dot{q}_V$ from section 3.1 is different from $\dot{q}_V$ in section 2.4 and figures 2.1(a), 2.2(a) and 2.2(b).

Since $\dot{q}_V(X_{\mathrm{PDE}}), \dot{q}_A(X_{\mathrm{NBC}}) \mid$ STAT is a Gaussian random vector, we can use corollary 1 to condition our GP prior u on the event

$$\begin{pmatrix} -\kappa\Delta u(X_{\mathrm{PDE}}) - \dot{q}_V(X_{\mathrm{PDE}}) \\ -\kappa\partial_{\nu(X_{\mathrm{NBC}})} u(X_{\mathrm{NBC}}) - \dot{q}_A(X_{\mathrm{NBC}}) \end{pmatrix} = -\kappa \begin{pmatrix} \Delta\left[\cdot\right](X_{\mathrm{PDE}}) \\ \partial_{\nu(X_{\mathrm{NBC}})}[\cdot](X_{\mathrm{NBC}}) \end{pmatrix}[u] - \begin{pmatrix} \dot{q}_V(X_{\mathrm{PDE}}) \\ \dot{q}_A(X_{\mathrm{NBC}}) \end{pmatrix}. \tag{3.4}$$

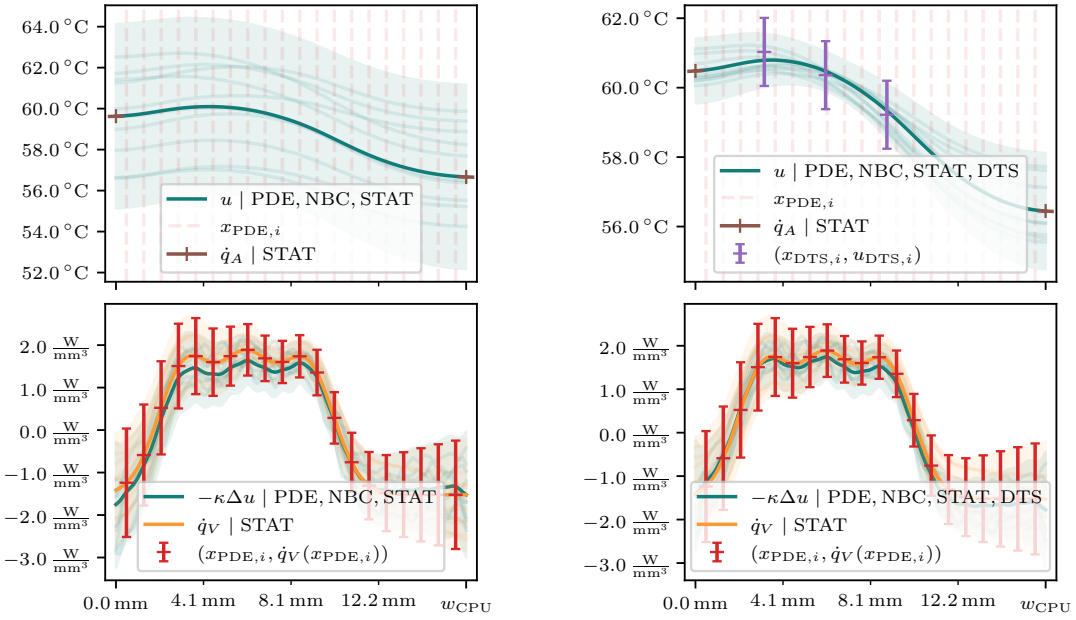
It is important to note that, due to the correlations between $\dot{q}_V(X_{\mathrm{PDE}}) \mid$ STAT and $\dot{q}_A(X_{\mathrm{NBC}}) \mid$ STAT, this is not equivalent to sequentially conditioning on the PDE and the boundary conditions as before. The resulting conditional GP $u \mid \mathrm{PDE}, \mathrm{NBC}, \mathrm{STAT}$



(a) GP prior over the volumetric heat source $\dot{q}_V$, which is inconsistent with the assumption of a stationary temperature distribution.

(b) Conditional GP $\dot{q}_V \mid \mathrm{STAT}$ obtained by conditioning the GP prior $\dot{q}_V$ from figure 3.2(a) on the stationarity constraint equation (3.3).

Figure 3.2.: Construction of a joint prior over the volumetric heat source $\dot{q}_V$ inside the CPU and the outgoing surface heat flux $\dot{q}_A$ on its sides, which is consistent with the assumption of a stationary temperature distribution. We start from independent Gaussian process priors over $\dot{q}_V$ (see figure 3.2(a)) and $\dot{q}_A$ and then condition both on the fact that the heat energy entering the system and the heat energy leaving the system (via the surface) are in balance, i.e. on equation (3.3). The result is a joint (correlated) prior $\dot{q}_V, \dot{q}_A \mid \mathrm{STAT}$, whose marginal $\dot{q}_V \mid \mathrm{STAT}$ is shown in figure 3.2(b).



(a) GP prior u conditioned on the PDE with uncertain right-hand side $\dot{q}_V$ and the Neumann boundary conditions $\dot{q}_A$, after enforcing equation (3.3) on the joint prior over $\dot{q}_V, \dot{q}_A$.

(b) Posterior belief about the temperature distribution in the CPU integrating empirical measurements, subjective knowledge (priors over u, $\dot{q}_V$, $\dot{q}_A$) and mechanistic knowledge (PDE, boundary conditions, stationarity condition equation (3.3))

Figure 3.3.: We integrate knowledge from the joint prior $\dot{q}_V, \dot{q}_A|\text{STAT}$ over the right-hand side of the PDE and the values of the Neumann boundary conditions into our belief about the temperature distribution by conditioning on said PDE and boundary conditions. Afterwards, we use standard GP regression to also take the direct measurements from section 3.2 into account. Note that, technically speaking, we do not solve a regular PDE, but rather a *stochastic partial differential equation* (SPDE) here.

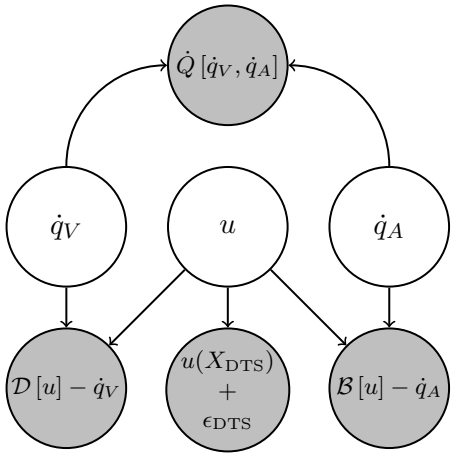


Figure 3.4.: Representation of the CPU model from section 3.3 as a *directed graphical model* [Bishop, 2006, Chapter 8]. As discussed in chapters 2 and 3, inference in this model becomes tractable if we approximate $\mathcal{D}\,[u] - \dot{q}_V = 0$ with $\mathcal{D}\,[u]\,(X_{\mathrm{PDE}}) - \dot{q}_V(X_{\mathrm{PDE}}) = 0$ and $\mathcal{B}\,[u] - \dot{q}_A = 0$ with $\mathcal{B}\,[u]\,(X_{\mathrm{NBC}}) - \dot{q}_A(X_{\mathrm{NBC}}) = 0$. The inference procedure described in section 3.3 is equivalent to the *junction tree algorithm* [Bishop, 2006, Section 8.4.6] applied to this approximation to the graphical model above. Hence, we perform exact inference in an approximate model.

is shown in figure 3.3(a). Comparing figures 3.1(a) and 3.3(a), we can see that, due to the uncertainty in the right-hand side $\dot{q}_A$ of the PDE, the samples exhibit much more spatial variation. Moreover, subjecting the sample paths of $-\kappa\Delta u \mid \mathrm{PDE}, \mathrm{NBC}, \mathrm{STAT}$ from the lower part of figure 3.3(a) to close scrutiny, one can observe that the GP posterior over u learns about the stationarity constraint imposed on $\dot{q}_V$, even from a finite number of observations. Finally, we can, as above, integrate the direct measurements of u from section 3.2 using standard GP regression. This yields the posterior GP shown in figure 3.3(b).

Note that the inference procedure outlined above is a special case of the so-called *junction tree algorithm* applied to (an approximation to) the directed graphical model shown in figure 3.4.

3.3.1. Implementation Details

While it is in principle possible to use automatic differentiation (AD) to compute the kernels $\mathcal{D}k$, $k\mathcal{D}^*$, $\mathcal{D}k\mathcal{D}^*$ and then evaluate equations (2.10) and (2.11) naively, we found that this is detrimental to the performance of the algorithm, especially when performing multiple successive conditioning steps as above. This mostly comes down to two problems. For one, in this example, $\mathcal{D}k\mathcal{D}^*$ contains fourth derivatives, which are expensive to compute when using AD. For another, with each additional conditioning step, the expressions for the conditional mean and covariance need to be substituted into equa-

tions (2.10) and (2.11), which leads to increasingly complex expressions. Our solution to both of these problems relies heavily on block-matrix inversion. Suppose that a GP prior $u \sim \mathcal{GP}(m, k)$ has already been conditioned on observations of the form $\mathcal{L}_1[u] + b_1 = y_1$, where $\mathcal{L}_1$ maps into $\mathbb{R}^{n_1}$ and $b_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$. Assume that we additionally want to condition on $\mathcal{L}_2[u] + b_2 = y_2$, where $\mathcal{L}_2$ maps into $\mathbb{R}^{n_2}$ and $b_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ such that $b_2 \perp\!\!\!\perp b_1$. Then, by corollary 1, the conditional mean and covariance functions of $u \mid \mathcal{L}_1[u] + b_1 = y_1, \mathcal{L}_2[u] + b_2 = y_2$ are given by

$$m_{1:i+1}(x) = m(x) + \begin{pmatrix}(k\mathcal{L}_1)^*(x) & (k\mathcal{L}_2^*)(x)\end{pmatrix} G^{-1} \begin{pmatrix} y_1 - (\mathcal{L}_1[m] - \mu_1) \\ y_2 - (\mathcal{L}_2[m] - \mu_2)\end{pmatrix} \tag{3.5}$$

$$k_{1:i+1}(x_1, x_2) = k(x_1, x_2) + \begin{pmatrix}(k\mathcal{L}_1)^*(x_1) & (k\mathcal{L}_2^*)(x_1)\end{pmatrix} G^{-1} \begin{pmatrix}(\mathcal{L}_1 k)(x_2) \\ (\mathcal{L}_2 k)(x_2)\end{pmatrix}, \tag{3.6}$$

with

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22}\end{pmatrix} \coloneqq \begin{pmatrix}\mathcal{L}_1 k \mathcal{L}_1^* + \Sigma_1 & \mathcal{L}_1 k \mathcal{L}_2^* \\ \mathcal{L}_2 k \mathcal{L}_1^* & \mathcal{L}_2 k \mathcal{L}_2^* + \Sigma_2\end{pmatrix}. \tag{3.7}$$

Since $u \mid \mathcal{L}_1[u] + b_1 = y_1$ is precomputed, we can assume to have access to $G_{11}^{-1} = (\mathcal{L}_1 k \mathcal{L}_1^* + \Sigma_1)^{-1}$, e.g. in the form of a Cholesky decomposition. We can leverage this to compute G^{-1} by block matrix inversion

$$G^{-1} = \begin{pmatrix} G_{11}^{-1} + G_{11}^{-1} G_{12} S^{-1} G_{12}^\top G_{11}^{-1} & -G_{11}^{-1} G_{12} S^{-1} \\ -S^{-1} G_{12}^\top G_{11}^{-1} & S^{-1}\end{pmatrix}, \tag{3.8}$$

where $S = G_{22} - G_{12}^\top G_{11}^{-1} G_{12}$ is the so-called *Schur complement* of G_{11} in G [Boyd and Vandenberghe, 2004]. Moreover, if a Cholesky factorization $G_{11} = L_{11} L_{11}^T$ is available, we can compute the Cholesky factorization of the full matrix G by

$$G = \begin{pmatrix} L_{11} & 0 \\ G_{12}^\top (L_{11}^\top)^{-1} & L_S\end{pmatrix} \begin{pmatrix} L_{11}^\top & L_{11}^{-1} G_{12} \\ 0 & L_S^\top\end{pmatrix}, \tag{3.9}$$

where $S = L_S L_S^\top$ is the Cholesky factorization of the Schur complement. Computing inverses and Cholesky factors in this two-step fashion has the same complexity as a direct computation. However, by implementing it this way, we have access to the intermediate conditional process right before the second process at no additional cost. By induction, this result applies to symmetric positive definite block matrices with arbitrary numbers of blocks.

Note that, in the expressions above, we only need to apply the linear operators to the prior kernel and mean function. We use this to alleviate the need for automatic differentiation. Namely, we manually implement $\mathcal{L}[m]$, $\mathcal{L}k$, $k\mathcal{L}^*$, and $\mathcal{L}k\mathcal{L}^*$ for all combinations of linear operator $\mathcal{L}$, prior mean m and prior covariance function k that we wish to use for inference, with optional fallbacks to an AD framework if a particular combination is not (yet) implemented.

These two techniques lead to a significant speed-up during inference.

3.4. Discussion

In this chapter, we showed how

- exact mechanistic knowledge in the form of a linear PDE subject to Neumann boundary conditions and an integral equation enforcing physical consistency (equation (3.2)),
- uncertain subjective knowledge given by the prior over the temperature distribution u and the priors over the right-hand side $\dot{q}_V$ of the PDE and the boundary heat flux $\dot{q}_A$, and
- noisy empirical measurements

can be fused into a common probabilistic model through Gaussian process inference. We have seen that our model meaningfully accounts for the uncertainties in all sources of information named above. All of this is possible due to the fact that we discard the paradigm of isolating a single solution function. Instead, our method produces an infinite set of solution candidates (the sample paths of the posterior GP) together with a probability measure quantifying our belief that any one of these functions is the true solution to the PDE. While it is possible to obtain such infinite sets of solution candidates in a non-probabilistic approach, the probability is crucial for this to be practically useful. To illustrate this, consider the conditional Gaussian processes from figures 3.1(a) and 3.3(a). When conditioning on the PDE at the collocation points with a known right-hand side, i.e. with no observation noise, we essentially enforce a hard constraint on the paths of the conditional GP. As a result, we remove solution candidates from the set of hypotheses provided by the samples of the prior. Assuming the right-hand side to be a Gaussian process amounts to adding Gaussian observation noise to the PDE observations. Since the Gaussian has support on the entire real line, we can not exclude any solution candidates in this case, which means that the set of solution candidates induced by the conditional GP is the same as the one induced by the prior. However, we can clearly see that the GP is influenced by the PDE observations (and the observations of the boundary conditions). This is due to the fact that the probability measure now prioritizes sample paths, which best explain the observations. In general it is computationally expensive or even intractable to operate on such infinite sets of candidate solutions. Fortunately, Gaussian processes induce probability measures on their sample path spaces for which the type of inference presented above is computationally efficient.

It is possible to obtain "canonical" point estimates of the solution from the probability measure, e.g. by computing the mean function, which is useful in low-uncertainty scenarios such as figure 2.2(b). Nevertheless, what makes the approach powerful is the structured uncertainty captured in samples or the covariance function. For instance, this structured uncertainty makes it possible to further narrow down the uncertain output of a Bayesian PDE solver using additional information in downstream computation. We can see an example of this in figure 3.3(b), where we use the conditional process from figure 3.3(a) as prior and condition on measurements of the solution which amounts to

standard GP regression. The prior, which is essentially the output of a Bayesian PDE solver, encodes the information that sample paths which agree with the SPDE and boundary conditions are more probable to be the solution. As above, we observe that most of the uncertainty lies in a subspace corresponding to constant functions, since these lie in the null space of both the differential and the boundary operator. The conditioning step then selects paths which additionally agree with the temperature measurements. This reduces the uncertainty in the subspace of constant functions, without interfering with the more certain parts of the prior. As a result, samples from the posterior GP still agree with the PDE and the boundary conditions up to the uncertainty in $\dot{q}_V$ and $\dot{q}_A$. If the output of the solver were a point estimate, it would not be evident how we could modify the output so as not to break the fact that it solves the PDE.

We have seen that computational pipelines such as the one presented in this example can be elegantly expressed as directed graphical models. Since our inference algorithm amounts to Gaussian belief propagation on the junction tree of this graphical model, it is evident that the Bayesian PDE solver is a local computation in the global inference procedure on the tree. This shows that GP-based PDE solvers can be implemented in a highly modular fashion, since the implementation of the solver does change based on what happens to the solution estimate and the input data in either upstream or downstream computations. All this information is already handily encoded in the structured uncertainties of the Gaussian processes.

All in all, we argue that Gaussian processes are a powerful modeling language, which makes it possible to implement highly modular computational pipelines aimed at solving potentially ill-posed problems involving linear PDEs by fusing diverse sources of empirical, mechanistic and prior information.

4. Gaussian Process Inference with Affine Observations of the Sample Paths

Throughout chapters 2 and 3 we conditioned Gaussian process priors on affine observations of their paths. More precisely, given a (multi-output) GP prior $f \sim \mathcal{GP}(m, k)$ with index set $\mathcal{X} \subset \mathbb{R}^d$, a linear operator $\mathcal{L}\colon \operatorname{paths}(f) \to \mathbb{R}^n$ acting on the paths of f, and a Gaussian random vector $g \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^n$ with $g \perp\!\!\!\perp f$, we computed the conditional random process

$$f \mid \mathcal{L}[f] + g = y \tag{4.1}$$

for some $y \in \mathbb{R}^n$. Formally, this object is defined as the family

$$(f \mid \mathcal{L}[f] + g = y) \coloneqq \{f_x \mid \mathcal{E}\}_{x \in \mathcal{X}}, \tag{4.2}$$

where $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ is the probability space on which both f and g are defined, $\mathcal{E}$ is the event $\mathcal{E} \coloneqq h^{-1}(\{y\}) \in \mathcal{B}(\Omega)$ and h is the random variable

$$h\colon \Omega \to \mathbb{R}^n, \omega \mapsto \mathcal{L}[f(\cdot, \omega)] + g(\omega). \tag{4.3}$$

We refer to appendices A.1 and A.2 for definitions of the objects mentioned above. For instance, in chapter 2, we use $\mathcal{L} \coloneqq (\mathcal{D}[\cdot](x_i))_{i=1}^n$, where $\mathcal{D}$ is a linear differential operator, as well as $\mathcal{L}\left[\tilde{f}\right] \coloneqq (\tilde{f}(x_i))_{i=1}^n$, and, in chapter 3, we additionally use

$$\mathcal{L}\left[\tilde{f}\right] = \int_D \tilde{f}(x)\,\mathrm{d}x. \tag{4.4}$$

It is well-known that h is a Gaussian random vector

$$h \sim \mathcal{GP}(\mathcal{L}[m] + \mu, \mathcal{L}k\mathcal{L}^* + \Sigma), \tag{4.5}$$

where $\mathcal{L}k\mathcal{L}^* \in \mathbb{R}^{n \times n}$ with

$$(\mathcal{L}k\mathcal{L}^*)_{ij} = \mathcal{L}\left[t \mapsto \mathcal{L}[k(t, \cdot)]_j\right]_i, \tag{4.6}$$

and that the conditional random process is a Gaussian process

$$f \mid \mathcal{L}[f] + g = y \sim \mathcal{GP}(m_{h=y}, k_{h=y}) \tag{4.7}$$

with conditional moments given by

$$m_{h=y}(x) = m(x) + \mathcal{L}[k(\cdot, x)]^\top (\mathcal{L}k\mathcal{L}^* + \Sigma)^{-1} (y - (\mathcal{L}[m] + \mu)), \tag{4.8}$$

and

$$k_{h=y}(x_1, x_2) = k(x_1, x_2) + \mathcal{L}\left[k(\cdot, x_1)\right]^\top \left(\mathcal{L}k\mathcal{L}^* + \Sigma\right)^{-1} \mathcal{L}\left[k(\cdot, x_2)\right] \tag{4.9}$$

and this result is widely-used in the literature (see e.g. Graepel [2003], Rasmussen and Williams [2006], Särkkä [2011], Särkkä et al. [2013], Cockayne et al. [2017], Raissi et al. [2017], Agrell [2019], Albert [2019], Krämer et al. [2022]). Unfortunately, to the best of our knowledge, no complete proof of the result has been published yet. Since the above are nontrivial claims about potentially ill-behaved infinite-dimensional objects, a proof would however be highly important, be it just to identify a precise set of assumptions about the objects at play, which are requires so that the result holds. For instance, it is possible that h is not a random variable (because it might not measurable), i.e. $\mathcal{E}$ might not be measurable.

To remedy this situation, a major contribution of this work are theorem 1 and corollaries 1 and 2 and their proof in appendix A, which provide a sequence of increasingly results capturing the claims above as a special case. Hence, besides being the theoretical basis for this work, theorem 1 and corollaries 1 and 2 also provide theoretical backing for many of the publications cited above.

Our results identify a set of mild assumptions, which are easy to verify and widely-applicable in practical applications. Assumption 1 constitutes the common set of assumptions shared by theorem 1 and corollaries 1 and 2. See section 4.1 for information on how to verify assumption 1 in a practical scenario.

Assumption 1. *Let*

$$f \sim \mathcal{GP}\left(m_f, k_f\right) \tag{4.10}$$

be a Gaussian process prior with index set $\mathcal{X}$ on the Borel probability space $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$, whose mean function and sample paths lie in a real separable Hilbert function space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with $\mathcal{H}_k \subset \mathcal{H}$ and with continuous point evaluation functionals. Let $\mathcal{L}\colon \mathcal{H} \to \mathcal{H}_\mathcal{L}$ be a bounded linear operator mapping the paths of f into a separable Hilbert space $\mathcal{H}_\mathcal{L}$.

We start our exposition here by presenting theorem 1, our most general result. Using theorem 1, it is possible to condition Gaussian processes on affine observations of their paths, which take values in arbitrary and potentially infinite-dimensional separable Hilbert spaces. For instance, this means that conditioning on observations of a whole function (instead of just a finite number of function evaluations) is possible, given that the assumptions of theorem 1 are fulfilled. The formulation of this theorem heavily relies on the theory of *Gaussian measures on separable Hilbert spaces*, some of which is detailed is appendix A.3 and appendix A.6.

Theorem 1 (Affine Gaussian Process Inference)**.** *Let assumption 1 hold. Then $\omega \mapsto f(\cdot, \omega)$ is an $\mathcal{H}$-valued Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with mean m and covariance operator $h \mapsto \mathcal{C}_f[h](x) = \langle k(x, \cdot), h\rangle_\mathcal{H}$. We also write $f \sim \mathcal{N}(m, \mathcal{C}_f)$. Let $g \sim \mathcal{N}(m_g, \mathcal{C}_g)$ be an $\mathcal{H}_\mathcal{L}$-valued Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $g \perp\!\!\!\perp f$. Then*

$$\begin{pmatrix} f \\ \mathcal{L}[f] + g \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_f \\ \mathcal{L}[m_f] + m_g \end{pmatrix}, \begin{pmatrix} \mathcal{C}_f & \mathcal{C}_f\mathcal{L}^* \\ \mathcal{L}\mathcal{C}_f & \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g \end{pmatrix}\right), \tag{4.11}$$

with values in $\mathcal{H} \times \mathcal{H}_{\mathcal{L}}$ *and hence*

$$\mathcal{L}[f] + g \sim \mathcal{N}(\mathcal{L}[m_f] + m_g, \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g). \tag{4.12}$$

If $\operatorname{ran}(\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)$ *is closed, then for all* $h \in \mathcal{H}_{\mathcal{L}}$

$$f \mid \mathcal{L}[f] + g = h \sim \mathcal{GP}\left(m_{f|\mathcal{L}[f]+g=h}, k_{f|\mathcal{L}[f]+g=h}\right), \tag{4.13}$$

where the conditional mean and covariance function are given by

$$m_{f|\mathcal{L}[f]+g=h(x)} = m_f(x) + \left\langle \mathcal{L}[k_f(\cdot, x)], (\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)^\dagger [h - (\mathcal{L}[m_f] + m_g)] \right\rangle_{\mathcal{H}_{\mathcal{L}}}, \tag{4.14}$$

and

$$k_{f|\mathcal{L}[f]+g=h(x_1,x_2)} = k_f(x_1, x_2) - \left\langle \mathcal{L}[k_f(\cdot, x_1)], (\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)^\dagger \mathcal{L}[k_f(\cdot, x_2)] \right\rangle_{\mathcal{H}_{\mathcal{L}}}, \tag{4.15}$$

respectively.

Unfortunately, especially in the context of PDEs, theorem 1 is difficult to apply in practice, since the operator $\mathcal{L}\mathcal{C}_f\mathcal{L}^*$ is infinite-dimensional and its pseudoinverse (if it exists) usually has no analytic form. However, as seen in chapters 2 and 3, its corollaries can, in practical scenarios, be applied to great effect. Corollary 1 enables affine observations, in which the GP sample paths enter through one or multiple continuous linear functionals. For example, we used corollary 1 in section 3.3 to condition on observations of a GPs. To state the result conveniently, we introduce some notation.

Notation 1. *Let* $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a positive-definite kernel and let* $\mathcal{L}_i\colon \mathcal{H}_k \to \mathbb{R}^{n_i}$ *for* $i = 1, 2$ *be bounded linear operators. We define the functions*

$$\mathcal{L}_1 k\colon \mathcal{X} \to \mathbb{R}^{n_1}, x \mapsto \mathcal{L}_1[k(\cdot, x)], \tag{4.16}$$

$$k\mathcal{L}_2^*\colon \mathcal{X} \to \mathbb{R}^{n_2}, x \mapsto \mathcal{L}_2[k(x, \cdot)], \tag{4.17}$$

$$\tag{4.18}$$

and the matrix[1] $\mathcal{L}_1 k \mathcal{L}_2^* \in \mathbb{R}^{n_1 \times n_2}$ *with entries*

$$(\mathcal{L}_1 k \mathcal{L}_2^*)_{ij} \coloneqq \mathcal{L}_2[(\mathcal{L}_1 k)_i]_j \tag{4.19}$$

$$= \mathcal{L}_1[(k\mathcal{L}_2^*)_j]_i. \tag{4.20}$$

[1] This notation is motivated by lemma A.7, which also shows that the two different ways to compute the entries of $\mathcal{L}_1 k \mathcal{L}_2^*$ are consistent.

Corollary 1. *Let assumption 1 hold for $\mathcal{H}_\mathcal{L} = \mathbb{R}^n$ and let $g \sim \mathcal{N}(\mu_g, \Sigma_g)$ be an $\mathbb{R}^n$-valued Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $g \perp\!\!\!\perp f$. Then*

$$\mathcal{L}[f] + g \sim \mathcal{N}(\mathcal{L}[m_f] + \mu_g, \mathcal{L}k_f\mathcal{L}^* + \Sigma_g) \tag{4.21}$$

and

$$f \mid \mathcal{L}[f] + g = h \sim \mathcal{GP}\left(m_{f|\mathcal{L}[f]+g=h}, k_{f|\mathcal{L}[f]+g=h}\right), \tag{4.22}$$

with conditional mean and covariance function given by

$$m_{f|\mathcal{L}[f]+g=h}(x) = m_f(x) + \mathcal{L}[k_f(x,\cdot)]^\top (\mathcal{L}k_f\mathcal{L}^* + \Sigma_g)^\dagger (h - (\mathcal{L}[m_f] + m_g)), \tag{4.23}$$

and

$$k_{f|\mathcal{L}[f]+g=h}(x_1, x_2) = k_f(x_1, x_2) - \mathcal{L}[k_f(x_1,\cdot)]^\top (\mathcal{L}k_f\mathcal{L}^* + \Sigma_g)^\dagger \mathcal{L}[k_f(\cdot, x_2)]. \tag{4.24}$$

Finally, we turn to corollary 2, which is the result that is most widely-used throughout the literature [Graepel, 2003, Särkkä, 2011, Särkkä et al., 2013, Cockayne et al., 2017, Raissi et al., 2017, Agrell, 2019, Albert, 2019, Krämer et al., 2022]. It shows how Gaussian processes can be conditioned on point evaluations of the image of their paths under a linear operator, provided that the linear operator is bounded and maps into a Hilbert function space, on which point evaluation is continuous. Moreover, it shows that, under these conditions, the image of the GP under the linear operator is itself a Gaussian process. Again, we introduce some notation to facilitate stating the result.

Notation 2. *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite kernel and, for $i = 1, 2$, let $\mathcal{L}_i\colon \mathcal{H}_k \to \mathbb{R}^{n_i}$ be a bounded linear operator mapping into a real Hilbert function space $\mathcal{H}_i \subset \mathbb{R}^{\mathcal{X}_i}$ with continuous point evaluation functionals. In analogy to notation 1, we define the bivariate functions*

$$k\mathcal{L}_2^*\colon \mathcal{X} \times \mathcal{X}_2 \to \mathbb{R},\ (x, x_2) \mapsto \mathcal{L}_2[k(x,\cdot)](x_2), \tag{4.25}$$

$$\mathcal{L}_1 k\colon \mathcal{X}_1 \times \mathcal{X} \to \mathbb{R},\ (x_1, x) \mapsto \mathcal{L}_1[k(\cdot, x)](x_1), \quad \text{and} \tag{4.26}$$

$$\mathcal{L}_1 k \mathcal{L}_2^*\colon \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}, (x_1, x_2) \mapsto \mathcal{L}_2[(\mathcal{L}_1 k)(x_1,\cdot)](x_2) = \mathcal{L}_1[(k\mathcal{L}_2^*)(\cdot, x_2)](x_1). \tag{4.27}$$

Corollary 2. *Let assumption 1 hold, where $\mathcal{H}_\mathcal{L} \subset \mathbb{R}^{\mathcal{X}'}$ is a space of real valued functions defined on $\mathcal{X}'$ such that the point evaluation functionals $\delta_{x'}\colon \mathcal{H}_\mathcal{L} \to \mathbb{R}, h \mapsto h(x)$ for all $x \in \mathcal{X}'$ are continuous. Let*

$$g \sim \mathcal{GP}(m_g, k_g) \tag{4.28}$$

be a Gaussian process with index set $\mathcal{X}'$ on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $g \perp\!\!\!\perp f$. Then

$$\mathcal{L}[f] + g \sim \mathcal{GP}(\mathcal{L}[m] + m_g, \mathcal{L}k_f\mathcal{L}^* + k_g), \tag{4.29}$$

and, for $X' = \{x_i'\}_{i=1}^n \subset \mathcal{X}'$ and $h \in \mathbb{R}^n$,

$$f \mid \mathcal{L}[f](X') + g(X') = h \sim \mathcal{GP}\left(m_{f|X',h}, k_{f|X',h}\right) \tag{4.30}$$

with

$$m_{f|X',h}(x) \coloneqq m_f(x) + (k_f\mathcal{L}^*)(x,X')(\mathcal{L}k_f\mathcal{L}^* + k_g)(X',X')^\dagger \left(h - \left(\mathcal{L}\left[m_f\right]\left(X'\right) + m_g(X')\right)\right) \tag{4.31}$$

and

$$k_{f|X',h}(x_1,x_2) \coloneqq k_f(x_1,x_2) - (k_f\mathcal{L}^*)(x_1,X')(\mathcal{L}k_f\mathcal{L}^* + k_g)(X',X')^\dagger(\mathcal{L}k_f)(X',x_2). \tag{4.32}$$

where

$$(k_f\mathcal{L}^*)(x,X') = \left((k_f\mathcal{L}^*)(x,x_i')\right)_{i=1}^n \in \mathbb{R}^{1\times n} \tag{4.33}$$

$$(\mathcal{L}k_f)(X',x_2) = \left((\mathcal{L}k_f)(x_i',x)\right)_{i=1}^n \in \mathbb{R}^n \tag{4.34}$$

$$(\mathcal{L}k_f\mathcal{L}^* + k_g)(X',X') = \left((\mathcal{L}k_f\mathcal{L}^*)(x_i',x_j') + k_g(x_i',x_j')\right)_{i,j=1}^n \in \mathbb{R}^{n\times n} \tag{4.35}$$

$$\mathcal{L}\left[m_f\right]\left(X'\right) = \left(\mathcal{L}\left[m_f\right](x_i)\right)_{i=1}^n \in \mathbb{R}^n \tag{4.36}$$

$$m_g(X') = \left(m_g(x_i)\right)_{i=1}^n \in \mathbb{R}^n. \tag{4.37}$$

If additionally $\mathcal{X} = \mathcal{X}'$, then

$$\begin{pmatrix} f \\ \mathcal{L}\left[f\right] + g \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m_f \\ \mathcal{L}\left[m_f\right] + m_g \end{pmatrix}, \begin{pmatrix} k_f & k_f\mathcal{L}^* \\ \mathcal{L}k_f & \mathcal{L}k_f\mathcal{L}^* + k_g \end{pmatrix}\right). \tag{4.38}$$

This corollary is is the theoretical basis for chapter 2 and most of chapter 3. Note that, for $\mathcal{L} = \mathrm{id}$, we recover standard GP regression as a special case in corollary 2.

Remark 4.1. *Theorem 1 and corollaries 1 and 2 also apply if the GPs involved are multi-output GPs. In this case, the sample paths are functions $I \times \mathcal{X} \to \mathbb{R}$ with $I = \{1,\dots,d\}$ by definition A.4. In order to apply linear operators defined on functions $\mathcal{X} \to \mathbb{R}^d$, we interpret a sample path $f(\cdot,\omega)\colon I \times \mathcal{X} \to \mathbb{R}$ as a function*

$$\tilde{f}(\cdot,\omega)\colon \mathcal{X} \to \mathbb{R}^d, x \mapsto \left(f((i,x),\omega)\right)_{i=1}^d \in \mathbb{R}^d. \tag{4.39}$$

4.1. On Prior Selection

A typical choice for the solution space U of a linear PDE, especially in the context of weak solutions (see chapter 5), are *Sobolev spaces* [Adams and Fournier, 2003]. Unfortunately, it is impossible to formulate a Gaussian process prior u, whose paths are elements of a Sobolev space U. This is due to the fact that Sobolev spaces are, technically speaking, not function spaces, but rather spaces of equivalence classes $[f]_\sim$ of functions, which are equal almost everywhere [Adams and Fournier, 2003]. By contrast, the path spaces of Gaussian processes are proper function spaces, which means that, in this setting, $\mathrm{paths}\,(u) \subseteq U$ is impossible.

Fortunately, if the path space can be continuously embedded in U, i.e. there is a continuous and injective linear operator $\iota\colon \mathrm{paths}\,(u) \to U$, commonly referred to as an

embedding, then the inference procedure above can still be applied. If such an embedding exists, we can interpret the paths of the GP as elements of U by applying ι implicitly. For instance, $\mathcal{D}[u]$ is then a shorthand notation for $\mathcal{D}[\iota[u]]$. Fortunately, since the embedding is assumed to be continuous, the conditions for GP inference with linear operator observations are still met when applying ι implicitly. The canonical choice for the embedding in the case of Sobolev spaces is $\iota[u] = [u]_{\sim_U}$.

Example 4.1 (Matérn covariances and Sobolev spaces)**.** *Kanagawa et al. [2018] show that, under certain assumptions, the sample spaces of GP priors with* Matérn covariance functions *[Rasmussen and Williams, 2006] are continuously embedded in Sobolev spaces whose smoothness depends on the parameter ν of the Matérn covariance function. To be precise, let $D \subset \mathbb{R}^d$ be open and bounded with Lipschitz boundary such that the cone condition [Adams and Fournier, 2003, Definition 4.6] holds. Denote by $k_{\nu,l}$ the Matérn kernel with smoothness parameter $\nu > 0$ and lengthscale $l > 0$. Then, with probability 1, the sample paths of a Gaussian process f with covariance function $k_{\nu,l}$ are contained in any RKHS $\mathcal{H}_{k_{\nu',l'}}$ with $l' > 0$ and*

$$0 < \underbrace{\nu' + \frac{d}{2}}_{=:m'} < \nu \tag{4.40}$$

[Kanagawa et al., 2018, Corollary 4.15 and Remark 4.15]. Moreover, if $m' \in \mathbb{N}$, then the RKHS $\mathcal{H}_{k_{\nu',l'}}$ is norm-equivalent to the Sobolev space $H^{m'}(D)$ [Kanagawa et al., 2018, Example 2.6]. This implies that the canonical embedding

$$\iota\colon \operatorname{paths}(f) \to H^{m'}(D), f(\cdot,\omega) \mapsto [f(\cdot,\omega)]_{\sim_{H^{s'}(D)}} \tag{4.41}$$

is continuous.

For $U = H^{m'}(D)$, the example above shows that the Matérn covariance function $k_{\nu,l}$ with $\nu = m' + \epsilon$ for any $\epsilon > 0$ leads to an admissible GP prior. The choice $\epsilon = \frac{1}{2}$ makes evaluating the covariance function particularly efficient [Rasmussen and Williams, 2006]. However, note that the elements of the Sobolev space $H^m(D)$ are only m-times weakly differentiable, which means that $H^2(D)$ is not an admissible choice in chapters 2 and 3.

Remark 4.2 (Sobolev Spaces and Strong Derivatives)**.** *The* Sobolev embedding theorem *[Adams and Fournier, 2003, Theorem 4.12] gives conditions under which the elements of a Sobolev space are embedded into Banach spaces of continuously differentiable functions. Let $D \subset \mathbb{R}^d$ be open and bounded with Lipschitz boundary such that the cone condition [Adams and Fournier, 2003, Definition 4.6] holds. Let $j \geq 0$, $m \geq 1$ be integers. If $m > \frac{d}{2}$, then there is a continuous embedding*

$$\iota\colon H^{j+m}(D) \to C_B^j(D), \tag{4.42}$$

where $C_B^j(D)$ is the space of continuously differentiable functions with bounded derivatives, which is a Banach space under the norm

$$\|f\|_{C_B^j(D)} = \max_{0 \leq |\alpha| \leq j} \sup_{x \in D} |D^\alpha f(x)|. \tag{4.43}$$

Moreover, point-evaluated partial derivatives on $C_B^j(D)$ are continuous linear functionals, since, for any multi-index $|\alpha'| \leq j$ and any $x' \in D$, we have

$$\left|D^{\alpha'}\left[f\right]\left(x'\right)\right| \leq \sup_{x\in D}\left|D^{\alpha'}f\left(x\right)\right| \leq \max_{0\leq|\alpha|\leq j}\sup_{x\in D}|D^{\alpha}f\left(x\right)| = \left\|f\right\|_{C_B^j(D)}. \tag{4.44}$$

Example 4.2 (Strong Derivatives in Matérn Sample Spaces)**.** *Under the assumptions of example 4.1, for a prior GP f with Matérn covariance function $k_{\nu,l}$ such that $\nu := m + k + \epsilon$, where $\epsilon > 0$ and*

$$k := \begin{cases} \frac{d}{2} + \frac{1}{2} & \text{if } d \text{ is odd,} \\ \frac{d}{2} + 1 & \text{if } d \text{ is even,} \end{cases} \tag{4.45}$$

we have the following chain of continuous embeddings

$$\text{paths}\left(f\right) \hookrightarrow H^{m+k}\left(D\right) \hookrightarrow C_B^m(D). \tag{4.46}$$

As noted in remark 4.2, point-evaluated partial derivatives of order $\leq m$ are continuous linear functionals on $C_B^m(D)$. It follows that a point-evaluated differential operator $\mathcal{D}\left[\cdot\right]\left(x\right)$ of order $\leq m$ is a continuous linear functional on $\text{paths}\left(f\right)$ *if the two continuous embeddings are prepended.*

In chapters 2 and 3, we have $d = 1$ and a GP prior with Matérn covariance function where $\nu = \frac{7}{2} = 2 + k + \frac{1}{2}$. It follows that point-evaluated differential operators of order ≤ 2 are continuous linear functionals. Hence, the assumptions of corollary 2 are fulfilled, which means that the inference procedure used in these chapters is supported by our theoretical results above.

5. Gaussian Process Approximation of Weak Solutions to Linear PDEs

Many models of physically plausible phenomena are expressed as functions u, which are not (continuously) differentiable or not even continuous. See Evans [2010, Section 1.3.2], Borthwick [2018, Section 1.2], or von Harrach [2021, Kapitel 1] for some examples. This means that these phenomena can not be the classical solution to a PDE. Moreover, it turns out that there are PDEs derived from established physical principles, which do not admit a classical solution at all. To solve both of these problems, one can weaken the notion of differentiability and the notion of a solution to a PDE. This leads to so-called *weak solutions*. In fact, many of the aforementioned phenomena are weak solutions to specific PDEs.

In the following, we will give an intuitive, yet superficial treatment of weak solution theory for linear PDEs by considering the weak formulation of the stationary heat equation for non-homogeneous media

$$-\operatorname{div}(\kappa \nabla u) = \dot{q}_V \tag{5.1}$$

as an example. Note that, for a constant κ, this equation turns into the version of the stationary heat equation used throughout chapters 2 and 3. Our exposition here is largely based on Evans [2010, Section 6.1.2] and we refer the reader there for additional information.

Let $D \subset \mathbb{R}^d$ be an open and bounded domain and assume that $u \in C^2(D)$, $\kappa \in L_\infty(D)$, and $\dot{q}_V \in L_2(D)$. If u is a solution to equation (5.1), then we can integrate both sides of the equation against a so-called *test function* $v \in C_c^\infty(D)$, i.e. an infinitely smooth function with compact support (see definition B.5), which results in

$$-\int_D \operatorname{div}(\kappa \nabla u)(x)\, v(x)\, \mathrm{d}x = \int_D \dot{q}_V(x) v(x)\, \mathrm{d}x. \tag{5.2}$$

Since both u and v are sufficiently differentiable, we can apply integration by parts to the first integral to obtain

$$\int_D \langle \kappa(x) \nabla u(x), \nabla v(x) \rangle \, \mathrm{d}x = \int_D \dot{q}_V(x) v(x)\, \mathrm{d}x, \tag{5.3}$$

since $v|_{\partial D} = 0$. Note that this expression does not only make sense if $u \in C^2(D)$, but also if u is once *weakly differentiable* (see [Evans, 2010, Section 5.2.1]) with $\nabla u \in L_2(D)^d$. Intuitively speaking, a weak derivative of a (classically non-differentiable) function "behaves

like a derivative" when integrated against a smooth test function. These relaxed requirements on u are exactly the defining properties of the Sobolev space $H^1(D) \supset C^2(D)$, i.e. it suffices that $u \in H^1(D)$. Denote by $H_0^1(D)$ the closure of $C_c^\infty(D)$ in $H^1(D)$. Then there is a sequence $(v_m)_{m=1}^\infty \subset C_c^\infty(D)$ with $v_m \to v$ (in $H^1(D)$) for every $v \in H_0^1(D)$ and equation (5.3) is continuous in v (in $H^1(D)$-norm). Hence, we can also relax the requirements on v to $v \in H_0^1(D) \supset C_c^\infty(D)$. Let

$$B[u,v] \coloneqq \int_D \langle \kappa(x) \nabla u(x), \nabla v(x) \rangle \, \mathrm{d}x. \tag{5.4}$$

Note that B is bilinear. Then, for $u \in H^1(D)$ and $v \in H_0^1(D)$, equation (5.3) is equivalent to

$$B[u,v] = \langle \dot{q}_V, v \rangle_{L_2}. \tag{5.5}$$

We define a *weak solution* of equation (5.1) as $u \in H^1(D)$ such that equation (5.5) for all $v \in W_0^{1,u}$. Moreover, equation (5.5) is commonly referred to as the *weak* or *variational formulation* of equation (5.1).

Definition 5.1. *A* weak formulation *of a PDE is an equation of the form*

$$B[u,v] = l[v], \tag{5.6}$$

where $B \colon U \times V \to \mathbb{R}$ is a bilinear form and $l \colon V \to \mathbb{R}$ is a continuous linear functional. A vector $u \in U$ is a weak solution *of the PDE if it solves equation* (5.6) *for all test functions $v \in V$.*

Remark 5.1. *We can recover the strong solution of the PDE from a weak formulation by choosing*

$$B[u,v] \coloneqq \langle v, \mathcal{D}[u] \rangle_V \qquad \text{and} \qquad l[v] \coloneqq \langle v, f \rangle_V. \tag{5.7}$$

This is due to the fact that, for any weak solution $u^\star \in U$ of the PDE, we have $\mathcal{D}[u^\star] - f \in V$ and hence

$$\begin{aligned} \|\mathcal{D}[u^\star] - f\|_V^2 &= \langle \mathcal{D}[u^\star] - f, \mathcal{D}[u^\star] - f \rangle_V && (5.8)\\ &= \langle \mathcal{D}[u^\star] - f, \mathcal{D}[u^\star] \rangle_V - \langle \mathcal{D}[u^\star] - f, f \rangle_V && (5.9)\\ &= \langle \mathcal{D}[u^\star] - f, f \rangle_V - \langle \mathcal{D}[u^\star] - f, f \rangle_V && (5.10)\\ &= 0, && (5.11) \end{aligned}$$

i.e. $\mathcal{D}[u^\star] - f = 0$. This implies that $u^\star$ is actually a strong solution of the PDE, because $\mathcal{D}[u^\star] = f$.

5.1. The Petrov-Galerkin Method[1]

Unfortunately, equation (5.6) is only rarely analytically solvable, so we need to find an approximate solution. A very common strategy is to replace U and V by finite

[1] This section is loosely based on [Fletcher, 1984] and [von Harrach, 2021].

dimensional subspaces

$$\hat{U} = \operatorname{span}(u_1, \ldots, u_m) \subset U \tag{5.12}$$

$$\hat{V} = \operatorname{span}(v_1, \ldots, v_n) \subset V \tag{5.13}$$

and solve the weak formulation in those, i.e. we now want to find $u \in \hat{U}$ such that equation (5.6) holds for all $v \in \hat{V}$. The functions in $\hat{U}$ are dubbed *trial functions*, while the functions in $\hat{V}$ are, perhaps confusingly, also referred to as test functions [Fletcher, 1984]. Note however that the functions in $\hat{V}$ do not need to have the same smoothness properties as the test functions from definition B.5. Due to the fact that $\hat{U}$ and $\hat{V}$ are finite-dimensional, this is formulation is equivalent to solving the linear system

$$\hat{B}\hat{c}^{\mathrm{PG}} = \hat{l}, \tag{5.14}$$

where $\hat{B} \in \mathbb{R}^{n \times m}$ and $\hat{l} \in \mathbb{R}^n$ are defined by

$$\hat{B}_{ij} \coloneqq B[u_j, v_i] \qquad \text{and} \tag{5.15}$$

$$\hat{l}_i \coloneqq l[v_i]. \tag{5.16}$$

Any solution $\hat{c}^{\mathrm{PG}}$ to this linear system corresponds to a solution

$$\hat{u}^{\mathrm{PG}} \coloneqq \sum_{i=1}^{m} c_i^{\mathrm{PG}} u_i \in \hat{U} \tag{5.17}$$

of equation (5.6) in $\hat{U}$ and $\hat{V}$. This approach to approximating the solution of a weak formulation is known as the *generalized Galerkin* or *Petrov-Galerkin method* [Fletcher, 1984]. If $n = m$ and $u_i = v_i$ for all $i = 1, \ldots, m$, then the method is know as the *Ritz* or *(Ritz-)Galerkin method*.

Among the most important properties of the Petrov-Galerkin method is the notion of "orthogonality" of the *residual* $r \coloneqq u^\star - \hat{u}^{\mathrm{PG}}$, i.e. the error of the approximation, to the subspace $\hat{U}$ w.r.t. the bilinear form B. To be precise, we have

$$B[r, v_i] = \underbrace{B[u^\star, v_i]}_{=l[v_i]} - \underbrace{B\left[\hat{u}^{\mathrm{PG}}, v_i\right]}_{=l[v_i]} = 0, \tag{5.18}$$

since $u^\star$ solves equation (5.6) for all $v \in V \supset \hat{V}$, $\hat{u}^{\mathrm{PG}}$ solves the equation for all $v \in \hat{V}$, and $v_i \in \hat{V}$.

The Petrov-Galerkin method gives rise to a whole family of accurate and versatile numerical methods for approximating weak solutions of linear PDEs, the most important subfamilies of which are *spectral methods*, *finite volume methods* and the ubiquitous *finite element method* (FEM).

Example 5.1 (Spectral methods)**.** *We obtain a canonical example of a spectral method if we choose truncated Fourier bases for $\hat{U}$ and $\hat{V}$, e.g.*

$$u_{2i} = v_{2i} = \cos(i\pi x) \tag{5.19}$$

$$u_{2i+1} = v_{2i+1} = \sin(i\pi x) \tag{5.20}$$

for $i = 0, \dots, n-1$.

Example 5.2 (Finite Element Methods)**.** *Generally speaking, finite element methods are (Petrov-)Galerkin methods, where the functions in the test and trial bases have compact support, i.e. they are nonzero only in a highly localized region of the domain. The archetype of a finite element method chooses piecewise linear test and trial functions, which are linear on each element of a triangulation of the domain. For instance, on a one-dimensional domain* $D = [-1, 1]$*, this amounts to fixing a grid* $-1 = x_0 < \cdots < x_{n+1} = 1$ *and then choosing the basis functions*

$$u_i(x) = v_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{if } x_{i-1} \le x \le x_i, \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & \text{if } x_i \le x \le x_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.21}$$

for $i = 1, \dots, n$*. Note that multiplying a coordinate vector* $c \in \mathbb{R}^n$ *with these basis functions indeed leads to a piecewise linear interpolation between the points*

$$(x_0, 0), (x_1, c_1), \dots, (x_n, c_n), (x_{n+1}, 0),$$

since, for $x \in [x_i, x_{i+1}]$*,*

$$\sum_{i=1}^{n} c_i u_i(x) = c_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + c_{i+1} \frac{x - x_i}{x_{i+1} - x_i} = \left(1 - \frac{x - x_i}{x_{i+1} - x_i}\right) c_i + \left(\frac{x - x_i}{x_{i+1} - x_i}\right) c_{i+1}.$$

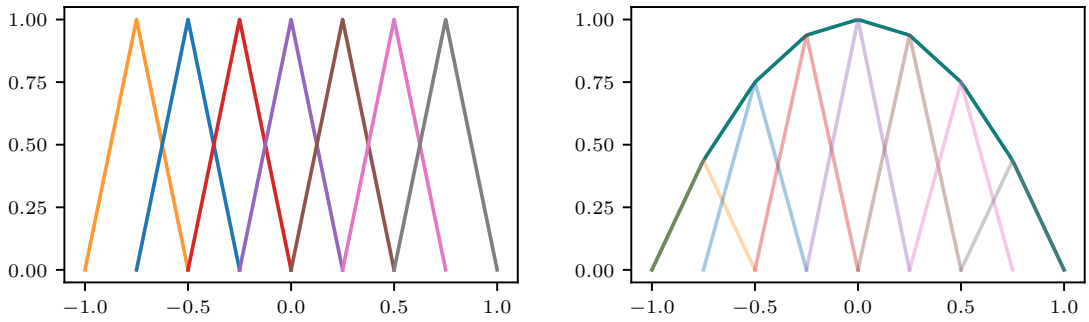
The basis functions and an element in their span are visualized in figure 5.1. It is also common to use piecewise polynomials of higher order as basis functions.

5.2. A Hierarchical Bayesian Framework for Approximating (Weak) Solutions to Linear PDEs

In this section, we will develop a unifying framework for GP-based methods aimed at approximating both weak and strong solutions to linear PDEs. Our framework is essentially a generalization of the Petrov-Galerkin approach applied to Gaussian process inference. We will show that our framework reproduces a large family of non-probabilistic numerical methods for linear PDEs including *symmetric collocation* [Fasshauer, 1997, 1999], the *Petrov-Galerkin method*, and hence *finite-element methods*, *spectral methods*, and *finite volume methods*.

Consider a linear PDE in weak formulation, i.e. we want to solve

$$B[u, v] = l[v] \qquad \forall v \in V \tag{5.22}$$



(a) Basis functions $u_i = u_i$ spanning the test and trial function spaces $\hat{U} = \hat{V}$. The functions are defined on the whole interval [-1, 1], but we only show the non-zero parts of the functions to avoid clutter in the figure above.

(b) An element from $\hat{U} = \hat{V}$, i.e. a linear combination of the basis functions on the left. The linear combination results in a function which takes the values of the coefficients at the peak locations of the respective basis functions and interpolates linearly between the points.

Figure 5.1.: Typical test and trial function spaces $\hat{U} = \hat{V}$ of piecewise linear functions used in the finite element method in one dimension.

for $u \in U$, where $l[v] = \langle f, v \rangle_V$ for some $f \in V$. We additionally require that B is continuous for fixed $v \in V$, i.e. for any $v \in V$ there must be a constant $C < \infty$ such that

$$B[u, v] \leq C \|u\|_U \tag{5.23}$$

for all $u \in U$. As shown in remark 5.1, this does not limit our method to approximation of weak solutions, since strong solutions can also be found using weak formulations. Let

$$\begin{pmatrix} u \\ f \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m_u \\ m_f \end{pmatrix}, \begin{pmatrix} k_{uu} & k_{uf} \\ k_{fu} & k_{ff} \end{pmatrix} \right) \tag{5.24}$$

be a multi-output Gaussian process prior over the weak solution u and the right-hand side f of the PDE, whose path space can be continuously embedded into $U \times V$ (see section 4.1 for more details on the latter assumption). Taking inspiration from section 5.1, we choose subspaces $\hat{U} \subset U$ and $\hat{V} = \operatorname{span}(v_1, \ldots, v_n) \subset V$ of trial and test functions. By applying a bounded projection $\mathcal{P}_{\hat{U}} \colon U \to \hat{U}$ onto $\hat{U}$, i.e. $\mathcal{P}_{\hat{U}}^2 = \mathcal{P}_{\hat{U}}$, $\left\|\mathcal{P}_{\hat{U}}\right\| < \infty$, and $\operatorname{ran}\left(\mathcal{P}_{\hat{U}}\right) = \hat{U}$, we obtain a canonical parametric approximation

$$u_{\hat{U}} \coloneqq \mathcal{P}_{\hat{U}}[u] \in \hat{U} \tag{5.25}$$

of the GP in the subspace $\hat{U}$. Afterwards, in analogy to the Petrov-Galerkin method, we can then solve the weak form in $\hat{U}$ and $\hat{V}$ by conditioning the prior u on the event

$$B\left[u_{\hat{U}}, v_i\right] - \langle f, v_i \rangle_V = 0 \qquad \forall i \in \{1, \ldots, n\}, \tag{5.26}$$

which we also denote by

$$B\left[u_{\hat{U}}, \hat{V}\right] - \left\langle f, \hat{V}\right\rangle_V = 0, \tag{5.27}$$

where

$$B\left[\cdot, \hat{V}\right]: U \to \mathbb{R}^n, u \mapsto \left(B\left[u_{\hat{U}}, v_i\right]\right)_{i=1}^n, \tag{5.28}$$

$$\left\langle f, \hat{V}\right\rangle_V: V \to \mathbb{R}^n, v \mapsto \left(\langle f, v_i\rangle_V\right)_{i=1}^n. \tag{5.29}$$

By corollary 1, the resulting conditional random process

$$u \mid B\left[u_{\hat{U}}, \hat{V}\right] - \left\langle f, \hat{V}\right\rangle_V = 0 \tag{5.30}$$

is again Gaussian. In the following, we will introduce the two main types of methods that can be derived from this general framework. Additionally, we will show that they recover certain classical methods in the conditional mean.

5.2.1. Gaussian Process Projection Methods

The simplest choice for the subspace $\hat{U} \subset U$ is arguably to set $\hat{U} = U$, which means that $\mathcal{P}_{\hat{U}} = \mathrm{id}_U$. We refer to this class of methods as *Gaussian process projection methods*, since evaluating $B[u, v_i]$ and $\langle f, v_i\rangle_V$ shares certain similarities with projections on v_i. This type of method is computationally feasible, as long as we can efficiently evaluate the expressions

$$\mathcal{L}\left[\begin{pmatrix} m_u \\ m_f \end{pmatrix}\right], \qquad \mathcal{L}\left[\begin{pmatrix} k_{uu} & k_{uf} \\ k_{fu} & k_{ff} \end{pmatrix}(x, \cdot)\right], \qquad \text{and} \qquad \mathcal{L}\begin{pmatrix} k_{uu} & k_{uf} \\ k_{fu} & k_{ff} \end{pmatrix}\mathcal{L}^*,$$

which appear in the conditional moments in corollary 1, where

$$\mathcal{L}: U \to \mathbb{R}^n, u \mapsto \left(B[u, v_i] - l[v_i]\right)_{i=1}^n. \tag{5.31}$$

This might not always be possible in closed-form, since B often involves computing integrals. In these cases one could fall back to a numeric quadrature method.

A prominent example of a method realized by choosing $\hat{U} = U$ in our framework is the *probabilistic meshfree method* used in chapters 2 and 3.

Example 5.3 (Symmetric Collocation)**.** *Point evaluation $\delta_x: V \to \mathbb{R}^k, v \mapsto v(x)$ on a function space $V \subset (\mathbb{R}^k)^D$ is a linear functional. If it is additionally continuous on the Hilbert space V, then, by Riesz' representation theorem [Yosida, 1995, Section III.6], there is a function[2] $\delta_x^* \in V$ such that $v(x) = \delta_x[v] = \langle \delta_x^*, v\rangle_V$ for all $v \in V$.*

Hence, if the given weak formulation corresponds to a PDE in strong formulation as in equation (5.7), all point evaluation functionals on V are continuous, $\hat{U} = U$ and $v_i = \delta_{x_i}^* \in V$, then we have

$$B[u, v_i] - l[v_i] = \mathcal{D}[u](x_i) - f(x_i), \tag{5.32}$$

[2] In a reproducing kernel Hilbert space $\mathcal{H}_k$, this function is given by $\delta_x^* = k(x, \cdot)$, since $f(x) = \langle k(x, \cdot), f\rangle_{\mathcal{H}_k}$ by the reproducing property.

and hence, we recover the probabilistic meshfree method *from [Cockayne et al., 2017] and chapters 2 and 3. Cockayne et al. [2017] show that the conditional mean of this approach reproduces* symmetric collocation *[Fasshauer, 1997, 1999], a non-probabilistic approximation method for strong solutions of PDEs, in the conditional mean.*

Unfortunately, the probabilistic meshfree method can only be applied in approximating strong solutions of linear PDEs, since the point evaluation functionals are usually not continuous on the spaces V considered for finding a weak solution.

However, more general Gaussian process projection methods are suitable for approximating weak solutions. For instance, a weak solution of the stationary heat equation in nonhomogeneous media from above can be approximated by choosing the piecewise linear functions from figure 5.1 as test basis functions.

5.2.2. Gaussian Process Galerkin Methods

Since our framework is heavily inspired by the Galerkin approach, a direct translation of Galerkin-type methods to the language of GP inference is relatively straightforward. Namely, if we choose $\hat{U}$ to be finite-dimensional, e.g. $\hat{U} = \operatorname{span}(u_1, \dots, u_m)$, then we recover a probabilistic version of the Petrov-Galerkin method from section 5.1. We dub the resulting class of methods *Gaussian process Galerkin methods*.

At first, Gaussian process Galerkin methods might seem inferior to Gaussian process projection methods, since former has a finite-dimensional trial function space, while the latter has an infinite-dimensional trial function space. However, note that the conditional mean of Gaussian process projection methods is also only updated by a linear combination of n functions, while the covariance function receives an at most rank n downdate. This means that, effectively, Gaussian process projection methods also have a finite-dimensional trial function space, which is implicitly constructed from the test function basis, the bilinear form B and the prior covariance function k_{uu}. In certain scenarios it is actually very beneficial to choose the trial basis manually. For instance, one might want the inference procedure to learn the low-frequency components first, so as to iteratively refine a global solution approximation.

Since $\hat{U}$ is finite-dimensional, there is a bounded linear operator $\mathcal{P}_{\mathbb{R}^m} \colon U \to \mathbb{R}^m$ such that

$$\mathcal{P}_{\hat{U}}[\tilde{u}] = \sum_{i=1}^{m} c_i u_i =: \mathcal{I}_{\mathbb{R}^m}^{\hat{U}}[c] \tag{5.33}$$

where the coefficients $c \coloneqq \mathcal{P}_{\mathbb{R}^m}[\tilde{u}]$ are the coordinates of $\mathcal{P}_{\hat{U}}[u]$ in $\hat{U}$ and $\mathcal{I}_{\mathbb{R}^m}^{\hat{U}} \colon \mathbb{R}^m \to \hat{U}$ is the canonical isomorphism between $\mathbb{R}^m$ and $\hat{U}$. Hence, we get the factorization

$$\mathcal{P}_{\hat{U}} = \mathcal{I}_{\mathbb{R}^m}^{\hat{U}} \mathcal{P}_{\mathbb{R}^m}. \tag{5.34}$$

The canonical choice for the projection $\mathcal{P}_{\hat{U}}$ would arguably be orthogonal projection w.r.t. the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{k'_{uu}}}$ of the sample space $\mathcal{H}_{k'_{uu}} = \operatorname{paths}(u)$ of u. However, this inner product is generally difficult to compute. Fortunately, we can use the L_2 inner

products or Sobolev inner products on the samples to induce a (usually non-orthogonal) projection $\mathcal{P}_{\hat{U}}$.

Example 5.4. *If the functions in U are square-integrable, the linear operator*

$$\mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i \coloneqq P^{-1}\left(\int_D u_i(x)u'(x)\,\mathrm{d}x\right)_{i=1}^m, \tag{5.35}$$

where

$$P_{ij} \coloneqq \int_D u_i(x)u_j(x)\,\mathrm{d}x, \tag{5.36}$$

induces a projection $\mathcal{P}_{\hat{U}} = \mathcal{I}^{\hat{U}}_{\mathbb{R}^m}\mathcal{P}_{\mathbb{R}^m}$ onto $\hat{U} \subset U$, even if $\langle\cdot,\cdot\rangle_U \neq \langle\cdot,\cdot\rangle_{L_2}$.

Proof.

$$\begin{aligned}
\mathcal{P}^2_{\hat{U}}\left[u'\right] &= \mathcal{P}_{\hat{U}}\left[\sum_{i=1}^m \mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i u_i\right] && (5.37)\\
&= \sum_{i=1}^m \mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i \mathcal{P}_{\hat{U}}\left[u_i\right] && (5.38)\\
&= \sum_{i=1}^m \mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i \sum_{j=1}^m \mathcal{P}_{\mathbb{R}^m}\left[u_i\right]_j u_j && (5.39)\\
&= \sum_{j=1}^m u_j \sum_{i=1}^m \mathcal{P}_{\mathbb{R}^m}\left[u_i\right]_j \mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i && (5.40)\\
&= \sum_{j=1}^m u_j \sum_{i=1}^m \left(\sum_{k=1}^m (P^{-1})_{jk}\left\langle u_k, u_i\right\rangle_{L_2}\right)\mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i && (5.41)\\
&= \sum_{j=1}^m u_j \sum_{i=1}^m \left(\sum_{k=1}^m (P^{-1})_{jk}P_{ki}\right)\mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i && (5.42)\\
&= \sum_{j=1}^m u_j \sum_{i=1}^m \left(P^{-1}P\right)_{ji}\mathcal{P}_{\mathbb{R}^m}\left[u'\right]_i && (5.43)\\
&= \sum_{j=1}^m u_j \mathcal{P}_{\mathbb{R}^m}\left[u'\right]_j && (5.44)\\
&= \mathcal{P}_{\hat{U}}\left[u'\right] && (5.45)
\end{aligned}$$

□

The factorization of $\mathcal{P}_{\hat{U}}$ from equation (5.34) allows a useful modification of the inference procedure proposed above. Note that, for $c \in \mathbb{R}^m$,

$$B\left[\mathcal{I}^{\hat{U}}_{\mathbb{R}^m}\left[c\right], v_i\right] = (\hat{B}c)_i, \tag{5.46}$$

where $\hat{B}$ is defined as in section 5.1. Let $\mathcal{L}_{\hat{V}} \colon V \to \mathbb{R}^n, v \mapsto (\langle v, v_i \rangle_V)_{i=1}^n$. Hence, we can equivalently apply our inference procedure to the model

$$(u, f) \sim \mathcal{GP}(m, k) \tag{5.47}$$
$$\hat{c}_u \coloneqq \mathcal{P}_{\mathbb{R}^m}[u] \tag{5.48}$$
$$\hat{l}_f \coloneqq \mathcal{L}_{\hat{V}}[f] \tag{5.49}$$

with observations

$$\hat{B}\hat{c}_u - \hat{l}_f = 0, \tag{5.50}$$

where

$$\begin{pmatrix} \hat{c}_u \\ \hat{l}_f \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathcal{P}_{\mathbb{R}^m}[m_u] \\ \mathcal{L}_{\hat{V}}[m_f] \end{pmatrix}, \begin{pmatrix} \mathcal{P}_{\mathbb{R}^m} k_{uu} \mathcal{P}^*_{\mathbb{R}^m} & \mathcal{P}_{\mathbb{R}^m} k_{uf} \mathcal{L}^*_{\hat{V}} \\ \mathcal{L}_{\hat{V}} k_{fu} \mathcal{P}^*_{\mathbb{R}^m} & \mathcal{L}_{\hat{V}} k_{ff} \mathcal{L}^*_{\hat{V}} \end{pmatrix} \right) \tag{5.51}$$

is the joint prior belief about the solution's coordinates in the subspace $\hat{U}$ and the discretized right-hand side of the PDE. Inference in this model is best performed hierarchically. First, we update our belief about the solution's coordinates in $\hat{U}$ by compute the conditional random variable

$$\hat{c}_u^\star \coloneqq \hat{c}_u \,\Big|\, \hat{B}\hat{c}_u - \hat{l}_f = 0\,, \tag{5.52}$$

which is also Gaussian. Next, we can reuse the Gram matrix that is part of the conditional moments of $\hat{c}_u^\star$ and the conditional mean itself to compute

$$\hat{u}^\star \coloneqq u \,\Big|\, \hat{B}\hat{c}_u - \hat{l}_f = 0\,. \tag{5.53}$$

This is useful, because it disentangles the errors due to discretization and those due to uncertainty in the right-hand side, which can be seen as follows.

$$u^\star_{\hat{U}} \coloneqq \mathcal{I}^{\hat{U}}_{\mathbb{R}^m}[\hat{c}_u^\star] = \mathcal{P}_{\hat{U}}[u] \,\Big|\, \hat{B}\hat{c}_u - \hat{l}_f = 0 \tag{5.54}$$

is a parametric Gaussian process modeling the projection of the solution estimate onto $\hat{U}$. If $m \geq n$, then the uncertainty in $u^\star_{\hat{U}}$ is solely due to the uncertainty in the right-hand side. Moreover, the uncertainty due to discretization error is modeled by the Gaussian process

$$\left(\mathrm{id} - \mathcal{P}_{\hat{U}}\right)[\hat{u}^\star]\,. \tag{5.55}$$

If the discretization error is small, then it might even be advisable to use $u^\star_{\hat{U}}$ as the solution estimate, since sampling from a parametric GP is, generally speaking, less expensive than sampling from a nonparametric GP.

Under certain conditions, Gaussian process Galerkin methods reproduce the classical Petrov-Galerkin method in the posterior mean. More precisely, for $\dim \hat{U} = \dim \hat{V}$, one can choose the prior covariance function k such that the mean function of the conditional GP $\hat{u}^\star$ is equal to the Petrov-Galerkin approximation $\hat{u}^{\mathrm{PG}}$ of the solution of the PDE.

Lemma 5.1. *If $\hat{B} \in \mathbb{R}^{n\times m}$ is invertible, $k_{ff} = 0$, $k_{fu} = 0$, and $\Sigma_{\hat{c}_u} \coloneqq \mathcal{P}_{\mathbb{R}^m} k_{uu} \mathcal{P}^*_{\mathbb{R}^m} \in \mathbb{R}^{m\times m}$ is invertible, then $m_{\hat{c}^\star_u} = \hat{c}^{PG}$, $U = \hat{U} \oplus \ker \mathcal{P}_{\hat{U}}$ and the conditional mean $m_{\hat{u}^\star}$ of*

$$\hat{u}^\star = u \,\Big|\, \hat{B}\hat{c}_u - \hat{l}_f = 0 \tag{5.56}$$

admits a unique additive decomposition

$$m_{\hat{u}^\star} = \hat{u}^{PG} + \hat{u}_{\ker \mathcal{P}_{\hat{U}}} \tag{5.57}$$

with $\hat{u}^{PG} \in \hat{U}$ and $\hat{u}_{\ker \mathcal{P}_{\hat{U}}} \in \ker \mathcal{P}_{\hat{U}}$.

Proof. Note that $\hat{B}\hat{c}_u = \hat{B}\mathcal{P}_{\mathbb{R}^m}[u]$ and $\hat{l}_f = \hat{l}$ as defined in section 5.1. By corollary 1, we have

$$m_{\hat{u}^\star}(x) = m(x) + (\hat{B}\mathcal{P}_{\mathbb{R}^m})[k(x,\cdot)]^\top \left((\hat{B}\mathcal{P}_{\mathbb{R}^m}) k (\hat{B}\mathcal{P}_{\mathbb{R}^m})^*\right)^{-1} \left(\hat{l} - \hat{B}\mathcal{P}_{\mathbb{R}^m}[m]\right) \tag{5.58}$$

$$= m(x) + \mathcal{P}_{\mathbb{R}^m}[k(x,\cdot)]^\top \hat{B}^\top \left(\hat{B}\Sigma_{\hat{U}}\hat{B}^\top\right)^{-1} \hat{B}\left(\hat{B}^{-1}\hat{l} - \mathcal{P}_{\mathbb{R}^m}[m]\right) \tag{5.59}$$

$$= m(x) + \mathcal{P}_{\mathbb{R}^m}[k(x,\cdot)]^\top \Sigma^{-1}_{\hat{U}} \left(\hat{B}^{-1}\hat{l} - \mathcal{P}_{\mathbb{R}^m}[m]\right). \tag{5.60}$$

Since $\mathcal{P}_{\hat{U}}$ is a bounded projection, we have

$$U = \operatorname{ran}\left(\mathcal{P}_{\hat{U}}\right) \oplus \ker\left(\mathcal{P}_{\hat{U}}\right) \tag{5.61}$$

(Rudin 1991, Section 5.16)

$$= \hat{U} \oplus \ker\left(\mathcal{P}_{\hat{U}}\right), \tag{5.62}$$

where each $u \in U$ decomposes uniquely into $u' = u'_{\hat{U}} + (u'_{\hat{U}})^c$ with $u'_{\hat{U}} \in \hat{U}$ and $(u'_{\hat{U}})^c \in \ker\left(\mathcal{P}_{\hat{U}}\right)$. It is clear that

$$u'_{\hat{U}} = \mathcal{P}_{\hat{U}}\left[u'\right],$$

and

$$\begin{aligned}(u'_{\hat{U}})^c &= \left(\mathrm{id} - \mathcal{P}_{\hat{U}}\right)\left[u'\right] \\ &= \mathcal{P}_{\ker\left(\mathcal{P}_{\hat{U}}\right)}\left[u'\right].\end{aligned}$$

This implies

$$m_{\hat{c}^\star_u} = \mathcal{P}_{\mathbb{R}^n}[m_{\hat{u}^\star}] \tag{5.63}$$

$$= \mathcal{P}_{\mathbb{R}^m}[m] + \underbrace{\mathcal{P}_{\mathbb{R}^m} k \mathcal{P}^*_{\mathbb{R}^m}}_{=\Sigma_{\hat{U}}} \Sigma^{-1}_{\hat{U}} \left(\hat{B}^{-1}\hat{l} - \mathcal{P}_{\mathbb{R}^m}[m]\right) \tag{5.64}$$

$$= \mathcal{P}_{\mathbb{R}^m}[m] + \hat{B}^{-1}\hat{l} - \mathcal{P}_{\mathbb{R}^m}[m] \tag{5.65}$$

$$= \hat{B}^{-1}\hat{l} \tag{5.66}$$

$$= \hat{c}^{\mathrm{PG}}. \tag{5.67}$$

Hence, we have

$$\mathcal{P}_{\hat{U}}\left[m_{\hat{u}^\star}\right] = \sum_{i=1}^{m} \left(\mathcal{P}_{\mathbb{R}^n}\left[m_{\hat{u}^\star}\right]\right)_i u_i = \sum_{i=1}^{m} \hat{c}_i^{\mathrm{PG}} u_i = \hat{u}^{\mathrm{PG}} \tag{5.68}$$

and since $U = \hat{U} \oplus \ker\left(\mathcal{P}_{\hat{U}}\right)$, the statement follows. □

Corollary 5.1. *If, additionally, $m_u \in \hat{U}$ and $\mathcal{P}_{\ker\left(\mathcal{P}_{\hat{U}}\right)} k_{uu} \mathcal{P}^*_{\mathbb{R}^n} = 0$, then the conditional mean $m_{\hat{u}^\star}$ is equal to the Petrov-Galerkin solution approximation, i.e.*

$$m_{\hat{u}^\star} = \hat{u}^{PG}. \tag{5.69}$$

6. Related Work

The idea of approaching problems from numerical mathematics by statistical inference and Bayesian inference in particular is pursued in the field of *probabilistic numerics* [Hennig et al., 2015, Cockayne et al., 2019, Oates and Sullivan, 2019, Owhadi et al., 2019, Hennig et al., 2022].

Solving PDEs by means of GP inference was previously explored in several publications, many of which come from this field of research. Graepel [2003], Särkkä [2011] propose GP-based collocation methods for solving general linear operator equations and Cockayne et al. [2017], Raissi et al. [2017] develop such methods for the specific case of linear PDEs. Both Cockayne et al. [2017], Raissi et al. [2017] then use these Bayesian PDE solvers to solve inverse problems, where the former perform full Bayesian inference over the parameters, while the latter resort to maximum likelihood estimation. Leveraging a finite-element discretization, Girolami et al. [2021] propose a Bayesian method for solving forward and inverse problems, where parameters of the PDE are noisy. This method is capable of solving PDEs in weak form, but it does not account for discretization error in the uncertainty estimation for the forward problem. Wang et al. [2021], Krämer et al. [2022] develop GP-based solvers for nonlinear PDEs by leveraging finite-difference approximations to the differential operator and linearization-based approximate inference. In addition, Krämer et al. [2022] employ Gauss-Markov priors, which allows for efficient inference by Bayesian filtering and smoothing. [Owhadi, 2015] proposes a mathematical framework, which frames numerical homogenization of weak form PDEs as Bayesian inference.

Symmetric collocation [Fasshauer, 1997, 1999] and Galerkin methods [Fletcher, 1984] are important classes of classical numerical methods for approximating the solutions to PDEs, which inspire the probabilistic extensions developed in this work.

Särkkä et al. [2013] demonstrate that spatio-temporal Gaussian process inference with affine observations of the spatial part of the sample paths amounts to infinite-dimensional filtering and smoothing problems. Owhadi and Scovel [2018] show how to condition Gaussian measures on an orthogonal direct sum of separable Hilbert spaces on observations of one of the summands.

7. Conclusion

This thesis explored how Gaussian processes can be used to fuse prior and mechanistic knowledge with empirical measurements, while accounting for uncertainties in a principled way. Chapter 2 showed how Gaussian process inference can be used to approximate the strong solution to linear PDEs by probabilistic collocation, while quantifying discretization error. A crucial insight for this is the interpretation of PDEs as an indirect observation of their unknown solution. Based on the recurring practical example developed in section 2.4, chapter 3 demonstrates how uncertainties in the input data can be propagated to the solution in a principled and natural way. We also saw how our solvers can be used as a modular building block in computational pipelines, which are very elegantly described as directed graphical models "with PDE nodes". The theoretical background for chapters 2 and 3 was introduced in chapter 4. Theorem 1 and corollaries 1 and 2 show when and how Gaussian processes can be conditioned on affine observations of their sample paths. Such affine observations include integral and (partial) derivative observations. This does not only provide theoretical backing for this work, but also for many previous publications, in which this result has been used without proof. Moreover, we detailed how prior GPs should be chosen in the context of Bayesian PDE solvers. Finally, chapter 5 detailed a general framework for PDE solvers based on Gaussian processes capable of approximating both weak and strong solutions of linear PDEs. We showed that our framework can be seen as a probabilistic generalization of the most important established non-probabilistic methods for linear PDEs, including symmetric collocation, and Galerkin-type methods such as spectral methods, finite element methods and finite volume methods. Namely, our framework reproduces the approximations to the solution computed by these methods in the posterior mean.

All in all, we conclude that probabilistic solvers for linear PDEs based on Gaussian process inference are a viable alternative to non-probabilistic solvers. As opposed to non-probabilistic solvers they address uncertainties in the input data and the solution prior by integrating them into the structured output uncertainty, instead of ignoring them. This is vital in practical applications in science and engineering, where uncertain input data is ubiquitous. We have also seen that the structural output uncertainty plays nicely with computational pipelines and enables highly modular implementations. Last but not least, the fact that classical methods are recovered in the posterior mean with zero-mean uncertainty due to discretization error might serve as a starting point for integrating probabilistic solvers into existing non-probabilistic pipelines without compromising predictions.

A. Proof of Theorem 1

Throughout this work we exploited that Gaussian processes can be conditioned on observations of their paths made through a linear operator or (multiple) linear functionals, which results in another Gaussian process with closed-form expressions for its mean and covariance function. For instance, we used this capability to inform a Gaussian process prior via

1. PDE observations (observations through a differential operator),
2. integral observations, and
3. observations of projections.

For *finite-dimensional* Euclidean vector spaces, this capability can be expressed by the following theorem, which has previously been proven in the literature.

Theorem A.1 (Linear-Gaussian Inference [Bishop, 2006])**.** *Consider the linear-Gaussian model*

$$x \sim \mathcal{N}(\mu, \Sigma), \qquad \text{and} \tag{A.1}$$
$$y \mid x \sim \mathcal{N}(Ax + b, \Lambda), \tag{A.2}$$

with $A \in \mathbb{R}^{n \times d}$, and $\Sigma \in \mathbb{R}^{d \times d}$ as well as $\Lambda \in \mathbb{R}^{n \times n}$ symmetric positive semidefinite. Then

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma A^\top \\ A\Sigma & A\Sigma A^\top + \Lambda \end{pmatrix}\right), \tag{A.3}$$

and hence

$$y \sim \mathcal{N}\left(A\mu + b, A\Sigma A^\top + \Lambda\right) \tag{A.4}$$
$$x \mid y \sim \mathcal{N}(\mu_{post}, \Sigma_{post}), \tag{A.5}$$

where

$$\mu_{post} = \mu - \Sigma A^\top \left(A\Sigma A^\top + \Lambda\right)^{-1} (y - (A\mu + b)) \tag{A.6}$$
$$= (\Sigma^{-1} + A^\top \Lambda^{-1} A)^{-1} (A^\top \Lambda^{-1}(y - b) + \Sigma^{-1}\mu), \quad \text{and} \tag{A.7}$$

$$\Sigma_{post} = \Sigma - \Sigma A^\top \left(A\Sigma A^\top + \Lambda\right)^{-1} A\Sigma \tag{A.8}$$
$$= (\Sigma^{-1} + A^\top \Lambda^{-1} A)^{-1}. \tag{A.9}$$

Remark A.1. *The likelihood of the linear-Gaussian model in equation* (A.2) *is best understood as* $y = Ax + \epsilon$*, where* $\epsilon \sim \mathcal{N}(b, \Lambda)$ *and* $x \perp\!\!\!\perp \epsilon$*.*

While the GP analogue of this theorem, i.e.

$$f \sim \mathcal{GP}(m_f, k_f) \tag{A.10}$$
$$g \sim \mathcal{GP}(m_g, k_g) \tag{A.11}$$
$$f \mid \mathcal{L}[f] + g = h \sim \mathcal{GP}\left(m_{f|h}, k_{f|h}\right) \tag{A.12}$$

is well-known and widely used (see e.g. [Graepel, 2003, Rasmussen and Williams, 2006, Särkkä, 2011, Särkkä et al., 2013, Cockayne et al., 2017, Raissi et al., 2017, Agrell, 2019, Albert, 2019, Krämer et al., 2022]), to the best of our knowledge, no complete proof of the result has been published. In particular, it is not clear which assumptions on the Gaussian process and the linear operator need to be met in order for the result to hold. In the following, we formalize and prove a generalization this result, theorem 1 and its corollaries 1 and 2, which grant a theoretical basis for all previously mentioned methodology using it.

A.1. Gaussian Processes

We start by reviewing basic properties of Gaussian processes.

Definition A.1. *A* Gaussian process (GP) f *with index set* $\mathcal{X}$ *is a family* $\{f_x\}_{x \in \mathcal{X}}$ *of* $\mathbb{R}$*-valued random variables on a common Borel probability space* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$[1]*, such that, for each finite set of indices* $x_1, \ldots, x_n$*, the joint distribution of* $f_{x_1}, \ldots, f_{x_n}$ *is Gaussian. We also write* $f(x) \coloneqq f_x$ *and* $f(x, \omega) \coloneqq f_x(\omega)$*.*

Definition A.2. *Let* f *be a Gaussian process on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with index set* $\mathcal{X}$*. The function*

$$m \colon \mathcal{X} \to \mathbb{R}, x \mapsto m(x) = \mathbb{E}_P[f_x] \tag{A.13}$$

is called the mean (function) *of* f *and the function*

$$k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}, (x_1, x_2) \mapsto k(x_1, x_2) = \mathrm{Cov}_P[f_{x_1}, f_{x_2}] \tag{A.14}$$

is called the covariance function *or* kernel *of* f*. We also often write* $f \sim \mathcal{GP}(m, k)$ *if* f *is a Gaussian process with mean* m *and kernel* k*.*

We commonly use Gaussian processes to model our belief about unknown functions, which can be motivated by interpreting their *sample paths* as function-valued random variables:

Definition A.3. *Let* f *be a Gaussian process on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with index set* $\mathcal{X}$*. For each* $\omega \in \Omega$*, the function*

$$f(\cdot, \omega) \colon \mathcal{X} \to \mathbb{R}^d, x \mapsto f_x(\omega) \tag{A.15}$$

[1] $\mathcal{B}(\Omega)$ denotes the Borel σ-algebra on Ω

is called a (sample) path *of the Gaussian process. We also write* $f_\cdot(\omega) \coloneqq f(\cdot, \omega)$. *The set* $\operatorname{paths}(f) \coloneqq \{f(\cdot, \omega) \colon \omega \in \Omega\}$ *containing all sample paths of* f *is referred to as the* path space *of* f.

Lemma A.1. *Let* f *be a Gaussian process on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with index set* $\mathcal{X}$. *Consider the function*

$$f_\cdot \colon \Omega \to \operatorname{paths}(f) \subset \mathbb{R}^{\mathcal{X}}, \omega \mapsto f_\cdot(\omega). \tag{A.16}$$

If there is a σ*-algebra on* $\operatorname{paths}(f)$ *such that* $f_\cdot$ *is measurable, then* $f_\cdot$ *is a function-valued random variable with values in* $\operatorname{paths}(f)$. *In the following, we will refer to function-valued random variables as* random functions*, in analogy to the concept of a random variable.*

When working with (deterministic) functions, we are very used to being able to e.g. add, scale, take limits, differentiate and integrate these functions. The same is possible for GPs, although there are several caveats that need to be taken into account, specifically for operations such as differentiation and integration.

Lemma A.2. *Let* $f = \{f_x\}_{x \in \mathcal{X}}$ *and* $g = \{g_x\}_{x \in \mathcal{X}}$ *be independent Gaussian processes on the same probability space* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with mean functions* m_f, m_g, *and covariance functions* k_f, k_g, *respectively. Let* h *be the family of random variables on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *induced by evaluating a linear combination of the sample paths of* f *and* g *at all points* $x \in \mathcal{X}$, *i.e.*

$$h = \{\omega \mapsto (\alpha f_\cdot(\omega) + \beta g_\cdot(\omega))(x)\}_{x \in \mathcal{X}} = \{\alpha f_x + \beta g_x\}_{x \in \mathcal{X}}. \tag{A.17}$$

with coefficients $\alpha, \beta \in \mathbb{R}$. *Then* h *is a Gaussian process with mean function* $m_h \coloneqq \alpha m_f + \beta m_g$ *and covariance function*

$$k_h(x_1, x_2) \coloneqq \alpha^2 k_f(x_1, x_2) + \beta^2 k_g(x_1, x_2).$$

Moreover, the sample paths of h *are linear combinations of the sample paths of* f *and* g, *i.e.* $h(\cdot, \omega) = \alpha f(\cdot, \omega) + \beta g(\cdot, \omega)$.

Proof. The linear combination of two independent Gaussian random variables or vectors is again a Gaussian random variable, which implies that h is a Gaussian process. Moreover, $m_h = \alpha m_f + \beta m_g$ follows by the linearity of expectation. Finally, using the covariance's bilinearity and symmetry properties, we have

$$\begin{aligned}
k_h(x_1, x_2) &= \operatorname{Cov}[h(x_1), h(x_2)] && \text{(A.18)}\\
&= \operatorname{Cov}[\alpha f(x_1) + \beta g(x_1), \alpha f(x_2) + \beta g(x_2)] && \text{(A.19)}\\
&= \alpha \operatorname{Cov}[f(x_1), \alpha f(x_2) + \beta g(x_2)] + \beta \operatorname{Cov}[g(x_1), \alpha f(x_2) + \beta g(x_2)] && \text{(A.20)}\\
&= \alpha^2 \operatorname{Cov}[f(x_1), f(x_2)] + \alpha\beta \underbrace{\operatorname{Cov}[f(x_1), g(x_2)]}_{=0} && \text{(A.21)}\\
&\quad + \beta\alpha \underbrace{\operatorname{Cov}[g(x_1), f(x_2)]}_{=0} + \beta^2 \operatorname{Cov}[g(x_1), g(x_2)] && \text{(A.22)}\\
&= \alpha^2 k_f(x_1, x_2) + \beta^2 k_g(x_1, x_2). && \text{(A.23)}
\end{aligned}$$

□

Note that the proof of lemma A.2 heavily draws on the fact that addition and scalar multiplication of functions are defined pointwise, since definition A.1 is tailored to such operations. Unfortunately, this basic characterization of a GP makes it hard to reason about applying operations such as limits, differentiation and integration, since they simultaneously operate on an (uncountably) infinite subset of the random variables. However, definition A.1 only provides information about finite subsets of these random variables. This problem extends to the more general classes of *linear operators*, i.e. linear maps between vector spaces of functions, and *linear functionals*, i.e. linear maps from a vector space of functions to its field of scalars (e.g. $\mathbb{R}$). Differentiation is an example of a linear operator, while limits and (definite) integrals are linear functionals. A solution to this problem is to treat the random function f. (if it exists) as a first class object, rather than accessing it via its evaluations as in definition A.1. This will grant access to all random variables in f simultaneously. To do so, we need to

1. gain an understanding of the structure of the GP's path space $\operatorname{paths}(f)$ in order to be able to decide whether f. is a random function, i.e. measurable. This will also come in handy, when applying linear operators to the GP, since its sample paths might not lie in the domain of any given linear operator, e.g. due to insufficient differentiability. In most practically relevant cases, $\operatorname{paths}(f) \subset \mathbb{R}^{\mathcal{X}}$ is a real separable Hilbert function space.

2. analyze the law or distribution of the random function f. in order to understand the belief about the sample paths encoded in P and f.. In most practically relevant cases, this will turn out to be a Gaussian measure on a separable Hilbert space, which is essentially the infinite-dimensional analogue of a multivariate Gaussian distribution on $\mathbb{R}^d$.

Fortunately, the first point has already been extensively adressed in the literature. See Kanagawa et al. [2018, Section 4] for an overview.

Remark A.2. *Let $f \sim \mathcal{GP}(m, k)$ be a Gaussian process with index set $\mathcal{X}$ and let $\mathcal{H}_k$ be the reproducing kernel Hilbert space (RKHS) of the covariance function or kernel k. In most practically relevant cases, the sample paths of f do almost surely not lie in $\mathcal{H}_k$. Rather, with probability 1, they are elements of a larger RKHS $\mathcal{H}_{k'} \supset \mathcal{H}_k$. We refer to [Kanagawa et al., 2018, Section 4] and Steinwart [2019] for more details on sample path properties. Assumption A.1 is a common set of assumptions for this to hold. Note that, in practice, $\mathcal{X}$ is often a compact (i.e. closed and bounded) subset of $\mathbb{R}^d$, k is continuous and ν is the Lebesgue measure, which already fulfills the first part of assumption A.1.*

Assumption A.1. *Let $\mathcal{X}$ be a compact metric space, $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous positive-definite kernel, and ν a finite Borel measure whose support is $\mathcal{X}$. Denote by $(\lambda_i, \phi_i)_{i \in I}$ the eigensystem obtained by applying Mercer's theorem [Kanagawa et al., 2018, Theorem 4.1] to k. Let $\mathcal{H}_{k^\theta} \supset \mathcal{H}_k$ the θ-th power of $\mathcal{H}_k$ [Steinwart and Scovel, 2012, Definition 4.1] with $\theta \in (0, 1)$ such that $\sum_{i \in I} \lambda_i^{1-\theta} < \infty$.*

Remark A.3. *The fact that this only holds only with probability 1 is not a problem due to the fact that we virtually always describe GP priors via m and k rather than by explicitly constructing the functions f_x for $x \in \mathcal{X}$. If there is $f(\cdot,\omega) \in \operatorname{paths}(f)$ for which $f(\cdot,\omega) \notin \mathcal{H}_{k'}$, we can simply define a new probability space $(\Omega', \mathcal{B}(\Omega'), P)$ with $\Omega' \coloneqq \{\omega \in \Omega \mid f(\cdot,\omega) \notin \mathcal{H}_{k'}\} = \Omega \cap f_\cdot^{-1}(\mathcal{H}_{k'})$ on which f will have the same mean and covariance function.*

In section 4.1, we have already seen that Sobolev spaces can be obtained as path spaces of Gaussian processes with Matérn covariance functions.

A.2. Multi-output Gaussian Processes

The sample paths of Gaussian processes as defined in definition A.1 are always real-valued. However, especially in the context of PDEs, vector-valued functions are ubiquitous, e.g. when dealing with vector fields such as the electric field. Fortunately, the index set of a Gaussian process can be chosen freely, which means that we can "emulate" vector-valued GPs. More precisely, a function $f\colon \mathcal{X} \to \mathbb{R}^{d'}$ is in some sense equivalent to a function $f'\colon \{1,\dots,d'\} \times \mathcal{X} \to \mathbb{R}, (i,x) \mapsto f'(i,x) = f_i(x)$. Applying this construction to a Gaussian process leads to the following definition of a *multi-output Gaussian process*:

Definition A.4 (Multi-output Gaussian Process)**.** *A d-output Gaussian process f with index set $\mathcal{X}$ on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ is a Gaussian process with index set $\mathcal{X}' \coloneqq \{1,\dots,d\} \times \mathcal{X}$ on the same probability space. With a slight abuse of notation, we write $f_x(\omega) \coloneqq (f_{(i,x)}(\omega))_{i=1}^d \in \mathbb{R}^d$, etc. We also write the mean and covariance functions m and k of f as $m\colon \mathcal{X} \to \mathbb{R}^d$ and $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$, where*

$$m(x) = \begin{pmatrix} m(1,x) \\ \vdots \\ m(d,x) \end{pmatrix} \quad \text{and} \quad k(x_1,x_2) = \begin{pmatrix} k((1,x_1),(1,x_2)) & \dots & k((1,x_1),(d,x_2)) \\ \vdots & \ddots & \vdots \\ k((d,x_1),(1,x_2)) & \dots & k((d,x_1),(d,x_2)) \end{pmatrix}.$$

Remark A.4. *If assumption A.1 holds for some single-output GP, i.e. $(\mathcal{X}, d_\mathcal{X})$ is a compact metric space and $\nu_\mathcal{X}$ is a finite Borel measure on $\mathcal{X}$, then it also holds in the multi-output GP case. specifically, we can equip $I = \{1,\dots,d\}$ with the discrete metric*

$$d_I(i_1,i_2) = \begin{cases} 0 & \text{if } i_1 = i_2 \\ 1 & \text{if } i_1 \neq i_2 \end{cases} \tag{A.24}$$

and the Dirac measure $\nu = \delta_\cdot(I)$. Then $(I \times \mathcal{X}, d)$ with

$$d((i_1,x_1),(i_2,x_2)) \coloneqq d_I(i_1,i_2) + d_\mathcal{X}(x_1,x_2)$$

is a compact metric space and $\nu_I \otimes \nu_\mathcal{X}$ is a finite Borel measure on $I \times \mathcal{X}$.

Multi-output Gaussian processes also give us the ability to reason about linear combinations of non-independent Gaussian processes.

Lemma A.3. *Let*

$$\begin{pmatrix} f \\ g \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m_f \\ m_g \end{pmatrix}, \begin{pmatrix} k_{ff} & k_{fg} \\ k_{gf} & k_{gg} \end{pmatrix}\right) \tag{A.25}$$

be a $2n$-output Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $f, g\colon \mathcal{X} \times \Omega \to \mathbb{R}^n$, $m_f, m_g\colon \mathcal{X} \to \mathbb{R}^n$ and $k_{ff}, k_{fg}, k_{gg}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n\times n}$, where $k_{gf} = k_{fg}^\top$. Let h be the family of random variables on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ induced by evaluating a linear combination of the sample paths of f and g at all points $x \in \mathcal{X}$, i.e.

$$h = \{\omega \mapsto (\alpha f_\cdot(\omega) + \beta g_\cdot(\omega))(i,x)\}_{(i,x)\in I\times\mathcal{X}} = \{\alpha f_{i,x} + \beta g_{i,x}\}_{(i,x)\in I\times\mathcal{X}}. \tag{A.26}$$

with $I = \{1, \cdots, n\}$ and coefficients $\alpha, \beta \in \mathbb{R}$. Then h is an n-output Gaussian process with mean function $m_h := \alpha m_f + \beta m_g$ and covariance function

$$k_h := \alpha^2 k_{ff} + \alpha\beta k_{fg} + \beta\alpha k_{gf} + \beta^2 k_{gg}.$$

Moreover, the sample paths of h are linear combinations of the sample paths of f and g, i.e. $h_\cdot(\omega) = \alpha f_\cdot(\omega) + \beta g_\cdot(\omega)$.

Proof. Analogous to the proof of lemma A.2. □

A.3. Gaussian Measures on Separable Hilbert Spaces

As stated before, we need to understand the distribution of the random functions $\omega \to f(\cdot,\omega)$ and $\omega \to \mathcal{L}[f(\cdot,\omega)]$ in order to use observations of GP sample paths through a linear operator $\mathcal{L}$ in inference. We will do so by analyzing the pushforward measure $\mu := P \circ f_\cdot^{-1}$. In many cases, this measure will turn out to be Gaussian probability measures on the (usually) infinite-dimensional Hilbert function space $\mathcal{H} := \mathrm{paths}(f)$ of sample paths (see proposition A.1 and lemma A.6).

Definition A.5 (Maniglia and Rhandi 2004, Definition 1.2.2)**.** *Let $\mathcal{H}$ be a real separable Hilbert space. A probability measure μ on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is called* Gaussian *if each $h^* \in \mathcal{H}^*$, i.e. each $\langle h, \cdot\rangle_\mathcal{H}$ with $h \in \mathcal{H}$, is a univariate Gaussian random variable. An $\mathcal{H}$-valued random variable is called Gaussian if its law is Gaussian.*

Just as for probability measures on Euclidean vector space $\mathbb{R}^n$, we can define a mean and covariance (operator) for this more general class of probability measures.

Definition A.6 (Maniglia and Rhandi 2004, Definition 1.2.1)**.** *Let μ be a Borel probability measure on a real separable Hilbert space $\mathcal{H}$. If the function $\langle h, \cdot\rangle_\mathcal{H}$ is μ-integrable for all $h \in \mathcal{H}$, and there is an $m \in \mathcal{H}$ such that*

$$\langle h, m\rangle_\mathcal{H} = \int_\mathcal{H} \langle h, h'\rangle_\mathcal{H} \, \mathrm{d}\mu(h') = \mathbb{E}_{h'\sim\mu}\left[\langle h, h'\rangle_\mathcal{H}\right] \tag{A.27}$$

for all $h \in \mathcal{H}$, then m is call the mean (vector) *of μ. If furthermore there is a positive symmetric linear operator $\mathcal{C}\colon \mathcal{H} \to \mathcal{H}$ such that*

$$\langle h_1, \mathcal{C}[h_2]\rangle_\mathcal{H} = \int_\mathcal{H} \langle h_1, h' - m\rangle_\mathcal{H} \langle h_2, h' - m\rangle_\mathcal{H} \, \mathrm{d}\mu(h') \tag{A.28}$$

$$= \operatorname{Cov}_{h' \sim \mu} \left[\left\langle h_1, h' \right\rangle_{\mathcal{H}}, \left\langle h_2, h' \right\rangle_{\mathcal{H}} \right] \tag{A.29}$$

for all $h_1, h_2 \in \mathcal{H}$*, then* $\mathcal{H}$ *is called the* covariance operator *of* μ*.*

Remark A.5. *The mean and the covariance operator of a Gaussian measure on a separable Hilbert space always exist and they identify the measure uniquely [Maniglia and Rhandi, 2004, Theorem 1.2.5]. Hence, we also often write* $\mathcal{N}(m, \mathcal{C})$ *to denote Gaussian measures on separable Hilbert spaces.*

Using the notion of a *Bochner integral* [Yosida, 1995, section V.5], we can also give an equivalent definition of the mean and covariance operator, which is more similar to the finite-dimensional counterpart.

Lemma A.4. *Let* $\mu = \mathcal{N}(m, \mathcal{C})$ *be a Gaussian measure on a real separable Hilbert space* $\mathcal{H}$*. Then the identity* $\mathrm{id}_{\mathcal{H}}$ *is Bochner* μ*-integrable and the mean* m *of* μ *is given by the following Bochner integral*

$$m = \int_{\mathcal{H}} h \, \mathrm{d}\mu(h). \tag{A.30}$$

Moreover, the function $h' \mapsto \langle h, h' - m \rangle_{\mathcal{H}} (h' - m)$ *is Bochner* μ*-integrable for any* $h \in \mathcal{H}$ *and the covariance operator* $\mathcal{C}$ *of* μ *is defined by*

$$\mathcal{C}[h] := \int_{\mathcal{H}} \left\langle h, h' - m \right\rangle_{\mathcal{H}} (h' - m) \, \mathrm{d}\mu\left(h'\right). \tag{A.31}$$

To prove lemma A.4, we will need the following theorem about the properties of Bochner integrals.

Theorem A.2 (Yosida [1995], Section V.5, Theorem 1 and Corollary 2)**.** *Let* $(\Omega, \mathcal{B}(\Omega), \mu)$ *be a measure space,* $(V, \|\cdot\|_V)$ *a Banach space and* $f \colon \Omega \to V$ *a strongly* $\mathcal{B}(\Omega)$*-measurable function. Let* $\mathcal{L} \colon V \to U$ *be a bounded linear operator with values in a Banach space* $(U, \|\cdot\|_U)$*.*

1. *f is Bochner μ-integrable if and only if $\|f(\cdot)\|_V$ is μ-integrable.*
2. *If f is Bochner μ-integrable, then $\mathcal{L}[u]$ is Bochner μ-integrable, and*

$$\int_B \mathcal{L}[u](\omega) \, \mathrm{d}\mu(\omega) = \mathcal{L}\left[\int_B u(\omega) \, \mathrm{d}\mu(\omega) \right] \tag{A.32}$$

for $B \in \mathcal{B}(\Omega)$*.*

Proof of Lemma A.4. By Maniglia and Rhandi [2004, Theorem 1.2.5], we have that

$$\int_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 \, \mathrm{d}\mu(h) < \infty. \tag{A.33}$$

This implies that $\|\cdot\|_{\mathcal{H}} \in L_2(\mathcal{H}, \mathcal{B}(\mathcal{H}), \mu)$. Hence,

$$\int_{\mathcal{H}} \|h\|_{\mathcal{H}} \, \mathrm{d}\mu(h) = \int_{\mathcal{H}} 1 \cdot \|h\|_{\mathcal{H}} \, \mathrm{d}\mu(h) \tag{A.34}$$

$$\leq \sqrt{\int_{\mathcal{H}} 1 \,\mathrm{d}\mu\,(h)} \cdot \sqrt{\int_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 \,\mathrm{d}\mu\,(h)} \tag{A.35}$$

(Cauchy-Schwarz inequality in $L_2(\mathcal{H}, \mathcal{B}(\mathcal{H}), \mu)$)

$$= \sqrt{\int_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 \,\mathrm{d}\mu\,(h)} \tag{A.36}$$

(μ is a probability measure)

$$< \infty. \tag{A.37}$$

By theorem A.2, it follows that $\mathrm{id}\colon \mathcal{H} \to \mathcal{H}$ is Bochner μ-integrable and that

$$\langle m, h\rangle_{\mathcal{H}} \coloneqq \int_{\mathcal{H}} \langle h, h'\rangle_{\mathcal{H}} \,\mathrm{d}\mu\,(h') = \left\langle h, \int_{\mathcal{H}} h' \,\mathrm{d}\mu\,(h') \right\rangle_{\mathcal{H}} \tag{A.38}$$

for $h \in \mathcal{H}$, since $\langle h, \cdot\rangle_{\mathcal{H}}$ is a continuous linear functional. Moreover, for $h \in \mathcal{H}$ we have

$$\int_{\mathcal{H}} \left\|\langle h, h' - m\rangle_{\mathcal{H}} (h' - m)\right\|_{\mathcal{H}} \,\mathrm{d}\mu\,(h') \tag{A.39}$$

$$= \int_{\mathcal{H}} \left|\langle h, h' - m\rangle_{\mathcal{H}}\right| \left\|h' - m\right\|_{\mathcal{H}} \,\mathrm{d}\mu\,(h') \tag{A.40}$$

$$\leq \|h\|_{\mathcal{H}} \int_{\mathcal{H}} \left\|h' - m\right\|_{\mathcal{H}}^2 \,\mathrm{d}\mu\,(h') \tag{A.41}$$

(Cauchy-Schwarz inequality in $\mathcal{H}$)

$$\leq \|h\|_{\mathcal{H}} \left(\int_{\mathcal{H}} \left\|h'\right\|_{\mathcal{H}}^2 \,\mathrm{d}\mu\,(h') + 2\|m\|_{\mathcal{H}} \int_{\mathcal{H}} \left\|h'\right\|_{\mathcal{H}} \,\mathrm{d}\mu\,(h') + \|m\|_{\mathcal{H}}^2 \int_{\mathcal{H}} 1 \,\mathrm{d}\mu\,(h') \right) \tag{A.42}$$

(Triangle inequality in $\mathcal{H}$)

$$= \|h\|_{\mathcal{H}} \int_{\mathcal{H}} \left\|h'\right\|_{\mathcal{H}}^2 \,\mathrm{d}\mu\,(h') + 2\|h\|_{\mathcal{H}} \|m\|_{\mathcal{H}} \int_{\mathcal{H}} \left\|h'\right\|_{\mathcal{H}} \,\mathrm{d}\mu\,(h') + \|h\|_{\mathcal{H}} \|m\|_{\mathcal{H}}^2 \tag{A.43}$$

(μ is a probability measure)

$$< \infty, \tag{A.44}$$

and hence, again by theorem A.2, the function $h' \mapsto \langle h, h' - m\rangle_{\mathcal{H}} (h' - m)$ is Bochner μ-integrable for any $h \in \mathcal{H}$. This means that

$$\langle h_1, \mathcal{C}[h_2]\rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle h_1, h' - m\rangle_{\mathcal{H}} \langle h_2, h' - m\rangle_{\mathcal{H}} \,\mathrm{d}\mu\,(h') \tag{A.45}$$

$$= \left\langle h_1, \int_{\mathcal{H}} \langle h_2, h' - m\rangle_{\mathcal{H}} (h' - m) \,\mathrm{d}\mu\,(h') \right\rangle_{\mathcal{H}} \tag{A.46}$$

for any $h_1, h_2 \in \mathcal{H}$, where we used the fact that $\langle h_1, \cdot\rangle_{\mathcal{H}}$ is a continuous linear functional for any $h_1 \in \mathcal{H}$ to invoke theorem A.2. □

A.4. Gaussian Measures on the Path Spaces of Gaussian Processes

We now have the tools to analyze the probability measure induced by the GP over its sample paths via the function-valued random variable $f_\cdot$.

Assumption A.2. *Let $f \sim \mathcal{GP}(m, k)$ be a Gaussian process with index set $\mathcal{X}$ on a Borel probability space $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$, whose mean and sample paths lie in a real separable Hilbert function space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with $\mathcal{H}_k \subset \mathcal{H}$, on which all point evaluation functionals are continuous, i.e. $m \in \mathcal{H}$ and $\mathrm{paths}(f) \subset \mathcal{H}$.*

Proposition A.1. *Let assumption A.2 hold. Then $\omega \to f(\cdot, \omega)$ is an $\mathcal{H}$-valued Gaussian random variable with mean m and covariance operator*

$$\mathcal{C}_k \colon \mathcal{H} \to \mathcal{H}, h \mapsto \mathcal{C}_k[h](x) = \langle k(x, \cdot), h \rangle_{\mathcal{H}}. \tag{A.47}$$

Proof. By definition, $f(x, \cdot)$ is a Gaussian random variable for every $x \in \mathcal{X}$. Hence, corollary 12 in [Berlinet and Thomas-Agnan, 2004, Chapter 4, Section 2, p.195] ensures that $f_\cdot \colon \Omega \to \mathcal{H}, \omega \mapsto f(\cdot, \omega)$ is Borel measurable, i.e. a random variable, and by theorem 91 in [Berlinet and Thomas-Agnan, 2004, Chapter 4, Section 3.1, p.196], its law μ is a Gaussian measure on $\mathcal{H}$.

Since μ is Gaussian and $\mathcal{H}$ is separable, by lemma A.4, it remains to show that m and $\mathcal{C}_k$ fulfill

$$m = \int_{\mathcal{H}} h \, \mathrm{d}\mu(h), \qquad \text{and} \tag{A.48}$$

$$\mathcal{C}_k[h] = \int_{\mathcal{H}} \langle h, h' - m \rangle_{\mathcal{H}} (h' - m) \, \mathrm{d}\mu(h') \tag{A.49}$$

for all $h \in \mathcal{H}$, which are both well-defined Bochner integrals. Consequently, for $x \in \mathcal{X}$, we find that

$$\begin{aligned} m(x) &= \mathbb{E}_{\mathrm{P}}[f_x] && \text{(A.50)} \\ &= \int_\Omega f(x, \omega) \, \mathrm{dP}(\omega) && \text{(A.51)} \\ &= \int_\Omega \delta_x [f(\cdot, \omega)] \, \mathrm{dP}(\omega) && \text{(A.52)} \\ &= \int_{\mathcal{H}} \delta_x[h] \, \mathrm{d}\mu(h) && \text{(A.53)} \\ &= \delta_x \left[\int_{\mathcal{H}} h \, \mathrm{d}\mu(h) \right], && \text{(A.54)} \end{aligned}$$

where the last equation holds by theorem A.2, since δ_x is continuous. Hence, by lemma A.4, $m \in \mathcal{H}$ is the mean of μ. Since δ_x is continuous for all $x \in \mathcal{X}$, by Riesz' representation

theorem [Yosida, 1995, Section III.6], there is $\delta_x^* \in \mathcal{H}$ such that $h(x) = \delta_x[h] = \langle \delta_x^*, h \rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$. It follows that, for $x_1, x_2 \in \mathcal{X}$, we have

$$
\begin{aligned}
k(x_1, x_2) &\coloneqq \operatorname{Cov}[f_{x_1}, f_{x_2}] && \text{(A.55)}\\
&= \int_\Omega (f(x_1, \omega) - m(x_1))(f(x_2, \omega) - m(x_2)) \,\mathrm{dP}(\omega) && \text{(A.56)}\\
&= \int_\Omega \left\langle \delta_{x_1}^*, f(\cdot, \omega) - m \right\rangle_{\mathcal{H}} \delta_{x_2}[f(\cdot, \omega) - m] \,\mathrm{dP}(\omega) && \text{(A.57)}\\
&= \int_{\mathcal{H}} \left\langle \delta_{x_1}^*, h' - m \right\rangle_{\mathcal{H}} \delta_{x_2}\left[h' - m\right] \,\mathrm{d}\mu\left(h'\right) && \text{(A.58)}\\
&= \delta_{x_2}\left[\int_{\mathcal{H}} \left\langle \delta_{x_1}^*, h' - m \right\rangle_{\mathcal{H}} (h' - m) \,\mathrm{d}\mu\left(h'\right)\right] && \text{(A.59)}
\end{aligned}
$$

where, again, the last equality holds due to theorem A.2, and hence

$$
k(x_1, \cdot) = \int_{\mathcal{H}} \left\langle \delta_{x_1}^*, h' - m \right\rangle_{\mathcal{H}} (h' - m) \,\mathrm{d}\mu\left(h'\right). \tag{A.60}
$$

For $h \in \mathcal{H}$ it follows that

$$
\begin{aligned}
\mathcal{C}_k[h](x) &\coloneqq \langle k(x, \cdot), h \rangle_{\mathcal{H}} && \text{(A.61)}\\
&= \left\langle h, \int_{\mathcal{H}} \left\langle \delta_x^*, h' - m \right\rangle_{\mathcal{H}} (h' - m) \,\mathrm{d}\mu\left(h'\right) \right\rangle_{\mathcal{H}} && \text{(A.62)}\\
&= \int_{\mathcal{H}} \left\langle \delta_x^*, h' - m \right\rangle_{\mathcal{H}} \left\langle h, h' - m \right\rangle_{\mathcal{H}} \,\mathrm{d}\mu\left(h'\right) && \text{(A.63)}\\
&\qquad\qquad \text{(by theorem A.2, since } \langle h, \cdot \rangle_{\mathcal{H}} \text{ is bounded)}\\
&= \int_{\mathcal{H}} \left\langle h, h' - m \right\rangle_{\mathcal{H}} \delta_x\left[h' - m\right] \,\mathrm{d}\mu\left(h'\right) && \text{(A.64)}\\
&\qquad\qquad \text{(reproducing property)}\\
&= \delta_x\left[\int_{\mathcal{H}} \left\langle h, h' - m \right\rangle_{\mathcal{H}} (h' - m) \,\mathrm{d}\mu\left(h'\right)\right], && \text{(A.65)}
\end{aligned}
$$

where the last equation holds by theorem A.2. This shows that $\mathcal{C}_k$ is indeed the covariance operator of μ. □

The correspondence from proposition A.1 also holds in reverse in the sense that, a Gaussian random variable h with values in a separable Hilbert space $\mathcal{H}$, and a set $\mathcal{X}^* \subset \mathcal{H}^*$ of continuous linear functionals on $\mathcal{H}$ induce a Gaussian process on the same probability space as f, whose paths are given by $(x^*, \omega) \to x^*[h(\omega)]$. Recall that, by the Riesz representation theorem [Yosida, 1995, Section III.6], every continuous linear functional on a Hilbert space $\mathcal{H}$, i.e. every element of the dual $\mathcal{H}^*$ of $\mathcal{H}$, can be represented as an inner product, i.e. for all $h^* \in \mathcal{H}^*$, there is $h \in \mathcal{H}$ such that $h^* = \langle h, \cdot \rangle_{\mathcal{H}}$.

Lemma A.5. *Let $h \sim \mathcal{N}(m, \mathcal{C})$ be a Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in a real separable Hilbert space $\mathcal{H}$. For every set $\mathcal{X} \subset \mathcal{H}$, the family*

$$f \coloneqq \{\omega \mapsto \langle x, h(\omega)\rangle_{\mathcal{H}}\}_{x \in \mathcal{X}} \tag{A.66}$$

is a Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with mean function $m_f \colon \mathcal{X} \to \mathbb{R}, x \mapsto \langle x, m\rangle_{\mathcal{H}}$ and covariance function

$$k_f(x_1, x_2) = \langle x_1, \mathcal{C}[x_2]\rangle_{\mathcal{H}}. \tag{A.67}$$

Proof. Since every $\langle x, \cdot\rangle_{\mathcal{H}} \in h$ is continuous, $\langle x, h(\omega)\rangle_{\mathcal{H}}$ is by definition a Gaussian random variable. Let $X = \{x_i\}_{i=1}^n \subset \mathcal{X}$ and $f_X \colon \Omega \to \mathbb{R}^n, \omega \mapsto (f_X(\omega))_i \coloneqq \langle x_i, h(\omega)\rangle_{\mathcal{H}}$. Then f_X is continuous and hence Borel measurable, since all norms on $\mathbb{R}^n$ are equivalent,

$$\|f_X(\omega)\|_2^2 = \sum_{i=1}^n \langle x_i, h(\omega)\rangle_{\mathcal{H}}^2 \leq \|h(\omega)\|_{\mathcal{H}}^2 \sum_{i=1}^n \|x_i\|_{\mathcal{H}}^2 < \infty. \tag{A.68}$$

for all $\omega \in \Omega$, and h is continuous. Moreover, $\langle v, f_X\rangle$ for $v \in \mathbb{R}^n$ is bounded and hence f_X is a Gaussian random variable on $\mathbb{R}^n$. All in all, it follows that f is a Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$. Finally, we have

$$\begin{aligned} m_f(x) &= \mathbb{E}_{\mathrm{P}}[f_x] && \text{(A.69)}\\ &= \int_\Omega f_x(\omega)\,\mathrm{dP}(\omega) && \text{(A.70)}\\ &= \int_\Omega \langle x, h(\omega)\rangle_{\mathcal{H}}\,\mathrm{dP}(\omega) && \text{(A.71)}\\ &= \langle x, m\rangle_{\mathcal{H}}. && \text{(A.72)} \end{aligned}$$

by equation (A.27) and

$$\begin{aligned} k_f(x_1, x_2) &= \mathrm{Cov}_{\mathrm{P}}[f_{x_1}, f_{x_2}] && \text{(A.73)}\\ &= \int_\Omega \langle x_1, h(\omega) - m\rangle_{\mathcal{H}} \langle x_2, h(\omega) - m\rangle_{\mathcal{H}}\,\mathrm{dP}(\omega) && \text{(A.74)}\\ &= \langle x_1, \mathcal{C}[x_2]\rangle_{\mathcal{H}}. && \text{(A.75)} \end{aligned}$$

by equation (A.28). All in all, we showed that $f \sim \mathcal{GP}(m_f, k_f)$. □

Note that, in general, the Gaussian processes resulting from lemma A.5 are different from the ones "entering" proposition A.1. The Hilbert space $\mathcal{H}$ does not have to be a function space, which means that the GP's sample paths can not lie in $\mathcal{H}$. Even if $\mathcal{H}$ is a function space, point evaluation might not be continuous or even defined on it, in which case it is also not possible to construct a GP whose sample paths lie in $\mathcal{H}$. Fortunately, if $\mathcal{H}$ is a function space on which all point evaluation functionals are defined, then this is possible:

Corollary A.1. *Let $h \sim \mathcal{N}(m, \mathcal{C})$ be a Gaussian measure on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in a real separable Hilbert function space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, on which all point evaluation functionals δ_x for $x \in \mathcal{X}$ are continuous. Then the family $f \coloneqq \{\omega \mapsto h(\omega)(x)\}_{x \in \mathcal{X}}$ is a Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with mean function m and covariance function*

$$k_f(x_1, x_2) = \mathcal{C}\left[\delta_{x_2}^*\right](x_1). \tag{A.76}$$

Moreover, the paths of f lie in $\mathcal{H}$ such that $f(\cdot, \omega) = h(\omega)$ for $\omega \in \Omega$.

A.5. Gaussian Processes are Closed Under Continuous Linear Transformations

We now have all the ingredients to analyze what happens if we apply linear operators such as differentiation or integration to the paths of a (multi-output) GP. It turns out that, for bounded linear operators, the resulting object is again a Gaussian process. To prove this result, we take advantage of the "equivalence" between Gaussian processes and Gaussian measures on the Hilbert space of sample paths. More precisely, we will first use proposition A.1 to convert a given GP into a function-valued Gaussian random variable whose values are the paths of the GP. This object is very amenable to applying the linear operator path-wise. The resulting random variable will also turn out to be a Gaussian random variable. If the linear operator maps into a Hilbert space of functions on which point evaluation is continuous, then we can convert the result of the previous operation back into a GP via corollary A.1.

We start by showing that applying a bounded linear operator to a Gaussian random variable yields another Gaussian random variable and compute its moments.

Lemma A.6. *Let $\mathcal{L}\colon \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator between real separable Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. Let $f \sim \mathcal{N}(m, \mathcal{C})$ be an $\mathcal{H}_1$-valued Gaussian random variable. Then $\mathcal{L}[f] \sim \mathcal{N}(\mathcal{L}[m], \mathcal{L}\mathcal{C}\mathcal{L}^*)$.*

Proof. First of all, $\mathcal{L}$ is bounded and hence continuous, which means that $\omega \mapsto \mathcal{L}[f(\omega)]$ is Borel measurable, i.e. a random variable. Let $g \coloneqq \mathcal{L}[f]$. Let μ_f be the law of f. Then the law μ_g of g is given by the push-forward of μ_f through $\mathcal{L}$, i.e. $\mu_g = \mu_f \circ \mathcal{L}^{-1}$. We need to show that μ_g is Gaussian, i.e. that $\mu_g \circ (h_2^*)^{-1}$ for any continuous linear functional $h_2^* \in \mathcal{H}_2^*$ is a Gaussian measure on $\mathbb{R}$ (by definition A.5). We have $\mu_g \circ (h_2^*)^{-1} = \mu_f \circ \mathcal{L}^{-1} \circ (h_2^*)^{-1} = \mu_f \circ (h_2^* \circ \mathcal{L})^{-1} = \mu_f \circ (h_1^*)^{-1}$, where $h_1^* \coloneqq h_2^* \circ \mathcal{L} \in \mathcal{H}_1^*$ (i.e. h_1^* is continuous), because $\mathcal{L}$ is bounded and hence continuous. Since μ_f is Gaussian, it follows by definition A.5 that $\mu_g \circ (h_2^*)^{-1} = \mu_f \circ (h_1^*)^{-1}$ is a Gaussian measure on $\mathbb{R}$. Consequently, μ_g is Gaussian.

Let μ_g and $\mathcal{C}_g$ be the mean and covariance operator of μ_g, respectively. We have

$$m_g = \int_{\mathcal{H}_2} h_2 \, \mathrm{d}\mu_g(h_2) \tag{A.77}$$

(by lemma A.4)

$$= \int_{\mathcal{H}_1} \mathcal{L}\left[h_1\right] \mathrm{d}\mu_f\left(h_1\right) \tag{A.78}$$

$$= \mathcal{L}\left[\int_{\mathcal{H}_1} h_1 \,\mathrm{d}\mu_f\left(h_1\right)\right] \tag{A.79}$$

(by theorem A.2, since $\mathcal{L}$ is bounded)

$$= \mathcal{L}\left[m\right], \tag{A.80}$$

where the last step also follows from lemma A.4. Moreover, for $h_2 \in \mathcal{H}_2$ it holds that

$$\mathcal{C}_g\left[h_2\right] = \int_{\mathcal{H}_2} \left\langle h_2, h_2' - m_g\right\rangle_{\mathcal{H}_2} \left(h_2' - m_g\right) \mathrm{d}\mu_g\left(h_2'\right) \tag{A.81}$$

(by lemma A.4)

$$= \int_{\mathcal{H}_1} \left\langle h_2, \mathcal{L}\left[h_1\right] - m_g\right\rangle_{\mathcal{H}_2} \left(\mathcal{L}\left[h_1\right] - m_g\right) \mathrm{d}\mu_f\left(h_1\right) \tag{A.82}$$

$$= \int_{\mathcal{H}_1} \left\langle h_2, \mathcal{L}\left[h_1 - m\right]\right\rangle_{\mathcal{H}_2} \mathcal{L}\left[h_1 - m\right] \mathrm{d}\mu_f\left(h_1\right) \tag{A.83}$$

$$= \int_{\mathcal{H}_1} \left\langle \mathcal{L}^*\left[h_2\right], h_1 - m\right\rangle_{\mathcal{H}_1} \mathcal{L}\left[h_1 - m\right] \mathrm{d}\mu_f\left(h_1\right) \tag{A.84}$$

$$= \mathcal{L}\left[\int_{\mathcal{H}_1} \left\langle \mathcal{L}^*\left[h_2\right], h_1 - m\right\rangle_{\mathcal{H}_1} \left(h_1 - m\right) \mathrm{d}\mu_f\left(h_1\right)\right] \tag{A.85}$$

(by theorem A.2, since $\mathcal{L}$ is bounded)

$$= \mathcal{L}\left[\mathcal{C}\left[\mathcal{L}^*\left[h_2\right]\right]\right] \tag{A.86}$$

(by lemma A.4)

$$= \mathcal{L}\mathcal{C}\mathcal{L}^*\left[h_2\right]. \tag{A.87}$$

This implies $\mathcal{C}_g = \mathcal{L}\mathcal{C}\mathcal{L}^*$. □

Next, we use proposition A.1 and lemmas A.5 and A.6 to show that, under certain conditions, applying a bounded linear operator to the paths of a Gaussian process induces another Gaussian process. To do so, we need to compute

$$\left(\mathcal{L}\mathcal{C}\mathcal{L}^*\right)\left[\delta_{x_2}^*\right]\left(x_1\right) = \left(\delta_{x_1} \circ \mathcal{L}\right)\mathcal{C}\left(\delta_{x_2} \circ \mathcal{L}\right)^*, \tag{A.88}$$

where $\mathcal{C}$ is the covariance operator associated with a Gaussian process (see proposition A.1). Recall notation 1 used to formulate corollary 1:

Notation 1. *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite kernel and let $\mathcal{L}_i\colon \mathcal{H}_k \to \mathbb{R}^{n_i}$ for $i = 1, 2$ be bounded linear operators. We define the functions*

$$\mathcal{L}_1 k\colon \mathcal{X} \to \mathbb{R}^{n_1}, x \mapsto \mathcal{L}_1\left[k(\cdot, x)\right], \tag{4.16}$$

$$k\mathcal{L}_2^*\colon \mathcal{X} \to \mathbb{R}^{n_2}, x \mapsto \mathcal{L}_2\left[k(x, \cdot)\right], \tag{4.17}$$

$$\tag{4.18}$$

and the matrix[2] $\mathcal{L}_1 k \mathcal{L}_2^* \in \mathbb{R}^{n_1 \times n_2}$ *with entries*

$$(\mathcal{L}_1 k \mathcal{L}_2^*)_{ij} \coloneqq \mathcal{L}_2 \left[(\mathcal{L}_1 k)_i\right]_j \tag{4.19}$$

$$= \mathcal{L}_1 \left[(k \mathcal{L}_2^*)_j\right]_i . \tag{4.20}$$

Lemma A.7. *Let* $\mathcal{L}_1 \colon \mathcal{H} \to \mathbb{R}^{n_1}$ *and* $\mathcal{L}_2 \colon \mathcal{H} \to \mathbb{R}^{n_2}$ *be bounded linear operators on a real separable Hilbert space* $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ *of real-valued functions defined on an arbitrary set* $\mathcal{X}$ *with continuous point evaluation functionals. Let*

$$\mathcal{K} \colon \mathcal{H} \to \mathcal{H}, h \mapsto \mathcal{K}[h](x) = \langle k(x, \cdot), h \rangle_{\mathcal{H}} , \tag{A.89}$$

be a self-adjoint operator on $\mathcal{H}$ *with symmetric kernel* $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that* $k(x, \cdot) \in \mathcal{H}$ *for all* $x \in \mathcal{X}$*. Then*

(i) we have

$$(\mathcal{L}_1 \mathcal{K})[h]_i = \langle (\mathcal{L}_1 k)_i, h \rangle_{\mathcal{H}} \tag{A.90}$$

for all $h \in \mathcal{H}$*, and*

(ii) $\mathcal{L}_1 \mathcal{K} \mathcal{L}_2^* \in \mathbb{R}^{n_1 \times n_2}$ *with*

$$(\mathcal{L}_1 \mathcal{K} \mathcal{L}_2^*)_{ij} = \mathcal{L}_2 \left[(\mathcal{L}_1 k)_i\right]_j \tag{A.91}$$

$$= \mathcal{L}_1 \left[(k \mathcal{L}_2^*)_j\right]_i \tag{A.92}$$

$$= (\mathcal{L}_1 k \mathcal{L}_2^*)_{ij} . \tag{A.93}$$

If $\mathcal{L} \coloneqq \mathcal{L}_1 = \mathcal{L}_2$*, then* $\mathcal{L} \mathcal{K} \mathcal{L}^*$ *is symmetric, and, if* $\mathcal{K}$ *is additionally positive-(semi)definite, then* $\mathcal{L} \mathcal{K} \mathcal{L}^*$ *is positive-(semi)definite.*

Proof.

- (A.90): $\mathcal{L}_1 [\cdot]_i$ is a bounded linear functional and hence, by the Riesz representation theorem [Yosida, 1995, Section III.6], there is $h_{\mathcal{L}_1, i} \in \mathcal{H}$ such that $\mathcal{L}_1 [h_1]_i = \langle h_{\mathcal{L}_1, i}, h \rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$. It follows that

$$\mathcal{L}_1 \mathcal{K}[h](x_1) = \mathcal{L}_1 [\mathcal{K}[h]](x_1) \tag{A.94}$$

$$= \langle h_{\mathcal{L}_1, i}, \mathcal{K}[h] \rangle_{\mathcal{H}} \tag{A.95}$$

$$= \langle \mathcal{K}[h_{\mathcal{L}_1, i}], h \rangle_{\mathcal{H}} , \tag{A.96}$$

  for all $h \in \mathcal{H}$, since $\mathcal{K}$ is self-adjoint, and

$$\mathcal{K}[h_{\mathcal{L}_1, i}](x) = \langle k(x, \cdot), h_{\mathcal{L}_1, i} \rangle_{\mathcal{H}} \tag{A.97}$$

$$= \mathcal{L}_1 [k(x, \cdot)]_i \tag{A.98}$$

[2] This notation is motivated by lemma A.7, which also shows that the two different ways to compute the entries of $\mathcal{L}_1 k \mathcal{L}_2^*$ are consistent.

$$= (\mathcal{L}_1 k)_i(x) \tag{A.99}$$

for all $x \in \mathcal{X}$ and hence

$$\mathcal{L}_1 \mathcal{K}[h]_i = \langle (\mathcal{L}_1 k)_i, h \rangle_{\mathcal{H}} \tag{A.100}$$

for all $h \in \mathcal{H}$.

- (A.91): For $e_j \in \mathbb{R}^{n_2}$ with $e_{j,i} = \delta_{ji}$, we have

$$\begin{aligned}
(\mathcal{L}_1 \mathcal{K} \mathcal{L}_2^*)_{ij} &= \mathcal{L}_1 \mathcal{K} \mathcal{L}_2^* [e_j]_i && \text{(A.101)} \\
&= \mathcal{L}_1 \mathcal{K} [\mathcal{L}_2^* [e_j]]_i && \text{(A.102)} \\
&= \langle (\mathcal{L}_1 k)_i, \mathcal{L}_2^* [e_j] \rangle_{\mathcal{H}} && \text{(A.103)} \\
&= \langle \mathcal{L}_2 [(\mathcal{L}_1 k)_i], e_j \rangle && \text{(A.104)} \\
&= \mathcal{L}_2 [(\mathcal{L}_1 k)_i]_j . && \text{(A.105)}
\end{aligned}$$

- (A.92):

$$\begin{aligned}
(\mathcal{L}_1 \mathcal{K} \mathcal{L}_2^*)_{ij} &= ((\mathcal{L}_1 \mathcal{K} \mathcal{L}_2^*)^\top)_{ji} && \text{(A.106)} \\
&= (\mathcal{L}_2 \mathcal{K} \mathcal{L}_1^*)_{ji} && \text{(A.107)} \\
& && (\mathcal{K} \text{ is self-adjoint}) \\
&= \mathcal{L}_1 [(\mathcal{L}_2 k)_j]_i , && \text{(A.108)}
\end{aligned}$$

  where the last equation follows from equation (A.91) with the roles of $\mathcal{L}_1$ and $\mathcal{L}_2$ reversed.

- $(\mathcal{L}\mathcal{K}\mathcal{L}^*)^\top = (\mathcal{L}^*)^* \mathcal{K}^* \mathcal{L}^* = \mathcal{L}\mathcal{K}\mathcal{L}^*$, since $\mathcal{K}$ is self-adjoint.

- $\langle x, \mathcal{L}\mathcal{K}\mathcal{L}^* x \rangle = \langle \mathcal{L}^*[x], \mathcal{K}[\mathcal{L}^*[x]] \rangle_{\mathcal{H}} \geq 0$, since $\mathcal{K}$ is positive-semidefinite, where the inequality is strict if $\mathcal{K}$ is (strictly) positive-definite.

□

Proposition A.2. *Let assumption A.2 hold and let $\mathcal{L}\colon \mathcal{H} \to \mathbb{R}^n$ be a bounded linear operator. Then*

$$\mathcal{L}[f] : \Omega \to \mathbb{R}^n, \omega \mapsto \mathcal{L}[f(\cdot, \omega)] \tag{A.109}$$

is a Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in $\mathbb{R}^n$ and

$$\mathcal{L}[f] \sim \mathcal{N}(\mathcal{L}[m], \mathcal{L}k\mathcal{L}^*). \tag{A.110}$$

Proof. By proposition A.1 we know that $\omega \mapsto f(\cdot, \omega)$ is an $\mathcal{H}$-valued Gaussian random variable with mean m and covariance operator $\mathcal{C}[h] = \langle k(x, \cdot), h \rangle_{\mathcal{H}}$. By lemma A.6, $\omega \mapsto \mathcal{L}[f(\cdot, \omega)]$ is an $\mathbb{R}^n$-valued Gaussian random variable with mean $\mathcal{L}[m]$ and covariance matrix $\mathcal{L}\mathcal{C}\mathcal{L}^*$. Finally, $\mathcal{L}\mathcal{C}\mathcal{L}^* = \mathcal{L}k\mathcal{L}^*$ by lemma A.7. □

Corollary A.2. *Let assumption A.2 hold and let $\mathcal{L}\colon \mathcal{H} \to \mathcal{H}_{\mathcal{L}}$ be a bounded linear operator mapping into another real separable Hilbert function space $\mathcal{H}_{\mathcal{L}} \subset \mathbb{R}^{\mathcal{X}_{\mathcal{L}}}$ with continuous point evaluation functionals. Then the family*

$$\mathcal{L}[f] \coloneqq \{\omega \to \mathcal{L}[f(\cdot,\omega)](x)\}_{x\in\mathcal{X}_{\mathcal{L}}} \tag{A.111}$$

of random variables is a Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $\mathcal{L}[f] \sim \mathcal{GP}(\mathcal{L}[m], \mathcal{L}k\mathcal{L}^*)$, whose paths lie in $\mathcal{H}_{\mathcal{L}}$ such that $\mathcal{L}[f](\cdot,\omega) = \mathcal{L}[f(\cdot,\omega)]$.

A.6. Joint Gaussian Measures on Separable Hilbert Spaces

In order to compute $f \mid \mathcal{L}[f] = h$ as in theorem 1, it is important to know the joint distribution of f and $\mathcal{L}[f]$. If f is a Gaussian random variable on a separable Hilbert space $\mathcal{H}_1$ and $\mathcal{L}$ maps into another separable Hilbert space $\mathcal{H}_2$, then this joint distribution lives in the Cartesian product of $\mathcal{H}_1$ and $\mathcal{H}_2$. We will refer to a Gaussian measure on a Cartesian product of separable Hilbert spaces as a joint Gaussian measure on separable Hilbert spaces. In the following, we will familiarize ourselves with the properties of joint Gaussian measures on separable Hilbert spaces.

Lemma A.8. *Let $\{\mathcal{H}_i\}_{i=1}^n$ be a finite family of real Hilbert spaces. Then the Cartesian product*

$$\mathcal{H}_\times \coloneqq \mathcal{H}_1 \times \cdots \times \mathcal{H}_n \tag{A.112}$$

with canonical addition and scalar multiplication

$$h + h' \coloneqq (h_1 + h_1', \ldots, h_n + h_n') \tag{A.113}$$

$$\alpha h \coloneqq (\alpha h_1, \ldots, \alpha h_n) \tag{A.114}$$

for $h, h' \in \mathcal{H}_\times$ and $\alpha \in \mathbb{R}$ is a vector space. Moreover, $\mathcal{H}_\times$ is a Hilbert space with respect to the inner product

$$\langle h, h' \rangle_{\mathcal{H}_\times} \coloneqq \sum_{i=1}^n \langle h_i, h_i' \rangle_{\mathcal{H}_i}. \tag{A.115}$$

If all every $\mathcal{H}_i$ for $i = 1, \ldots, n$ is separable, then $\mathcal{H}_\times$ is separable. Let $\Pi_i\colon \mathcal{H}_\times \to \mathcal{H}_i, h \mapsto h_i$ for $i \in \{1, \ldots, n\}$. Then Π_i is bounded and

$$\Pi_i^*[h_i] = (\underbrace{0, \ldots, 0}_{i-1 \text{ times}}, h_i, 0, \ldots, 0). \tag{A.116}$$

Proof. Each $\mathcal{H}_i$ is a Banach space w.r.t. the norm $\|\cdot\|_{\mathcal{H}_i} \coloneqq \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}_i}}$ induced by the inner product. Then, by Adams and Fournier [2003, Theorem 1.23], $\mathcal{H}_\times$ with addition and scalar multiplication as defined above is a vector space. Obviously, $\langle \cdot, \cdot \rangle_{\mathcal{H}_\times}$ is an inner product, which induces the norm $\|\cdot\|_{\mathcal{H}_\times}$ with

$$\|h\|_{\mathcal{H}_\times}^2 \coloneqq \langle h, h \rangle_{\mathcal{H}_\times} = \sum_{i=1}^n \langle h_i, h_i \rangle_{\mathcal{H}_i} = \sum_{i=1}^n \|h_i\|_{\mathcal{H}_i}^2. \tag{A.117}$$

By Adams and Fournier [2003, Theorem 1.23], $\mathcal{H}_\times$ is a Banach space w.r.t. $\|\cdot\|_{\mathcal{H}_\times}$, implying that it is a Hilbert space w.r.t. $\langle\cdot,\cdot\rangle_{\mathcal{H}_\times}$. Moreover, again by Adams and Fournier [2003, Theorem 1.23] $\mathcal{H}_\times$ is separable if all $\mathcal{H}_i$ for $i = 1,\dots,n$ are separable. Let $h = (h_1,\dots,h_n) \in \mathcal{H}_\times$. Then

$$\|\Pi_i[h]\|_{\mathcal{H}_i}^2 = \|h_i\|_{\mathcal{H}_i}^2 \leq \sum_{j=1}^n \|h_j\|_{\mathcal{H}_j}^2 = \|h\|_{\mathcal{H}_\times} \tag{A.118}$$

and, for $h' \in \mathcal{H}_\times$ and $h_i \in \mathcal{H}_i$,

$$\left\langle h_i, \Pi_i\left[h'\right]\right\rangle_{\mathcal{H}_i} = \left\langle h_i, h_i'\right\rangle_{\mathcal{H}_i} = \sum_{j=1}^n \left\langle \delta_{ji} h_i, h_i'\right\rangle_{\mathcal{H}_i} = \left\langle (\underbrace{0,\dots,0}_{i-1 \text{ times}}, h_i, 0,\dots,0), h'\right\rangle_{\mathcal{H}_\times}. \tag{A.119}$$

□

Remark A.6. *For linear operators $\mathcal{L}\colon \mathcal{H} \to \mathcal{H}'$ between Cartesian products $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$ and $\mathcal{H}' = \mathcal{H}_1' \times \cdots \times \mathcal{H}_m'$, we introduce the notation*

$$\mathcal{L}[(h_1,\dots,h_n)] = (\mathcal{L}_{11}[h_1] + \cdots + \mathcal{L}_{1n}[h_n], \dots, \mathcal{L}_{m1}[h_1] + \cdots + \mathcal{L}_{mn}[h_n]) \tag{A.120}$$

$$=: \begin{pmatrix} \mathcal{L}_{11} & \dots & \mathcal{L}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{L}_{m1} & \dots & \mathcal{L}_{mn} \end{pmatrix} [(h_1,\dots,h_n)], \tag{A.121}$$

with $\mathcal{L}_{ij} \coloneqq \Pi_i' \mathcal{L} \Pi_j^*\colon \mathcal{H}_j \to \mathcal{H}_i'$, where Π_i' is the projection Π_i in lemma A.8 corresponding to $\mathcal{H}'$.

Just as in the finite-dimensional case, we can use orthogonal projections to margialize out variables in a random vector whose law is a joint Gaussian measure on separable Hilbert spaces. .

Proposition A.3 (Marginalization in Joint Gaussian Measures)**.** *Let $\{\mathcal{H}_i\}_{i=1}^n$ be a family of real separable Hilbert spaces and $\mathcal{H} \coloneqq \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$. Let $f = (f_1,\dots,f_n)$ be an $\mathcal{H}$-valued Gaussian random variable with mean $m = (m_1,\dots,m_n)$ and covariance operator*

$$\mathcal{C} \coloneqq \begin{pmatrix} \mathcal{C}_{11} & \dots & \mathcal{C}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{C}_{n1} & \dots & \mathcal{C}_{nn} \end{pmatrix}. \tag{A.122}$$

For $i_1,\dots,i_k \in \{1,\dots,n\}$ we have

$$(f_{i_1},\dots,f_{i_k}) \sim \mathcal{N}\left((m_{i_1},\dots,m_{i_k}), \begin{pmatrix} \mathcal{C}_{i_1,i_1} & \dots & \mathcal{C}_{i_1,i_k} \\ \vdots & \ddots & \vdots \\ \mathcal{C}_{i_k,i_1} & \dots & \mathcal{C}_{i_k,i_k} \end{pmatrix}\right). \tag{A.123}$$

Proof. The linear operator $\Pi_{i_1,\dots,i_k} \colon \mathcal{H} \to \mathcal{H}_{i_1} \times \cdots \times \mathcal{H}_{i_k}, h \mapsto (h_{i_1}, \dots, h_{i_k})$ is bounded and

$$\Pi^*_{i_1,\dots,i_k} \left[(h_{i_1}, \dots, h_{i_k})\right]_i = \begin{cases} h_{i_l} & \text{if } \exists l \in \{1, \dots, k\} \colon i = i_l, \\ 0 & \text{otherwise.} \end{cases} \tag{A.124}$$

Hence, the result follows from lemma A.6. □

The statistical independence properties of joint Gaussian measures on Hilbert spaces are also essentially analogous to the finite-dimensional case.

Proposition A.4 (Independence in Joint Gaussian Measures)**.** *Let $\{\mathcal{H}_i\}_{i=1}^n$ be a family of real separable Hilbert spaces and $\mathcal{H} \coloneqq \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$. Let $\{f_i\}_{i=1}^n$ be a family of random variables on a common Borel probability space $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$, where $f_i \sim \mathcal{N}(m_i, \mathcal{C}_i)$ with values in $\mathcal{H}_i$. If the random variables $\{f_i\}_{i=1}^n$ are independent, then the random variable*

$$f \colon \Omega \to \mathcal{H}, \omega \mapsto (f_1(\omega), \dots, f_n(\omega)) \tag{A.125}$$

on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ is Gaussian with mean $(m_1, \dots, m_n)$ and covariance operator

$$\mathcal{C} \coloneqq \begin{pmatrix} \mathcal{C}_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathcal{C}_n \end{pmatrix}, \tag{A.126}$$

i.e. $\mathcal{C}_{ij} = \delta_{ij}\mathcal{C}_j$.

Proof. Consider the case, where $n = 2$. Let $f_1 \perp\!\!\!\perp f_2$. Then there is $h^* \in \mathcal{H}^*$ such that $\omega \mapsto h^*[f(\omega)]$ is not Gaussian. By the Riesz representation theorem [Yosida, 1995, Section III.6], there is an $h \in \mathcal{H}$ such that $h^* = \langle h, \cdot \rangle_{\mathcal{H}} = \langle h_1, \cdot \rangle_{\mathcal{H}_1} + \langle h_2, \cdot \rangle_{\mathcal{H}_2}$. The random variables $\langle h_1, f_1 \rangle_{\mathcal{H}_1}$ and $\langle h_1, f_2 \rangle_{\mathcal{H}_1}$ are by definition Gaussian. Moreover, they are independent, since $f_1 \perp\!\!\!\perp f_2$. Since the sum of independent Gaussian random variables is Gaussian, $h^*[f]$ is Gaussian. It follows that f is Gaussian. Since Π_i is bounded, by lemma A.6, we have that $m = (m_1, \dots, m_n)$ and $C_{ii} = \mathcal{C}_i$. Let μ the law of f. Then, for $h_j \in \mathcal{H}_j$,

$$\begin{aligned}
\mathcal{C}_{ij}[h_j] &= \Pi_i \mathcal{C}_{ij} \Pi_j^*[h_j] && \text{(A.127)} \\
&= \Pi_i \left[\int_{\mathcal{H}} \left\langle \Pi_j^*[h_j], h' - m \right\rangle_{\mathcal{H}} (h' - m) \, \mathrm{d}\mu(h') \right] && \text{(A.128)} \\
&= \int_{\mathcal{H}} \left\langle h_j, \Pi_j[h' - m] \right\rangle_{\mathcal{H}} \Pi_i[h' - m] \, \mathrm{d}\mu(h') && \text{(A.129)} \\
& && \text{(theorem A.2, since } \Pi_i \text{ bounded)} \\
&= \int_{\mathcal{H}} \left\langle h_j, \Pi_j[h'] - \Pi_j[m] \right\rangle_{\mathcal{H}} \left(\Pi_i[h'] - \Pi_i[m]\right) \mathrm{d}\mu(h') && \text{(A.130)} \\
&= \int_{\mathcal{H}_j} \int_{\mathcal{H}_i} \left\langle h_j, h_j' - m_j \right\rangle_{\mathcal{H}} \left(h_i' - m_i\right) \mathrm{d}\mu \circ \Pi_i^{-1}(h_i') \, \mathrm{d}\mu \circ \Pi_j^{-1}(h_j') && \text{(A.131)}
\end{aligned}$$

$$(f_i \perp\!\!\!\perp f_j)$$

$$= \int_{\mathcal{H}_j} \left\langle h_j, h_j' - m_j \right\rangle_{\mathcal{H}} \underbrace{\int_{\mathcal{H}_i} (h_i' - m_i) \, \mathrm{d}\mu \circ \Pi_i^{-1}\left(h_i'\right)}_{=0} \, \mathrm{d}\mu \circ \Pi_j^{-1}\left(h_j'\right) \tag{A.132}$$

$$= 0. \tag{A.133}$$

The statement for general n follows by induction and the observation that the Hilbert spaces $(\mathcal{H}_1 \times \cdots \times \mathcal{H}_n) \times \mathcal{H}_{n+1}$ and $\mathcal{H}_1 \times \cdots \times \mathcal{H}_n \times \mathcal{H}_{n+1}$ are isometrically isomorphic. □

Joint Gaussian measures on separable Hilbert spaces also enable us to reason about sums or linear combinations of two Gaussian random variables on separable Hilbert spaces.

Corollary A.3. *Let $\mathcal{H}_1, \mathcal{H}_2$ be real separable Hilbert spaces and*

$$(f, g) \sim \mathcal{N}\left((m_f, m_g), \begin{pmatrix} \mathcal{C}_{ff} & \mathcal{C}_{fg} \\ \mathcal{C}_{fg}^* & \mathcal{C}_{gg} \end{pmatrix}\right) \tag{A.134}$$

an $\mathcal{H}_1 \times \mathcal{H}_2$-valued Gaussian random variable. Let $\mathcal{L}_1 \colon \mathcal{H}_1 \to \mathcal{H}'$ and $\mathcal{L}_2 \colon \mathcal{H}_2 \to \mathcal{H}'$ be bounded linear operators mapping into another real separable Hilbert space $\mathcal{H}'$. Then

$$\mathcal{L}_1[f] + \mathcal{L}_2[g] \tag{A.135}$$

is an $\mathcal{H}'$-valued Gaussian random variable with mean $\alpha m_f + \beta m_g$ and covariance operator

$$\mathcal{L}_1 \mathcal{C}_{ff} \mathcal{L}_1^* + \mathcal{L}_1 \mathcal{C}_{fg} \mathcal{L}_2^* + \mathcal{L}_2 \mathcal{C}_{gf} \mathcal{L}_1^* + \mathcal{L}_2 \mathcal{C}_{gg} \mathcal{L}_2^*. \tag{A.136}$$

Proof. The linear operator $\mathcal{L}' \colon \mathcal{H}_1 \times \mathcal{H}_2 \to \mathcal{H}', (h_1, h_2) \mapsto \mathcal{L}'[(h_1, h_2)] \coloneqq \mathcal{L}_1[h_1] + \mathcal{L}_2[h_2]$ is bounded, since

$$\begin{aligned}
\left\|\mathcal{L}'[(h_1, h_2)]\right\|_{\mathcal{H}'} &= \|\mathcal{L}_1[h_1] + \mathcal{L}_2[h_2]\|_{\mathcal{H}'} && \text{(A.137)} \\
&\leq \|\mathcal{L}_1[h_1]\|_{\mathcal{H}'} + \|\mathcal{L}_2[h_2]\|_{\mathcal{H}'} && \text{(A.138)} \\
&\leq \|\mathcal{L}_1\| \, \|h_1\|_{\mathcal{H}'} + \|\mathcal{L}_2\| \, \|h_2\|_{\mathcal{H}'} && \text{(A.139)} \\
&\leq \max\{\|\mathcal{L}_1\|, \|\mathcal{L}_2\|\} \left(\|h_1\|_{\mathcal{H}_1} + \|h_2\|_{\mathcal{H}_2}\right) && \text{(A.140)} \\
&= \max\{\|\mathcal{L}_1\|, \|\mathcal{L}_2\|\} \, \|(h_1, h_2)\|_{\mathcal{H}_1 \times \mathcal{H}_2}, && \text{(A.141)}
\end{aligned}$$

and $\max\{\|\mathcal{L}_1\|, \|\mathcal{L}_2\|\} < \infty$, because $\mathcal{L}_1$ and $\mathcal{L}_2$ bounded. Moreover,

$$(\mathcal{L}')^*\left[h'\right] = (\mathcal{L}_1^*\left[h'\right], \mathcal{L}_2^*\left[h'\right]). \tag{A.142}$$

Hence, the result follows from lemma A.6. □

We can now investigate the joint distribution of f and $\mathcal{L}[f]$, or, more generally, $\mathcal{L}[f] + g$, where f and g are Gaussian random variables and L is a bounded linear operator.

Proposition A.5. *Let $f \sim \mathcal{N}(m_f, \mathcal{C}_f)$ be a Gaussian random variable on a Borel probability space $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in a real separable Hilbert space $\mathcal{H}_1$. Let $\mathcal{L}\colon \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator mapping into another real separable Hilbert space $\mathcal{H}_2$. Let $g \sim \mathcal{N}(m_g, \mathcal{C}_g)$ be an $\mathcal{H}_2$-valued Gaussian random variable on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with $g \perp\!\!\!\perp f$. Then*

$$\mathcal{L}[f] + g \sim \mathcal{N}(\mathcal{L}[m_f] + m_g, \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g) \tag{A.143}$$

with values in $\mathcal{H}_2$ and

$$(f, \mathcal{L}[f] + g) \sim \mathcal{N}\left((m_f, \mathcal{L}[m_f] + m_g), \begin{pmatrix} \mathcal{C}_f & \mathcal{C}_f\mathcal{L}^* \\ \mathcal{L}\mathcal{C}_f & \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g \end{pmatrix}\right) \tag{A.144}$$

with values in $\mathcal{H}_1 \times \mathcal{H}_2$.

Proof. Let $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$. Since, $\mathcal{L}$ is bounded, the linear operator

$$\tilde{\mathcal{L}} \coloneqq \begin{pmatrix} \mathrm{id}_{\mathcal{H}_1} \\ \mathcal{L} \end{pmatrix} : \mathcal{H}_1 \to \mathcal{H} \qquad \text{with} \qquad \tilde{\mathcal{L}}^* = \begin{pmatrix} \mathrm{id}_{\mathcal{H}_1} & \mathcal{L}^* \end{pmatrix} \tag{A.145}$$

is bounded. Moreover, we know from lemma A.8 that $\Pi_2^*\colon \mathcal{H}_2 \to \mathcal{H}$ is bounded. Since $g \perp\!\!\!\perp f$, proposition A.4 implies that

$$\omega \mapsto (f(\omega), g(\omega)) \sim \mathcal{N}\left((m_f, m_g), \begin{pmatrix} \mathcal{C}_f & 0 \\ 0 & \mathcal{C}_g \end{pmatrix}\right) \tag{A.146}$$

on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in $\mathcal{H}$. Note that

$$\omega \mapsto (f(\omega), \mathcal{L}[f(\omega)] + g(\omega)) = \hat{\mathcal{L}}[f(\omega)] + \Pi_2^*[g(\omega)]. \tag{A.147}$$

Hence, result about $(f, \mathcal{L}[f] + g)$ follows from corollary A.3 and the result about $\mathcal{L}[f] + g$ follows by applying proposition A.3 to $(f, \mathcal{L}[f] + g)$. □

As for regular Gaussian measures on separable Hilbert spaces, we can establish a correspondence between joint Gaussian measures on separable Hilbert spaces and multi-output Gaussian processes. We will first give the general construction and then apply it to the joint Gaussian measure in proposition A.5.

Proposition A.6. *Let $\{\mathcal{H}_i \subset \mathbb{R}^{\mathcal{X}}\}_{i=1}^n$ be a family of real separable Hilbert spaces of real-valued functions on a common domain $\mathcal{X}$ and let $\mathcal{H} \coloneqq \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$. Let*

$$f = (f_1, \ldots, f_n) \sim \mathcal{N}\left((m_1, \ldots, m_n), \begin{pmatrix} \mathcal{C}_{11} & \ldots & \mathcal{C}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{C}_{n1} & \ldots & \mathcal{C}_{nn} \end{pmatrix}\right) \tag{A.148}$$

on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ with values in $\mathcal{H}$. If the point evaluation functionals on $\mathcal{H}_i$ for all $i \in \{1, \ldots, n\}$ are continuous, then the family $\tilde{f} \coloneqq \{\omega \mapsto f_i(\omega)(x)\}_{(i,x)\in I\times\mathcal{X}}$ with $I =$

$\{1, \ldots, n\}$ *is an* n*-output Gaussian process with index set* $\mathcal{X}$ *on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$*. Its with mean and covariance functions are given by* $\tilde{m}(i, x) = m_i(x)$ *and*

$$\tilde{k}((i_1, x_1), (i_2, x_2)) = \mathcal{C}_{i_1, i_2}\left[\delta^*_{x_2}\right](x_1), \tag{A.149}$$

respectively. Moreover, there is an isometry $\mathcal{I}$ *between the path space of* $\tilde{f}$ *and* $\mathcal{H}$ *such that, for all* $\omega \in \Omega$*, we have* $\mathcal{I}\left[\tilde{f}.(\omega)\right] = f(\omega)$*.*

Proof. Let

$$\iota \colon \mathcal{H} \to \tilde{\mathcal{H}} \subset \mathbb{R}^{\mathcal{X}'}, h \mapsto \iota[h](i, x) = h_i(x) \tag{A.150}$$

with $\tilde{\mathcal{H}} \coloneqq \iota[\mathcal{H}]$, which is linear and bijective. Then $\tilde{\mathcal{H}}$ with pointwise addition and scalar multiplication and inner product

$$\left\langle \tilde{h}, \tilde{h}' \right\rangle_{\tilde{\mathcal{H}}} = \left\langle \tilde{h}(1, \cdot), \tilde{h}'(1, \cdot) \right\rangle_{\mathcal{H}_1} + \left\langle \tilde{h}(2, \cdot), \tilde{h}'(2, \cdot) \right\rangle_{\mathcal{H}_2} \tag{A.151}$$

is a Hilbert space. Moreover,

$$\begin{aligned}
\|\iota[h]\|^2_{\tilde{\mathcal{H}}} &= \langle \iota[h], \iota[h] \rangle_{\tilde{\mathcal{H}}} && \text{(A.152)}\\
&= \langle \iota[h](1, \cdot), \iota[h](1, \cdot) \rangle_{\mathcal{H}_1} + \langle \iota[h](2, \cdot), \iota[h](2, \cdot) \rangle_{\mathcal{H}_2} && \text{(A.153)}\\
&= \langle h_1, h_1 \rangle_{\mathcal{H}_1} + \langle h_2, h_2 \rangle_{\mathcal{H}_2} && \text{(A.154)}\\
&= \langle h, h \rangle_{\mathcal{H}} && \text{(A.155)}\\
&= \|h\|^2_{\mathcal{H}}, && \text{(A.156)}
\end{aligned}$$

i.e. ι is bounded (it is even an isometry). By lemma A.6, it follows that $\iota[f]$ is a Gaussian random variable with mean $(i, x) \mapsto \iota[m](i, x) = m_i(x)$ and covariance operator $\iota \mathcal{C} \iota^*$. Since the point evaluation functionals on all $\mathcal{H}_i$ are continuous, it follows that the point evaluation functionals on $\mathcal{H}'$ are continuous. Hence, by corollary A.1, $\tilde{f}$ is indeed a Gaussian process on $I \times \mathcal{X}$ with mean function $\tilde{m}$ and covariance function

$$\begin{aligned}
((i_1, x_1), (i_2, x_2)) \mapsto \iota \mathcal{C} \iota^* \left[\delta^*_{(i_2, x_2)}\right](i_1, x_1) && \text{(A.157)}\\
= \iota\left[\mathcal{C}\Pi_{i_2}\left[\delta^*_{x_2}\right]\right](i_1, x_1) && \text{(A.158)}\\
= \iota\left[\left(\mathcal{C}_{1, i_2}\left[\delta^*_{x_2}\right], \ldots, \mathcal{C}_{n, i_2}\left[\delta^*_{x_2}\right]\right)\right](i_1, x_1) && \text{(A.159)}\\
= \mathcal{C}_{i_1, i_2}\left[\delta^*_{x_2}\right](x_1). && \text{(A.160)}
\end{aligned}$$

□

Corollary A.4. *Let the assumptions of corollary A.2 hold with* $m_f \coloneqq m$ *and* $k_f \coloneqq k$*, and let* $g \sim \mathcal{GP}(m_g, k_g)$ *with* $g \perp\!\!\!\perp f$ *and paths in* $\mathcal{H}_{\mathcal{L}}$*. Then the family*

$$h \coloneqq \{\omega \mapsto \mathcal{L}[f(\cdot, \omega)](x) + g(x, \omega)\}_{x \in \mathcal{X}_{\mathcal{L}}} \tag{A.161}$$

is a Gaussian process on $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with*

$$\mathcal{L}[f] + g \sim \mathcal{GP}(\mathcal{L}[m_f] + m_g, \mathcal{L} k_f \mathcal{L}^* + k_g), \tag{A.162}$$

whose paths lie in $\mathcal{H}_{\mathcal{L}}$ *such that* $h(\cdot,\omega) = \mathcal{L}[f(\cdot,\omega)] + g(\cdot,\omega)$. *If additionally* $\mathcal{X}_{\mathcal{L}} = \mathcal{X}$, *then the family*

$$\begin{pmatrix} f & \mathcal{L}[f] + g \end{pmatrix}^\top \coloneqq \{\omega \mapsto (f(\cdot,\omega), \mathcal{L}[f(\cdot,\omega)] + g(\cdot,\omega))_i(x)\}_{(i,x)\in\{1,2\}\times\mathcal{X}} \tag{A.163}$$

is a 2-output Gaussian process with index set $\mathcal{X}$ *on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with*

$$\begin{pmatrix} f \\ \mathcal{L}[f] + g \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m_f \\ \mathcal{L}[m_f] + m_g \end{pmatrix}, \begin{pmatrix} k_f & k_f\mathcal{L}^* \\ \mathcal{L}k_f & \mathcal{L}k_f\mathcal{L}^* + k_g \end{pmatrix}\right) \tag{A.164}$$

and paths in $\mathcal{H} \times \mathcal{H}_{\mathcal{L}}$ *such that*

$$\begin{pmatrix} f \\ \mathcal{L}[f] + g \end{pmatrix}(\cdot,\omega) = \begin{pmatrix} f(\cdot,\omega) \\ \mathcal{L}[f(\cdot,\omega)] + g(\cdot,\omega) \end{pmatrix}. \tag{A.165}$$

A.7. Gaussian Processes are Closed Under Conditioning on Affine Observations

The final ingredient needed to prove theorem 1 is a way to condition joint Gaussian measures on one of their "components". Owhadi and Scovel [2018] show how to condition Gaussian measures on an orthogonal direct sum of separable Hilbert spaces on observations in one of the two subspaces, i.e. they show how to compute $x \mid x_2 = t$, where $x = x_1 + x_2$ is a Gaussian random variable with values in $\mathcal{H}_1 \oplus \mathcal{H}_2$. We will extend this result, since Owhadi and Scovel [2018] don't give explicit expressions for the conditional mean and covariance operator.

But first of all, we need to investigate how our joint Gaussian measure from proposition A.5 fits into the formalism of orthogonal direct sums of separable Hilbert spaces.

Remark A.7. *The Cartesian product* $\mathcal{H}_\times$ *of Hilbert spaces from lemma A.8*

$$\mathcal{H}_\times = \hat{\mathcal{H}}_1 \oplus \cdots \oplus \hat{\mathcal{H}}_n, \tag{A.166}$$

is an orthogonal direct sum of Hilbert spaces

$$\hat{\mathcal{H}}_i \coloneqq \{(\underbrace{0,\ldots,0}_{i-1 \text{ times}}, h, 0, \ldots, 0) \mid h_i \in \mathcal{H}_i\} \tag{A.167}$$

for $i = 1, \ldots, n$. *Moreover, the mapping*

$$\hat{\mathcal{I}}_{\mathcal{H}_i}^{-1} \colon \hat{\mathcal{H}}_i \to \mathcal{H}_i, \hat{h}_i \to h_i \coloneqq (\hat{h}_i)_i \tag{A.168}$$

is an isometry and hence bounded.

We will now introduce some notation, which makes working with orthogonal direct sums of Hilbert spaces easier.

Remark A.8. *Let $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_n$ be an orthogonal direct sum of Hilbert spaces. For every $h \in \mathcal{H}$, we introduce the notation*

$$h = h_1 + \cdots + h_n =: \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} \tag{A.169}$$

with $h_i \coloneqq \Pi_i [h]$, where Π_i denotes orthogonal projection onto $\mathcal{H}_i$. For a linear operator $\mathcal{L} \colon \mathcal{H} \to \mathcal{H}'$ mapping into another orthogonal direct sum $\mathcal{H}' = \mathcal{H}'_1 \oplus \cdots \oplus \mathcal{H}'_m$ of Hilbert spaces, we commonly use block matrix notation, i.e.

$$\mathcal{L} [h_1 + \cdots + h_n] = \sum_{i=1}^{m} \mathcal{L}_{i1} [h_1] + \cdots + \mathcal{L}_{in} [h_n] \tag{A.170}$$

$$=: \begin{pmatrix} \mathcal{L}_{11} & \dots & \mathcal{L}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{L}_{m1} & \dots & \mathcal{L}_{mn} \end{pmatrix} [h_1 + \cdots + h_n] \tag{A.171}$$

with $\mathcal{L}_{ij} \colon \mathcal{H}_j \to \mathcal{H}_i, h_j \mapsto \Pi_i \mathcal{L} [h_j]$.

We now prove a special case of theorem 3.3 in Owhadi and Scovel [2018], which gives explicit expressions for the mean and covariance operator of the conditional measure. These expressions are well-known for conditional Gaussian measures on finite-dimensional Euclidean vector spaces.

Theorem A.3. *Let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be an orthogonal direct sum of real separable Hilbert spaces. Let $x = x_1 + x_2$ be an $\mathcal{H}$-valued Gaussian random variable with mean $m = m_1 + m_2$, and covariance operator*

$$\mathcal{C} \coloneqq \begin{pmatrix} \mathcal{C}_{11} & \mathcal{C}_{12} \\ \mathcal{C}_{12}^* & \mathcal{C}_{22} \end{pmatrix} : \mathcal{H} \to \mathcal{H} \tag{A.172}$$

such that $\operatorname{ran}(\mathcal{C}_{22})$ is closed. Then $x \mid x_2 = t$ for any $t \in \mathcal{H}_2$ is an $\mathcal{H}$-valued Gaussian random variable with mean

$$m_{x|x_2=t} \coloneqq \begin{pmatrix} m_1 + \mathcal{C}_{12} \mathcal{C}_{22}^{\dagger} [t - m_2] \\ t \end{pmatrix}, \tag{A.173}$$

and covariance operator

$$\mathcal{C}_{x|x_2=t} \coloneqq \begin{pmatrix} \mathcal{C}_{11} - \mathcal{C}_{12} \mathcal{C}_{22}^{\dagger} \mathcal{C}_{12}^* & 0 \\ 0 & 0 \end{pmatrix}. \tag{A.174}$$

Proof. By theorem 3.3 in Owhadi and Scovel [2018], $x \mid x_2 = t$ is Gaussian (since its law is Gaussian), its covariance operator is the short of $\mathcal{C}$ to $\mathcal{H}_2$ [Owhadi and Scovel,

2018, Anderson and Trapp, 1975] and its mean is given by $\begin{pmatrix} m_1 + \hat{Q}^*(t - m_2) & t \end{pmatrix}^\top$ for any $\mathcal{C}$-symmetric oblique projection

$$Q \coloneqq \begin{pmatrix} 0 & 0 \\ \hat{\mathcal{Q}} & \mathrm{id}_{\mathcal{H}_2} \end{pmatrix} \in \mathcal{P}(\mathcal{C}, \mathcal{H}_2), \tag{A.175}$$

onto $\mathcal{H}_2$ if $\mathcal{C}$ is compatible with $\mathcal{H}_2$, i.e. if $\mathcal{P}(\mathcal{C}, \mathcal{H}_2) \neq \emptyset$. In the following, we will show that $m_{x|x_2=t}$ and $C_{x|x_2=t}$ are indeed the mean and covariance operator described by the theorem.

The covariance operator $\mathcal{C}$ of x is in the trace class [Maniglia and Rhandi, 2004, Lemma 1.1.4] and hence bounded [Yosida, 1995, Section X.2]. This implies that the operators $\mathcal{C}_{12}$ and $\mathcal{C}_{22}$ are bounded, since orthogonal projection onto $\mathcal{H}_1$ and $\mathcal{H}_2$ is bounded. Moreover, $\mathcal{C}_{22}$ is self-adjoint and positive, which means that its square root $\sqrt{\mathcal{C}_{22}}$ exists and is also bounded, self-adjoint and positive [Bernau, 1968, Theorem 4]. By theorem 3 and the corollary of lemma 1 in Anderson and Trapp [1975], we have

$$\mathrm{ran}\,(\mathcal{C}_{12}^*) \subset \mathrm{ran}\left(\sqrt{\mathcal{C}_{22}}\right) = \mathrm{ran}\,(\mathcal{C}_{22}) \tag{A.176}$$

and hence $\mathrm{ran}\left(\sqrt{\mathcal{C}_{22}}\right)$ is closed, because $\mathrm{ran}\,(\mathcal{C}_{22})$ is closed by assumption. Consequently, by theorem 3 in Ben-Israel and Greville [2003, Section 8.3], the Moore-Penrose pseudoinverses $\mathcal{C}_{22}^\dagger, \mathcal{C}_{22} \colon \mathcal{H}_2 \to \mathcal{H}_2$ exist and are bounded. Additionally, by the closed graph theorem [Yosida, 1995, Section II.6], $\mathcal{C}_{22}^\dagger$ and $\mathcal{C}_{22}$ are closed linear operators, which together with theorem 2 (g) and (i) in Ben-Israel and Greville [2003, Section 8.3] implies $(\mathcal{C}_{22}^\dagger)^* = (\mathcal{C}_{22}^*)^\dagger = \mathcal{C}_{22}^\dagger$, $(\sqrt{\mathcal{C}_{22}}^\dagger)^* = (\sqrt{\mathcal{C}_{22}}^*)^\dagger = \sqrt{\mathcal{C}_{22}}^\dagger$, and

$$(\sqrt{\mathcal{C}_{22}}^\dagger)^* \sqrt{\mathcal{C}_{22}}^\dagger = \sqrt{\mathcal{C}_{22}}^\dagger \sqrt{\mathcal{C}_{22}}^\dagger = \left(\sqrt{\mathcal{C}_{22}}\sqrt{\mathcal{C}_{22}}\right)^\dagger = \mathcal{C}_{22}^\dagger. \tag{A.177}$$

Let $\hat{\mathcal{Q}} \coloneqq \mathcal{C}_{22}^\dagger \mathcal{C}_{12}^*$, i.e. $\hat{\mathcal{Q}}^* = \mathcal{C}_{12}(\mathcal{C}_{22}^\dagger)^* = \mathcal{C}_{12}\mathcal{C}_{22}^\dagger$. We will now show that $\mathcal{Q}$ as defined above with this choice of $\hat{\mathcal{Q}}$ is a $\mathcal{C}$-symmetric oblique projection onto $\mathcal{H}_2$. From the boundedness of $\mathcal{C}_{22}^\dagger$ and $\mathcal{C}_{12}$, it follows that $\hat{\mathcal{Q}}$ and $\mathcal{Q}$ are bounded. Evidently, $\mathcal{Q}$ is idempotent, i.e. $\mathcal{Q}^2 = \mathcal{Q}$, and $\mathrm{ran}\,(\mathcal{Q}) = \mathcal{H}_2$, since $\mathrm{ran}\left(\hat{\mathcal{Q}}\right) \subset \mathrm{ran}\left(\mathcal{C}_{22}^\dagger\right) \subset \mathcal{H}_2$ and $\mathrm{ran}\,(\mathrm{id}_{\mathcal{H}_2}) = \mathcal{H}_2$. The adjoint $\mathcal{Q}^*$ of $\mathcal{Q}$ is defined by

$$\left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \mathcal{Q}\left[\begin{pmatrix} h_1' \\ h_2' \end{pmatrix}\right] \right\rangle_{\mathcal{H}} = \left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} 0 \\ \hat{\mathcal{Q}}\,[h_1'] + h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} \tag{A.178}$$

$$= \left\langle \begin{pmatrix} h_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \hat{\mathcal{Q}}\,[h_1'] + h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} + \left\langle \begin{pmatrix} 0 \\ h_2 \end{pmatrix}, \begin{pmatrix} 0 \\ \hat{\mathcal{Q}}\,[h_1'] + h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} \tag{A.179}$$

$$= \left\langle h_2, \hat{\mathcal{Q}}\left[h_1'\right] + h_2' \right\rangle_{\mathcal{H}_2} \tag{A.180}$$

$$(h_1 \in \mathcal{H}_1 \text{ and } \hat{\mathcal{Q}}\,[h_1'] + h_2' \in \mathcal{H}_2 \text{ are orthogonal in } \mathcal{H})$$

$$= \left\langle h_2, \hat{\mathcal{Q}}\left[h_1'\right] \right\rangle_{\mathcal{H}_2} + \left\langle h_2, h_2' \right\rangle_{\mathcal{H}_2} \tag{A.181}$$

$$= \left\langle \hat{\mathcal{Q}}^* [h_2], h_1' \right\rangle_{\mathcal{H}_1} + \langle h_2, h_2' \rangle_{\mathcal{H}_2} \tag{A.182}$$

$$= \left\langle \begin{pmatrix} \hat{\mathcal{Q}}^* [h_2] \\ 0 \end{pmatrix}, \begin{pmatrix} h_1' \\ 0 \end{pmatrix} \right\rangle_{\mathcal{H}} + \left\langle \begin{pmatrix} 0 \\ h_2 \end{pmatrix}, \begin{pmatrix} h_1' \\ 0 \end{pmatrix} \right\rangle_{\mathcal{H}} \tag{A.183}$$

$$+ \left\langle \begin{pmatrix} \hat{\mathcal{Q}}^* [h_2] \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} + \left\langle \begin{pmatrix} 0 \\ h_2 \end{pmatrix}, \begin{pmatrix} 0 \\ h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} \tag{A.184}$$

($\mathcal{H}_1$ and $\mathcal{H}_2$ are orthogonal subspaces)

$$= \left\langle \begin{pmatrix} \hat{\mathcal{Q}}^* [h_2] \\ h_2 \end{pmatrix}, \begin{pmatrix} h_1' \\ h_2' \end{pmatrix} \right\rangle_{\mathcal{H}} \tag{A.185}$$

for all $h_1 + h_2, h_1' + h_2' \in \mathcal{H}$, i.e.

$$\mathcal{Q}^* \left[\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right] = \begin{pmatrix} \mathcal{C}_{12} (\mathcal{C}_{22}^\dagger)^* [h_2] \\ h_2 \end{pmatrix} = \begin{pmatrix} \mathcal{C}_{12} \mathcal{C}_{22}^\dagger [h_2] \\ h_2 \end{pmatrix}. \tag{A.186}$$

It follows that, for all $h_1 + h_2 \in \mathcal{H}$ with $h_2 \in \operatorname{ran}(\mathcal{C}_{22})$,

$$\mathcal{Q}^* \mathcal{C} \left[\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right] = \mathcal{Q}^* \left[\begin{pmatrix} \mathcal{C}_{11} [h_1] + \mathcal{C}_{12} [h_2] \\ \mathcal{C}_{12}^* [h_1] + \mathcal{C}_{22} [h_2] \end{pmatrix} \right] \tag{A.187}$$

$$= \begin{pmatrix} \mathcal{C}_{12} \mathcal{C}_{22}^\dagger \left[\mathcal{C}_{12}^* [h_1] + \mathcal{C}_{22} [h_2] \right] \\ \mathcal{C}_{12}^* [h_1] + \mathcal{C}_{22} [h_2] \end{pmatrix} \tag{A.188}$$

$$= \begin{pmatrix} \mathcal{C}_{12} \left[\mathcal{C}_{22}^\dagger \mathcal{C}_{12}^* [h_1] + \mathcal{C}_{22}^\dagger \mathcal{C}_{22} [h_2] \right] \\ \mathcal{C}_{12}^* [h_1] + \mathcal{C}_{22} [h_2] \end{pmatrix} \tag{A.189}$$

$$= \begin{pmatrix} \mathcal{C}_{12} \left[\mathcal{C}_{22}^\dagger \mathcal{C}_{12}^* [h_1] + h_2 \right] \\ \mathcal{C}_{22} \left[\mathcal{C}_{22}^\dagger \mathcal{C}_{12}^* [h_1] + h_2 \right] \end{pmatrix} \tag{A.190}$$

($\mathcal{C}_{22} \mathcal{C}_{22}^\dagger|_{\operatorname{ran}(\mathcal{C}_{22})} = \mathrm{id}_{\operatorname{ran}(\mathcal{C}_{22})}$ and $\operatorname{ran}(\mathcal{C}_{21}^*) \subset \operatorname{ran}(\mathcal{C}_{22})$ by equation (A.176))

$$= \mathcal{C} \left[\begin{pmatrix} 0 \\ \mathcal{C}_{22}^\dagger \mathcal{C}_{12}^* [h_1] + h_2 \end{pmatrix} \right] \tag{A.191}$$

$$= \mathcal{C} \mathcal{Q} \left[\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right]. \tag{A.192}$$

Additionally, for $h_1 + h_2 \in \mathcal{H}$ with $h_2 \notin \operatorname{ran}(\mathcal{C}_{22})$, there are $h_2^{\parallel} \in \operatorname{ran}(\mathcal{C}_{22})$ and $h_2^{\perp} \in \ker(\mathcal{C}_{22})$ such that $h_2 = h_2^{\parallel} + h_2^{\perp}$ and hence $h_2^{\perp} \in \operatorname{ran}(\mathcal{C}_{22})^{\perp} \subset \operatorname{ran}(\mathcal{C}_{12}^*)^{\perp} = \ker(\mathcal{C}_{12})$, i.e. $h_2^{\perp} \in \ker(\mathcal{C})$. Consequently,

$$\mathcal{Q}^* \mathcal{C} \left[\begin{pmatrix} h_1 \\ h_2^{\parallel} + h_2^{\perp} \end{pmatrix} \right] = \mathcal{Q}^* \mathcal{C} \left[\begin{pmatrix} h_1 \\ h_2^{\parallel} \end{pmatrix} \right] + \mathcal{Q}^* \mathcal{C} \left[\begin{pmatrix} 0 \\ h_2^{\perp} \end{pmatrix} \right] \tag{A.193}$$

$$= \mathcal{Q}^* \mathcal{C} \left[\begin{pmatrix} h_1 \\ h_2^{\parallel} \end{pmatrix} \right] + \mathcal{Q}^* [0] \tag{A.194}$$

$$(h_2^\perp \in \ker(\mathcal{C}))$$

$$= \mathcal{C}\mathcal{Q}^* \begin{bmatrix} \begin{pmatrix} h_1 \\ h_2^\parallel \end{pmatrix} \end{bmatrix} \tag{A.195}$$

$$(h_2^\parallel \in \operatorname{ran}(\mathcal{C}_{22}))$$

$$= \mathcal{C}\mathcal{Q}^* \begin{bmatrix} \begin{pmatrix} h_1 \\ h_2^\parallel \end{pmatrix} \end{bmatrix} + \mathcal{C} \begin{bmatrix} \begin{pmatrix} 0 \\ h_2^\perp \end{pmatrix} \end{bmatrix} \tag{A.196}$$

$$(h_2 \in \ker(\mathcal{C}))$$

$$= \mathcal{C}\mathcal{Q}^* \begin{bmatrix} \begin{pmatrix} h_1 \\ h_2^\parallel \end{pmatrix} \end{bmatrix} + \mathcal{C}\mathcal{Q} \begin{bmatrix} \begin{pmatrix} 0 \\ h_2^\perp \end{pmatrix} \end{bmatrix} \tag{A.197}$$

$$(\mathcal{Q}\left[h_2^\perp\right] = h_2^\perp \in \mathcal{H}_2)$$

$$= \mathcal{C}\mathcal{Q} \begin{bmatrix} \begin{pmatrix} h_1 \\ h_2^\parallel + h_2^\perp \end{pmatrix} \end{bmatrix}. \tag{A.198}$$

Hence $\mathcal{Q}^*\mathcal{C} = \mathcal{C}\mathcal{Q}$. All in all, we showed that $\mathcal{Q} \in \mathcal{P}(\mathcal{C}, \mathcal{H}_2) \neq \emptyset$, which, by theorem 3.3 in Owhadi and Scovel [2018], implies that $m_{x|x_2=t}$ is indeed the mean of $x \mid x_2 = t$.

Now let $\mathcal{A} \coloneqq \sqrt{\mathcal{C}_{22}}^\dagger \mathcal{C}_{12}^*$. Then

$$\sqrt{\mathcal{C}_{22}}\mathcal{A} = \sqrt{\mathcal{C}_{22}}\sqrt{\mathcal{C}_{22}}^\dagger \mathcal{C}_{12}^* = \mathcal{C}_{12}^*, \tag{A.199}$$

since $\sqrt{\mathcal{C}_{22}}\sqrt{\mathcal{C}_{22}}^\dagger|_{\operatorname{ran}(\sqrt{\mathcal{C}_{22}})} = \mathrm{id}_{\operatorname{ran}(\sqrt{\mathcal{C}_{22}})}$ [Ben-Israel and Greville, 2003, Section 8.3, Definition 1] and $\operatorname{ran}(\mathcal{C}_{12}^*) \subset \operatorname{ran}\left(\sqrt{\mathcal{C}_{22}}\right)$ by equation (A.176). Thus, by theorem 3 in Anderson and Trapp [1975], it follows that the short of $\mathcal{C}$ to $\mathcal{H}_2$ is given by

$$\begin{pmatrix} \mathcal{C}_{11} - \mathcal{A}^*\mathcal{A} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathcal{C}_{11} - \mathcal{C}_{12}(\sqrt{\mathcal{C}_{22}}^\dagger)^*\sqrt{\mathcal{C}_{22}}^\dagger \mathcal{C}_{12}^* & 0 \\ 0 & 0 \end{pmatrix} \tag{A.200}$$

$$= \begin{pmatrix} \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^\dagger \mathcal{C}_{12}^* & 0 \\ 0 & 0 \end{pmatrix} \tag{A.201}$$

$$\text{(by equation (A.177))}$$

$$= \mathcal{C}_{x|x_2=t}. \tag{A.202}$$

□

Remark A.9. *The requirement that* $\operatorname{ran}(\mathcal{C}_{22})$ *needs to be closed might seem a bit opaque at first. However, a simple sufficient criterion to check whether it holds is to check, whether* $\operatorname{ran}(\mathcal{C}_{22})$ *is finite-dimensional, e.g. a subspace of* $\mathbb{R}^n$ *or some function space spanned by a finite number of functions.*

This criterion is actually sufficient, which can be seen as follows. It is well-known that the pseudoinverse of a bounded operator is bounded if and only the range of the operator is closed [Clason, 2021, Theorem 3.7]. The marginal covariance operator $\mathcal{C}_{22}$ *is*

in the trace class and hence compact [Maniglia and Rhandi, 2004, Yosida, 1995]. For compact operators, an even stronger result holds. Namely, the pseudoinverse of a compact operator is bounded if and only if its range is finite-dimensional [Clason, 2021, Corollary 3.8]. Since the pseudoinverse and the inverse of an invertible operator coincide, this also means that, by the bounded inverse theorem, compact operators can only be invertible if they map between finite-dimensional spaces.

Corollary A.5. *Under the assumptions of theorem A.3, $x_1 \mid x_2 = t$ for any $t \in \mathcal{H}_2$ is an $\mathcal{H}_1$-valued Gaussian random variable with mean*

$$m_{x_1|x_2=t} \coloneqq m_1 + \mathcal{C}_{12}\mathcal{C}_{22}^{\dagger}\left[t - m_2\right] \tag{A.203}$$

and covariance operator

$$\mathcal{C}_{x_1|x_2=t} \coloneqq \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^{\dagger}\mathcal{C}_{12}^{*}. \tag{A.204}$$

Proof. The operator $\Pi_1 \colon \mathcal{H} \to \mathcal{H}_1, h = h_1 + h_2 \mapsto h_1$ is bounded, since

$$\left\|\Pi_1\left[\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}\right]\right\|_{\mathcal{H}} = \left\|\begin{pmatrix} h_1 \\ 0 \end{pmatrix}\right\|_{\mathcal{H}} = \left\|h_1\right\|_{\mathcal{H}_1}. \tag{A.205}$$

Moreover, $\Pi_1^*\left[h_1\right] = \begin{pmatrix} h_1 & 0 \end{pmatrix}^{\top}$. Hence, the result follows from lemma A.6. □

We can now apply corollary A.5 to the joint Gaussian measure from proposition A.5 to condition on linear operator observations.

Corollary A.6. *Let $\mathcal{L} \colon \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator between real separable Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. Let $f \sim \mathcal{N}\left(m_f, \mathcal{C}_f\right)$ be an $\mathcal{H}_1$-valued Gaussian random variable and let $g \sim \mathcal{N}\left(m_g, \mathcal{C}_g\right)$ be an $\mathcal{H}_2$-valued Gaussian random variable with $g \perp\!\!\!\perp f$. If $\operatorname{ran}\left(\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g\right)$ is closed, then $f \mid \mathcal{L}\left[f\right] + g = t$ for all $t \in \mathcal{H}_2$ is an $\mathcal{H}_1$-valued Gaussian random variable with mean*

$$m_{f|\mathcal{L}[f]+g=t} \coloneqq m_f + \mathcal{C}\mathcal{L}^*\left(\mathcal{L}\mathcal{C}\mathcal{L}^* + \mathcal{C}_g\right)^{\dagger}\left[t - \left(\mathcal{L}\left[m_f\right] + m_g\right)\right] \tag{A.206}$$

and covariance operator

$$\mathcal{C}_{f|\mathcal{L}[f]+g=t} = \mathcal{C}_f - \mathcal{C}_f\mathcal{L}^*\left(\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g\right)^{\dagger}\mathcal{L}\mathcal{C}_f. \tag{A.207}$$

Proof. By proposition A.5, $\omega \mapsto \left(f(\omega), \mathcal{L}\left[f(\omega)\right] + g(\omega)\right)$ is a Gaussian random variable on $\mathcal{H}_1 \times \mathcal{H}_2$ with mean $\left(m_f, \mathcal{L}\left[m_f\right] + m_g\right)$ and covariance operator

$$\begin{pmatrix} \mathcal{C}_f & \mathcal{C}_f\mathcal{L}^* \\ \mathcal{L}\mathcal{C}_f & \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g \end{pmatrix}. \tag{A.208}$$

By remark A.7, $\mathcal{H}_1 \times \mathcal{H}_2$ is a direct sum of Hilbert spaces $\hat{\mathcal{H}}_1 \times \hat{\mathcal{H}}_2$. Hence, by corollary A.5, $\hat{\mathcal{I}}_{\mathcal{H}_1}\left[f\right] \mid \hat{\mathcal{I}}_{\mathcal{H}_2}\left[\mathcal{L}\left[f\right] + g\right] = \hat{\mathcal{I}}_{\mathcal{H}_2}\left[t\right]$ is a $\hat{\mathcal{H}}_1$-valued Gaussian random variable with mean

$$\hat{\mathcal{I}}_{\mathcal{H}_1}\left[m_f\right] + \left(\hat{\mathcal{I}}_{\mathcal{H}_1}\mathcal{C}_f\mathcal{L}^*\hat{\mathcal{I}}_{\mathcal{H}_2}^{-1}\right)\left(\hat{\mathcal{I}}_{\mathcal{H}_2}\left(\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g\right)\hat{\mathcal{I}}_{\mathcal{H}_2}^{-1}\right)^{\dagger}\left[\hat{\mathcal{I}}_{\mathcal{H}_2}\left[t\right] - \hat{\mathcal{I}}_{\mathcal{H}_2}\left[\mathcal{L}\left[m_f\right] + m_g\right]\right] \tag{A.209}$$

$$= \hat{\mathcal{I}}_{\mathcal{H}_1} \left[m_f + \mathcal{C}_f \mathcal{L}^* \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right)^\dagger \left[t - \left(\mathcal{L} \left[m_f \right] + m_g \right) \right] \right] \tag{A.210}$$

and covariance operator

$$\left(\hat{\mathcal{I}}_{\mathcal{H}_1} \mathcal{C}_f \hat{\mathcal{I}}_{\mathcal{H}_1}^{-1} \right) - \left(\hat{\mathcal{I}}_{\mathcal{H}_1} \mathcal{C}_f \mathcal{L}^* \hat{\mathcal{I}}_{\mathcal{H}_2}^{-1} \right) \left(\hat{\mathcal{I}}_{\mathcal{H}_2} \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right) \hat{\mathcal{I}}_{\mathcal{H}_2}^{-1} \right)^\dagger \left(\hat{\mathcal{I}}_{\mathcal{H}_2} \mathcal{L} \mathcal{C}_f \hat{\mathcal{I}}_{\mathcal{H}_1}^{-1} \right) \tag{A.211}$$

$$= \hat{\mathcal{I}}_{\mathcal{H}_1} \left(\mathcal{C}_f - \mathcal{C}_f \mathcal{L}^* \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right)^\dagger \mathcal{L} \mathcal{C}_f \right) \hat{\mathcal{I}}_{\mathcal{H}_1}^{-1}. \tag{A.212}$$

Note that $\hat{\mathcal{I}}_{\mathcal{H}_2} \left[\mathcal{L} \left[f \right] + g \right] = \hat{\mathcal{I}}_{\mathcal{H}_2} \left[t \right]$ if and only if $\mathcal{L} \left[f \right] + g = t$ and hence

$$\left(\hat{\mathcal{I}}_{\mathcal{H}_1} \left[f \right] \mid \hat{\mathcal{I}}_{\mathcal{H}_2} \left[\mathcal{L} \left[f \right] + g \right] = \hat{\mathcal{I}}_{\mathcal{H}_2} \left[t \right] \right) = \left(\hat{\mathcal{I}}_{\mathcal{H}_1} \left[f \right] \mid \mathcal{L} \left[f \right] + g = t \right). \tag{A.213}$$

By applying the bounded operator $\hat{\mathcal{I}}_{\mathcal{H}_1}^{-1}$ to the latter (via lemma A.6), the result follows. □

If the random variable f in corollary A.6 is induced by a Gaussian process, then the law of $f \mid \mathcal{L} \left[f \right] + g = t$ is a Gaussian measure over the paths of another Gaussian process.

Corollary A.7 (Conditioning a GP on Affine Observations)**.** *Let assumption A.2 hold with $m_f \coloneqq m$ and $k_f \coloneqq k$, and let $\mathcal{C}_f \coloneqq \mathcal{C}_k$ be defined as in proposition A.1. Let $\mathcal{L} \colon \mathcal{H} \to \mathcal{H}_\mathcal{L}$ be a bounded linear operator mapping into a separable Hilbert space $\mathcal{H}_\mathcal{L}$ and let $g \sim \mathcal{N} \left(m_g, \mathcal{C}_g \right)$ be an $\mathcal{H}_\mathcal{L}$-valued Gaussian random variable on $\left(\Omega, \mathcal{B} \left(\Omega \right), \mathrm{P} \right)$ with $g \perp\!\!\!\perp f$. If $\operatorname{ran} \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right)$ is closed, then the family $\{ f_x \}_{x \in \mathcal{X}}$ is a Gaussian process on the Borel probability space $\left(\Omega, \mathcal{B} \left(\Omega \right), \mathrm{P}(\cdot \mid \mathcal{L} \left[f. \right] + g = t) \right)$ with mean function*

$$m_{f \mid \mathcal{L}[f] + g = t} \coloneqq m_f(x) - \left\langle \mathcal{L} \left[k_f(x, \cdot) \right], \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right)^\dagger \left[t - \left(\mathcal{L} \left[m_f \right] + m_g \right) \right] \right\rangle_{\mathcal{H}_\mathcal{L}} \tag{A.214}$$

and covariance function

$$k_{f \mid \mathcal{L}[f] + g = t}(x_1, x_2) \coloneqq k_f(x_1, x_2) - \left\langle \mathcal{L} \left[k_f(x_1, \cdot) \right], \left(\mathcal{L} \mathcal{C}_f \mathcal{L}^* + \mathcal{C}_g \right)^\dagger \mathcal{L} \left[k_f(\cdot, x_2) \right] \right\rangle_{\mathcal{H}_\mathcal{L}}. \tag{A.215}$$

With a slight abuse of notation, we write

$$f \mid \mathcal{L} \left[f \right] + g = t \sim \mathcal{GP} \left(m_{f \mid \mathcal{L}[f] + g = t}, k_{f \mid \mathcal{L}[f] + g = t} \right). \tag{A.216}$$

Proof. This follows by corollaries A.1 and A.6 and lemma A.9. □

Lemma A.9. *Let $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel, $\mathcal{H} \subset \mathbb{R}^\mathcal{X}$ a real Hilbert function space with $\mathcal{H}_k \subset \mathcal{H}$ and continuous point evaluation functionals, $\mathcal{L} \colon \mathcal{H} \to \mathcal{H}_\mathcal{L}$ a linear operator mapping into a real Hilbert space $\mathcal{H}_\mathcal{L}$, and*

$$\mathcal{K} \colon \mathcal{H} \to \mathcal{H}, h \mapsto \mathcal{K} \left[h \right] (x) \coloneqq \left\langle k(x, \cdot), h \right\rangle_\mathcal{H}. \tag{A.217}$$

Then

$$\mathcal{L} \mathcal{K} \left[\delta_x^* \right] = \mathcal{L} \left[k(\cdot, x) \right], \quad \text{and} \tag{A.218}$$

$$\mathcal{K} \mathcal{L}^* \left[\cdot \right] (x) = \left\langle \mathcal{L} \left[k(x, \cdot) \right], \cdot \right\rangle_{\mathcal{H}_\mathcal{L}}, \tag{A.219}$$

where $\delta_x^* \in \mathcal{H}$ such that $h(x) = \delta_x(h) = \left\langle \delta_x^*, h \right\rangle_\mathcal{H}$ for $h \in \mathcal{H}$.

Proof. We have

$$\mathcal{K}\left[\delta_{x_1}^*\right](x_2) = \left\langle k(x_2,\cdot), \delta_{x_1}^*\right\rangle_{\mathcal{H}} = k(x_2,x_1) \tag{A.220}$$

for $x_1, x_2 \in \mathcal{X}$. Hence, $\mathcal{L}\mathcal{K}\left[\delta_x^*\right] = \mathcal{L}\left[\mathcal{K}\left[\delta_x^*\right]\right] = \mathcal{L}\left[k(\cdot,x)\right]$. Moreover, for $h \in \mathcal{H}_{\mathcal{L}}$, it holds that

$$\begin{aligned}\mathcal{K}\mathcal{L}^*\left[h\right](x) &= \mathcal{K}\left[\mathcal{L}^*\left[h\right]\right](x) && \text{(A.221)}\\ &= \left\langle k(x,\cdot), \mathcal{L}^*\left[h\right]\right\rangle_{\mathcal{H}_{\mathcal{L}}} && \text{(A.222)}\\ &= \left\langle \mathcal{L}\left[k(x,\cdot)\right], h\right\rangle_{\mathcal{H}_{\mathcal{L}}}. && \text{(A.223)}\end{aligned}$$

□

A.8. Theorem 1 and its Corollaries

With all the above, theorem 1 is now merely a corollary of previous results, particularly propositions A.1 and A.5 and corollary A.7. Nevertheless, to summarize the results of this appendix, we formulate it here as a separate theorem and consider the entirety of appendix A as its proof. This also offers a unique interface for practitioners wishing to ground their GP inference algorithms in our theoretical evidence. To this end, we formulate two corollaries, which, together with theorem 1, provide the theoretical basis for most GP inference performed in practice. All three results share a common set of assumptions.

Assumption 1. *Let*

$$f \sim \mathcal{GP}\left(m_f, k_f\right) \tag{4.10}$$

be a Gaussian process prior with index set $\mathcal{X}$ on the Borel probability space $\left(\Omega, \mathcal{B}\left(\Omega\right), \mathrm{P}\right)$, whose mean function and sample paths lie in a real separable Hilbert function space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with $\mathcal{H}_k \subset \mathcal{H}$ and with continuous point evaluation functionals. Let $\mathcal{L}\colon \mathcal{H} \to \mathcal{H}_{\mathcal{L}}$ be a bounded linear operator mapping the paths of f into a separable Hilbert space $\mathcal{H}_{\mathcal{L}}$.

In the most general case, the linear operator $\mathcal{L}$ maps into a space, which is either not a function space or a function space on which point evaluation is not a continuous functional. This happens for instance when applying the differential operator of highest possible order on a Sobolev path space, since then the resulting object will be an L_2 function, which is not pointwise defined.

Theorem 1 (Affine Gaussian Process Inference)**.** *Let assumption 1 hold. Then $\omega \mapsto f(\cdot,\omega)$ is an $\mathcal{H}$-valued Gaussian random variable on $\left(\Omega, \mathcal{B}\left(\Omega\right), \mathrm{P}\right)$ with mean m and covariance operator $h \mapsto \mathcal{C}_f\left[h\right](x) = \left\langle k(x,\cdot), h\right\rangle_{\mathcal{H}}$. We also write $f \sim \mathcal{N}\left(m, \mathcal{C}_f\right)$. Let $g \sim \mathcal{N}\left(m_g, \mathcal{C}_g\right)$ be an $\mathcal{H}_{\mathcal{L}}$-valued Gaussian random variable on $\left(\Omega, \mathcal{B}\left(\Omega\right), \mathrm{P}\right)$ with $g \perp\!\!\!\perp f$. Then*

$$\begin{pmatrix} f \\ \mathcal{L}\left[f\right] + g \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_f \\ \mathcal{L}\left[m_f\right] + m_g \end{pmatrix}, \begin{pmatrix} \mathcal{C}_f & \mathcal{C}_f\mathcal{L}^* \\ \mathcal{L}\mathcal{C}_f & \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g \end{pmatrix}\right), \tag{4.11}$$

with values in $\mathcal{H} \times \mathcal{H}_{\mathcal{L}}$ *and hence*

$$\mathcal{L}[f] + g \sim \mathcal{N}(\mathcal{L}[m_f] + m_g, \mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g). \tag{4.12}$$

If $\operatorname{ran}(\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)$ *is closed, then for all* $h \in \mathcal{H}_{\mathcal{L}}$

$$f \mid \mathcal{L}[f] + g = h \sim \mathcal{GP}\left(m_{f|\mathcal{L}[f]+g=h}, k_{f|\mathcal{L}[f]+g=h}\right), \tag{4.13}$$

where the conditional mean and covariance function are given by

$$m_{f|\mathcal{L}[f]+g=h(x)} = m_f(x) + \left\langle \mathcal{L}[k_f(\cdot, x)], (\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)^\dagger [h - (\mathcal{L}[m_f] + m_g)] \right\rangle_{\mathcal{H}_{\mathcal{L}}}, \tag{4.14}$$

and

$$k_{f|\mathcal{L}[f]+g=h(x_1,x_2)} = k_f(x_1, x_2) - \left\langle \mathcal{L}[k_f(\cdot, x_1)], (\mathcal{L}\mathcal{C}_f\mathcal{L}^* + \mathcal{C}_g)^\dagger \mathcal{L}[k_f(\cdot, x_2)] \right\rangle_{\mathcal{H}_{\mathcal{L}}}, \tag{4.15}$$

respectively.

The first corollary deals with the case, where we observe the GP through a finite number of linear functionals. This happens when conditioning on integral observations or on (Galerkin) projections as in chapter 5.

Corollary 1. *Let assumption 1 hold for* $\mathcal{H}_{\mathcal{L}} = \mathbb{R}^n$ *and let* $g \sim \mathcal{N}(\mu_g, \Sigma_g)$ *be an* $\mathbb{R}^n$*-valued Gaussian random variable on* $(\Omega, \mathcal{B}(\Omega), \mathrm{P})$ *with* $g \perp\!\!\!\perp f$*. Then*

$$\mathcal{L}[f] + g \sim \mathcal{N}(\mathcal{L}[m_f] + \mu_g, \mathcal{L}k_f\mathcal{L}^* + \Sigma_g) \tag{4.21}$$

and

$$f \mid \mathcal{L}[f] + g = h \sim \mathcal{GP}\left(m_{f|\mathcal{L}[f]+g=h}, k_{f|\mathcal{L}[f]+g=h}\right), \tag{4.22}$$

with conditional mean and covariance function given by

$$m_{f|\mathcal{L}[f]+g=h}(x) = m_f(x) + \mathcal{L}[k_f(x, \cdot)]^\top (\mathcal{L}k_f\mathcal{L}^* + \Sigma_g)^\dagger (h - (\mathcal{L}[m_f] + m_g)), \tag{4.23}$$

and

$$k_{f|\mathcal{L}[f]+g=h}(x_1, x_2) = k_f(x_1, x_2) - \mathcal{L}[k_f(x_1, \cdot)]^\top (\mathcal{L}k_f\mathcal{L}^* + \Sigma_g)^\dagger \mathcal{L}[k_f(\cdot, x_2)]. \tag{4.24}$$

Finally, we address the archetypical case, in which both the prior f and the prior predictive $\mathcal{L}[f] + g$ are Gaussian processes. This happens if the linear operator maps into a function space, in which point evaluation is continuous. In this article, this case occurred in chapters 2 and 3, where we inferred the strong solution of a PDE from observations of the PDE residual at a finite number of domain points.

Corollary 2. *Let assumption 1 hold, where $\mathcal{H}_{\mathcal{L}} \subset \mathbb{R}^{\mathcal{X}'}$ is a space of real valued functions defined on $\mathcal{X}'$ such that the point evaluation functionals $\delta_{x'} \colon \mathcal{H}_{\mathcal{L}} \to \mathbb{R}, h \mapsto h(x)$ for all $x \in \mathcal{X}'$ are continuous. Let*

$$g \sim \mathcal{GP}\left(m_g, k_g\right) \tag{4.28}$$

be a Gaussian process with index set $\mathcal{X}'$ on $\left(\Omega, \mathcal{B}\left(\Omega\right), \mathrm{P}\right)$ with $g \perp\!\!\!\perp f$. Then

$$\mathcal{L}\left[f\right] + g \sim \mathcal{GP}\left(\mathcal{L}\left[m\right] + m_g, \mathcal{L}k_f\mathcal{L}^* + k_g\right), \tag{4.29}$$

and, for $X' = \{x'_i\}_{i=1}^n \subset \mathcal{X}'$ and $h \in \mathbb{R}^n$,

$$f \mid \mathcal{L}\left[f\right]\left(X'\right) + g(X') = h \sim \mathcal{GP}\left(m_{f|X',h}, k_{f|X',h}\right) \tag{4.30}$$

with

$$m_{f|X',h}(x) \coloneqq m_f(x) + (k_f\mathcal{L}^*)(x, X')(\mathcal{L}k_f\mathcal{L}^* + k_g)(X', X')^\dagger \left(h - \left(\mathcal{L}\left[m_f\right]\left(X'\right) + m_g(X')\right)\right) \tag{4.31}$$

and

$$k_{f|X',h}(x_1, x_2) \coloneqq k_f(x_1, x_2) - (k_f\mathcal{L}^*)(x_1, X')(\mathcal{L}k_f\mathcal{L}^* + k_g)(X', X')^\dagger(\mathcal{L}k_f)(X', x_2). \tag{4.32}$$

where

$$(k_f\mathcal{L}^*)(x, X') = \left((k_f\mathcal{L}^*)(x, x'_i)\right)_{i=1}^n \in \mathbb{R}^{1\times n} \tag{4.33}$$

$$(\mathcal{L}k_f)(X', x_2) = \left((\mathcal{L}k_f)(x'_i, x)\right)_{i=1}^n \in \mathbb{R}^n \tag{4.34}$$

$$(\mathcal{L}k_f\mathcal{L}^* + k_g)(X', X') = \left((\mathcal{L}k_f\mathcal{L}^*)(x'_i, x'_j) + k_g(x'_i, x'_j)\right)_{i,j=1}^n \in \mathbb{R}^{n\times n} \tag{4.35}$$

$$\mathcal{L}\left[m_f\right]\left(X'\right) = \left(\mathcal{L}\left[m_f\right](x_i)\right)_{i=1}^n \in \mathbb{R}^n \tag{4.36}$$

$$m_g(X') = \left(m_g(x_i)\right)_{i=1}^n \in \mathbb{R}^n. \tag{4.37}$$

If additionally $\mathcal{X} = \mathcal{X}'$, then

$$\begin{pmatrix} f \\ \mathcal{L}\left[f\right] + g \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m_f \\ \mathcal{L}\left[m_f\right] + m_g \end{pmatrix}, \begin{pmatrix} k_f & k_f\mathcal{L}^* \\ \mathcal{L}k_f & \mathcal{L}k_f\mathcal{L}^* + k_g \end{pmatrix}\right). \tag{4.38}$$

B. Linear Partial Differential Equations

Definition B.1 (Multi-index)**.** *Using a d-dimensional* multi-index $\alpha \in \mathbb{N}_0^d$*, we can represent (mixed) partial derivatives of arbitrary order as*

$$\frac{\partial^{|\alpha|}}{\partial x^\alpha} \coloneqq \frac{\partial^{|\alpha|}}{\partial x_1^{(\alpha_1)} \dots \partial x_d^{(\alpha_d)}}, \tag{B.1}$$

where $|\alpha| \coloneqq \sum_{i=1}^d \alpha_i$*. If the variables w.r.t. which we differentiate are clear from the context, we also denote this (mixed) partial derivative by* D^α*.*

Definition B.2 (Linear differential operator)**.** *A* linear differential operator $\mathcal{D}\colon U \to V$ *of order* k *between spaces* U, V *of real-valued functions defined on some domain* $\Omega \subset \mathbb{R}^d$ *is a linear operator that linearly combines partial derivatives up to* k*-th order of its input function, i.e.*

$$\mathcal{D}[u] \coloneqq \sum_{\alpha \in \mathbb{N}_0^d, |\alpha| \leq k} A_\alpha D^\alpha u, \tag{B.2}$$

where $A_\alpha \in \mathbb{R}$ *for every multi-index* α*.*

Definition B.3 (Heat equation [Lienhard and Lienhard, 2020, Evans, 2010])**.** *Let* $\Omega \subset \mathbb{R}^d$ *be an open and bounded region and* $T > 0$*. The* heat equation *is given by*

$$\rho c_p \frac{\partial u}{\partial t} - \operatorname{div}(k \nabla u) = \dot{q}_V, \tag{B.3}$$

where $k \in \mathbb{R}^{d \times d}$*,* $\rho, c_p, k_{ij} \in L_\infty(\Omega \times (0,T])$*, and* $\dot{q}_V \in L_2(\Omega \times (0,T])$*.*

Definition B.4 (Elliptic PDE in nondivergence form)**.** *Let* $\Omega \subset \mathbb{R}^d$ *be an open and bounded region. The equation*

$$-\operatorname{div}(A \nabla u) + b^T \nabla u + cu = f, \tag{B.4}$$

where $A_{ij}, b_i, c \in L_\infty(\Omega)$ *and* $f \in L_2(\Omega)$*.*

B.1. Weak Derivatives and Sobolev Spaces

Definition B.5 (Test Function)**.** *Let* $D \subset \mathbb{R}^d$ *be open and let*

$$C_c^\infty(D) \coloneqq \{\phi \in C^\infty(D, \mathbb{R}) \mid \operatorname{supp}(\phi) \subset U \text{ is compact}\} \tag{B.5}$$

be the space of smooth functions with compact support in D*. A function* $\phi \in C_c^\infty(D)$ *is dubbed* test function *and we refer to* $C_c^\infty(D)$ *as the* space of test functions*.*

Theorem B.1 (Sobolev Spaces[1]). *Let $D \subset \mathbb{R}^d$ be open, $m \in \mathbb{N}_{>0}$, and $p \in [1, \infty) \cup \{\infty\}$. The functional*

$$\|u\|_{m,p,D} \coloneqq \begin{cases} \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^p_{L_p(D)}\right)^{1/p} & \text{if } p < \infty, \\ \max_{|\alpha| \leq m} \|D^\alpha u\|_{L_\infty(D)} & \text{if } p = \infty. \end{cases} \tag{B.6}$$

is called a Sobolev norm. *A Sobolev norm $\|u\|_{m,p,D}$ is a norm on subspaces of $L_p(D)$, on which the right-hand side is well-defined and finite. A* Sobolev space of order m *is defined as the subspace*

$$W^{m,p}(D) \coloneqq \{u \in L_p(D) \mid D^\alpha u \in L_p(D) \text{ for } |\alpha| \leq m\}. \tag{B.7}$$

of L_p, where the D^α are weak partial derivatives. Sobolev spaces $W^{m,p}(D)$ are Banach spaces under the Sobolev norm $\|\cdot\|_{m,p}$. The Sobolev space $H^m(D) \coloneqq W^{2,m}(D)$ is a separable Hilbert space with inner product

$$\langle u_1, u_2 \rangle_{m,D} \coloneqq \sum_{|\alpha| \leq m} \langle D^\alpha u_1, D^\alpha u_2 \rangle_{L_2(D)} \tag{B.8}$$

and norm

$$\|\cdot\|_{m,D} \coloneqq \sqrt{\langle \cdot, \cdot \rangle_{m,D}} = \|\cdot\|_{m,2,D}\,. \tag{B.9}$$

[1] This theorem is a summary of [Adams and Fournier, 2003, Definitions 3.1 and 3.2 and Theorems 3.3 and 3.6]

Bibliography

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Elsevier, 2nd edition, 2003. ISBN 9780080541297.

Christian Agrell. Gaussian processes with linear operator inequality constraints. *Journal of Machine Learning Research*, 20(135):1–36, 2019. URL http://jmlr.org/papers/v20/19-065.html.

Christopher G. Albert. Gaussian processes for data fulfilling linear differential equations. *Proceedings of the 39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 33(1), 2019. ISSN 2504-3900. doi:10.3390/proceedings2019033005.

W. N. Anderson, Jr. and G. E. Trapp. Shorted operators. II. *SIAM Journal on Applied Mathematics*, 28(1):60–71, 1975. doi:10.1137/0128007.

Adi Ben-Israel and Thomas N.E. Greville. *Generalized Inverses: Theory and Applications*. CMS Books in Mathematics. Springer, New York, 2nd edition, 2003. ISBN 978-0-387-21634-8. doi:10.1007/b97366.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, first edition, 2004. ISBN 978-1-4613-4792-7. doi:10.1007/978-1-4419-9096-9.

S. J. Bernau. The square root of a positive self-adjoint operator. *Journal of The Australian Mathematical Society*, 8(1):17–36, February 1968. doi:10.1017/S1446788700004560.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, first edition, 2006. ISBN 978-0387-31073-2.

David Borthwick. *Introduction to Partial Differential Equations*. Universitext. Springer, first edition, 2018. ISBN 978-3-319-48936-0. doi:10.1007/978-3-319-48936-0.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 9780511804441. doi:https://doi.org/10.1017/CBO9780511804441.

Christian Clason. Regularization of inverse problems. Lecture Notes, February 2021. URL https://arxiv.org/abs/2001.00617.

Jon Cockayne, Chris Oates, Tim Sullivan, and Mark Girolami. Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. In Geert Verdoolaege, editor, *Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1853 of *AIP Conference Proceedings*, pages 060001–1 – 060001–8, 2017. doi:10.1063/1.4985359.

Jon Cockayne, Chris J. Oates, T. J. Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019. doi:10.1137/17M1139357.

Wolfgang Demtröder. *Experimentalphysik 2: Elektrizität und Optik*. Springer-Lehrbuch. Springer Spektrum, sixth edition, 2013. ISBN 978-3-642-29944-5. doi:10.1007/978-3-642-29944-5.

Wolfgang Demtröder. *Experimentalphysik 1: Mechanik und Wärme*. Springer-Lehrbuch. Springer Spektrum, seventh edition, 2015. ISBN 978-3-662-46415-1. doi:10.1007/978-3-662-46415-1.

Kris Dumont, Jan Vierendeels, Rado Kaminsky, Guido van Nooten, Pascal Verdonck, and Danny Bluestein. Comparison of the hemodynamic and thrombogenic performance of two bileaflet mechanical heart valves using a CFD/FSI model. *Journal of Biomechanical Engineering*, 129(4):558–565, August 2007. doi:10.1115/1.2746378.

Lawrence C. Evans. *Partial Differential Equations: Second Edition*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2nd edition, 2010. ISBN 978-0-82-184974-3. URL `https://bookstore.ams.org/gsm-19-r`.

Gregory E. Fasshauer. Solving partial differential equations by collocation with radial basis functions. In Alain Le Méhauté, Christophe Rabut, and Larry L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 131–138. Vanderbilt University Press, Nashville, TN, 1997. ISBN 9780826512949.

Gregory E. Fasshauer. Solving differential equations with radial basis functions: multilevel methods and smoothing. *Advances in Computational Mathematics*, 11:139–159, November 1999. doi:10.1023/A:1018919824891.

C. A. J. Fletcher. *Computational Galerkin Methods*. Scientific Computation. Springer, Berlin, Heidelberg, 1 edition, 1984. ISBN 978-3-642-85949-6. doi:10.1007/978-3-642-85949-6.

Mark Girolami, Eky Febrianto, Yin Ge, and Fehmi Cirak. The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions. *Computer Methods in Applied Mechanics and Engineering*, 275:113533, 2021. doi:10.1016/j.cma.2020.113533.

Thore Graepel. Solving noisy linear operator equations by Gaussian processes: Application to ordinary and partial differential equations. In *Proceedings of the 20th International Conference on Machine Learning*, pages 234–241. AAAI Press, 2003.

Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A*, 471(2179), 2015. doi:10.1098/rspa.2015.0142.

Philipp Hennig, Michael A. Osborne, and Hans P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, June 2022. ISBN 9781316681411. doi:10.1017/9781316681411.

David S. Holder, editor. *Electrical Impedance Tomography: Methods, History and Applications*. Institute of Physics Medical Physics Series. Institute of Physics Publishing, Bristol, 2005. ISBN 0750309520.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Nicholas Krämer, Jonathan Schmidt, and Philipp Hennig. Probabilistic numerical method of lines for time-dependent partial differential equations. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 625–639. PMLR, 2022. URL https://proceedings.mlr.press/v151/kramer22a.html.

John H. Lienhard, IV and John H. Lienhard, V. *A Heat Transfer Textbook*. Phlogiston Press, Cambridge, MA, 5th edition, 2020. URL http://ahtt.mit.edu.

Stefania Maniglia and Abdelaziz Rhandi. Gaussian measures on separable Hilbert spaces and applications, January 2004.

Pierre Michaud. A simple model of processor temperature for deterministic turbo clock frequency. resreport RR-9308, Inria Rennes, 2019. URL https://hal.inria.fr/hal-02391970.

Claus Müller. *Spherical Harmonics*, volume 17 of *Lecture Notes in Mathematics*. Springer, Berlin, 1st edition, 1966. ISBN 9783540036005. doi:10.1007/BFb0094775.

Chris J. Oates and Tim J. Sullivan. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29:1335–1351, 2019. doi:10.1007/s11222-019-09902-z.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, 6th edition, 2003. ISBN 978-3-642-14394-6. doi:10.1007/978-3-642-14394-6.

Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015. doi:10.1137/140974596.

Houman Owhadi and Clint Scovel. Conditioning Gaussian measure on Hilbert space. *Journal of Mathematical and Statistical Analysis*, 1(109), 2018.

Houman Owhadi, Clint Scovel, and Florian Schäfer. Statistical numerical approximation. *Notices of the American Mathematical Society*, 66(10):1608–1617, 2019. doi:10.1090/noti1963.

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. ISSN 0021-9991. doi:10.1016/j.jcp.2017.07.050.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, London, England, 2006. ISBN 026218253X.

Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, second edition, 1991. ISBN 978-0-07-054236-5.

Simo Särkkä. Linear operators and stochastic partial differential equations in Gaussian process regression. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 151–158, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. doi:10.1007/978-3-642-21738-8_20.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*, volume 10 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2019. ISBN 9781316510087. doi:10.1017/9781108186735.

Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013. doi:10.1109/MSP.2013.2246292.

Ingo Steinwart. Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51:361–395, 2019. doi:10.1007/s11118-018-9715-5.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417, 2012. doi:10.1007/s00365-012-9153-3.

Filip Tronarp, Hans Kersting, Simo Särkkä, and Philipp Hennig. Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: A new perspective. *Statistics and Computing*, 29:1297–1315, 2019. doi:10.1007/s11222-019-09900-1.

Bastian von Harrach. Numerik partieller differentialgleichungen. Lecture Notes, 2021. URL `https://www.math.uni-frankfurt.de/~harrach/lehre/Numerik_PDGL.pdf`.

Junyang Wang, Jon Cockayne, Oksana Chkrebtii, Tim J. Sullivan, and Chris J. Oates. Bayesian numerical methods for nonlinear partial differential equations. *Statistics and Computing*, 31(55), 2021. doi:10.1007/s11222-021-10030-w.

Kôsaku Yosida. *Functional Analysis*, volume 123 of *Classics in Mathematics*. Springer, 6th edition, 1995. ISBN 978-3-540-58654-8. doi:10.1007/978-3-642-61859-8.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Tübingen, 15.07.2022

Ort, Datum	Unterschrift