Efficient Gaussian process updating under linear operator data for uncertainty reduction on implicit sets in Bayesian inverse problems

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Cédric Travelletti
from **Ayent**

Supervisor of the doctoral thesis:
Prof. Dr. David Ginsbourger

Institut für Mathematische Statistik und Versicherungslehre
University of Bern
Switzerland

Efficient Gaussian process updating under linear operator data for uncertainty reduction on implicit sets in Bayesian inverse problems

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Cédric Travelletti
from **Ayent**

Supervisor of the doctoral thesis:
Prof. Dr. David Ginsbourger

Institut für Mathematische Statistik und Versicherungslehre
University of Bern
Switzerland

Accepted by the Faculty of Science

Bern, 02.06.2023

The Dean
Prof. Dr. Marco Herwegh

Efficient Gaussian process updating under linear operator data for uncertainty reduction on implicit sets in Bayesian inverse problems

Cédric Travelletti

Abstract

This thesis aims at developing sequential uncertainty reduction techniques for set estimation in Bayesian inverse problems. Sequential uncertainty reduction (SUR) strategies provide a statistically principled way of designing data collection plans that optimally reduce the uncertainty on a given quantity of interest. This thesis focusses on settings where the quantity of interest is a set that is implicitly defined by conditions on some unknown function and one is only able to observe the values of linear operators applied to the function. This setting corresponds to the one encoutered in linear inverse problems and proves to be challenging for SUR techniques. Indeed, SUR relies on having a probabilistic model for the unknown function under consideration, and these models become untractable for moderately sized problem. We start by introducing an implicit representation for covariance matrices of Gaussian processes (GP) to overcome this limitation, and demonstrate how it allows one to perform SUR for excursion set estimation in a real-world 3D gravimetric inversion problem on the Stromboli volcano. In a second time, we focus on extending vanilly SUR to multivariate problems. To that end, we introduce the concept of 'generalized locations', which allows us to rewrite the co-kriging equations in a form-invariant way and to derive semi-analytical formulae for multivariate SUR criteria. Those approaches are demonstrated on a river plume estimation problem. After having extended SUR for inverse problems to large-scale and multivariate settings, we devote our attention to improving the realism of the models by including user-defined trends. We show how this can be done by extending universal kriging to inverse problems and also provide fast k-fold cross-validation formulae. Finally, in order to provide theoretical footing for the developed approaches, show how the conditional law of a GP can be seen as a disintegration of a corresponding Gaussian measure under some suitable condition.

Acknowledgements

Since this thesis has been completed in a linguistic environment that was constantly straddling the border between French and English, the following acknowledgements will also perform a similar stunt.

Let us start with the English part. First I would like to thank Niklas Linde for his dedication to the project underpinning this thesis. Niklas has been following this thesis from its inception and his insightful comments on the practical aspects of the methods developed hereafter have proven invaluable. When I lost sight of the big picture and got trapped in theoretical detours, Niklas always insisted that our work should stay in touch with the real-world applications. If this thesis is of any practical value, I owe it to him.
I would also like to thank the other members of the Linde research group at UniL: Lea, Macarena, Shiran and Guillaume. Thanks for having been there along the trip. I am grateful for the many comments that you formulated during our frequent meetings and whish you all the best for your future career.
I would also like to thank Wolfgang Polonik for having followed this thesis during 4 years and acted as a theoretical advisor. I hope that we will able to collaborate on the applications of this thesis to filament estimation in the future.

Next, I would like to thank the various people that have welcomed me in their research group for academic stays.
A big thank you to Jo Eidsvik, Trygve Fossum, Yaolin Ge and Martin Outzen Berild for hosting me at NTNU Trondheim during the troubled times of Covid. I tremendously enjoyed my stay there and was able to grow as a researcher and human being. I will fondly remember your kindness and hope that I will be able to keep a "Norwegian touch" in the way I lead my life.
I would also like to express my gratitude to Peter Frazier, Poompol Buathong, Yujia Zhang, Jiayue Wan and Raul Astudillo for their warm welcome during a research stay at Cornell University. I greatly appreciated your dedication to help come up with new ideas for path planning in sequential uncertainty reduction strategies. Thank you for having taken so much of your time to help me progress on that front, even if this research direction is not yet ready to be presented in this thesis. Thanks Poompol for having been my guide around Cornell and thank you Yujia and Jiayue for the trip to Niagara falls. And again thank you Peter for having offered me the opportunity to discover the US academic systemand the USA in general, I really think that this will help me in the future.

I would like to thank Olivier Roustant for helpful feedback and comments and Sébastien Petit for pointing out relevant references for this thesis. I would also

like to thank the people that organized the various conferences and research events in which I participated and also the people I met there: you have enriched my intellectual horizon.

Finally, I extend my gratitude to Arnaud Doucet. Thank you Arnaud for having acted as a referee for this thesis. Your comments have helped my improve my presentation and increased the quality of this work.

Et maintenant pour les remerciements en français:

Tout d'abord un grand merci à mon superviseur de thèse: David Ginsbourger. Merci d'avoir cru en moi et de m'avoir donné une chance, malgré mon parcours en zigzag. Merci pour les innombrables fou rires, les découvertes culinaires, les débats philosophiques endiablés et surtout pour m'avoir transmis (une partie) de ton savoir scientifique. J'ai énormément appris de tes idées et leur originalité me donne aujourd'hui un point de vue intellectuel que je n'aurais pu acquérir nulle part ailleurs. Merci aussi pour ton humanité tout au long de ces 4 anées et demi: tu m'as toujours soutenu dans les moments à vide, et par delà les quelques incompréhensions que nous pouvons avoir tu t'es constamment battu pour que cette thèse soit un succès. Merci encore.
Merci à ma "jumelle de doctorat", Athénaïs Gautier. Ce fut un plaisir de partager ces années avec toi et je n'ose pas imaginer ce qu'elles auraient été si tu n'avais pas été là. Merci pour les délires, les conférences rock n'roll, les excès de Chartreuse, les bains dans l'Isère et les soirés bernoises.

Un grand merci à mes parents Barbara et Raoul, à mon frère David ainsi qu'à toute ma famille. Comme je l'ai déjà dit, sans eux je serais incapable de me gérer, et donc encore moins de faire un doctorat. Merci pour m'avoir nourri et blanchi durant la plus grand partie de cette thèse :) et surtout merci pour votre soutien sans faille tout au long de mon interminable parcours académique.

Merci à mes amis, les récents et ceux de toujours. Merci Guillaume, mon frère scientifique depuis tant d'années. Bien que tu aies choisi la voie de la Vérité, c'est toujours un plaisir incomparable de délirer avec toi, que ce sur soit sur les maths, la philo ou des choses moins avouables. Merci pour les grand moments d'extrémisme et toutes ces soirés où nous avons été, parfois au risque de nos vies, chercher la Limite. Enfin merci pour l'inspiration sans cesse renouvelée que me donnent tes tribulations. Merci à Mathieu: plus qu'un ami, un rival. Merci pour ton infinie culture et pour ta façon d'acquiescer à la Vie. Ce fut un privilège de te compter parmi mes amis pendant ces années de thèse et ton amitié m'a toujouts tiré vers le haut.
Merci aux Siffleurs: Émile, Gabi et Wilson. Merci aux Timbrés: Elisa et Julot. Merci à Joseph pour les meilleurs gags que j'aie entendus ici bas. Merci à Noémie et Pauline, à Bastien, à Maxime pour les voyages incomparables, à Gaëtan, Simon, Jordan, à Célien pour la grimpe, à Yann pour *incantation à la neige*. Merci à Sarah pour l'inspiration dans le dernier sprint de la rédaction de ce travail.

Enfin merci au monde de la nuit, qui m'as permis de garder les pieds sur terre et de rester humain. Merci à la station d'Anzère et à la Jeunesse d'Ayent. Merci au Zazabar et à ceux qui sont son âme: Zaza et Léo. Même si ça peut paraître improb-

able, ce lieu à joué un rôle déterminant dans le travail qui suit. Merci à Morgane et Antho pour leur sourire et pour m'avoir (trop de fois?) supporté jusqu'à des heures tardives. Et merci à tous ceux que j'ai pu rencontrer lors de ces nuits de débauche.

Finalement, merci à toutes ces petites choses qui rendent la vie plus douce. Sans ordre précis: Blink 182, Sergio Leone, le Rawyl et tout le calcaire en général, Rebatet, The Doors, la Chartreuse, le Gamay, l'Itre du Bouis. Et surtout, merci à mon village: Ayent.

Contents

List of Figures

List of Tables

List of Symbols

General Inverse Problems:

x	a generic spatial location
ρ	the unknown function to recover in an inverse problem
D	the domain of the problem
G	the forward operator
w	one generic point involved in the forward
$\boldsymbol{W}$	points involved in the forward
r	dimension of the forward (number of points involved)
$\mathbb{F}, \mathbb{Y}$	model and data spaces
$\mathcal{A}, \mathcal{C}$	sigma algebras
$\boldsymbol{Y}$	the random version of the data vector
y	a given observed realization of the data vector
q	dimension of the data vector

Chapter 3:

ℓ	a linear form
f	a generic element of the Banach space $\mathbb{F}$
Z	a Gaussian process with sample paths in $\mathbb{F}$
μ_Z	the associated Gaussian measure on $\mathbb{F}$

Chapter 4:

Γ^*	the excursion set to be estimated
T	the excursion threshold
Γ	a random excursion set
n	current data collection stage (sequential strategies)
i	dummy stage index (sequential strategies)
u	visited sampling locations (sequential strategies)

Chapter 5:

η	a centred GP (fluctuations)
f	a trend function
$F_{\boldsymbol{W}}$	matrix of trend functions evaluated at $\boldsymbol{W}$
β	trend coefficient
$\boldsymbol{\beta}_{\text{prior}}, \boldsymbol{\Sigma}$	prior mean and covariance (trend coefficients)
$\boldsymbol{i}$	list of left-out data indices (cross-validation)
$-\boldsymbol{i}$	list of the remaining data indices
$\boldsymbol{E_i}$	random cross-validation residual when predicting data subset $\boldsymbol{i}$
$\boldsymbol{e_i}$	observed realization of $\boldsymbol{E_i}$
$\tilde{K}$	cross-validation helper matrix

Chapter 6:

p	dimension of the response/GP
i	a generic response index
χ	a generic generalized location
$\boldsymbol{\chi}$	a batch of generalized locations
$\boldsymbol{\xi}$	generalized locations involved in the forward
Φ	the standard Gaussian cdf

Chapter 1

Introduction

This thesis focuses on sequential estimation of implicit sets in Bayesian inverse problems. Broadly speaking, an inverse problem is the task of recovering some unknown function describing some physical quantity from indirect observations thereof. Such problems arise in various areas of the natural sciences as well as in machine learning. In this thesis, we focus on situations where one is not interested in reconstructing the unknown physical phenomenon in itself, but would rather like to estimate regions that are implicitly characterized by properties thereof. These can include, for example, regions where the unknown function takes high values (excursion sets) or regions of high curvature (transition regions). Our goal is to propose adaptive data collection plans that leverage previous knowledge about the physical process at hand to sequentially collect new data and adapt the future data gathering to optimally decrease the uncertainty about the target region. Given the current state of the art in Bayesian inverse problems, such a task is fraught with computational challenges and Chapter 4 is devoted to overcoming these. Then, Chapter 5 shows how prior domain-specific knowledge can be used to produce more realistic models. Finally, Chapter 6 extends our framework to multivariate physical phenomena. Most of the approaches developed in this thesis rely on theoretical developments introduced in Chapter 3. We next give a detailed summary of our contributions for each chapter.

Chapter 3: Our first contribution is the construction of a theoretical framework for the updating of Gaussian processes under linear operator data. Such updating situations arise frequently in Bayesian inverse problems, which is the core motivation for the work done in this chapter. Although update formulae for GPs have been known for some time (Chevalier et al., 2014b; Emery, 2009; Gao et al., 1996; Barnes and Watson, 1992), these approaches silently take various properties of the conditional distribution of the process for granted, which is of minor harm in the case of pointwise observations but can be problematic in the presence of linear operator data.

To enable a rigorous treatment of the conditional law under linear operator observations, we build upon the theory of disintegrations of Gaussian measures. In passing, we clarify some links between Gaussian processes and Gaussian measures and provide characterizations of the type linear operators that can be assimilated for a given GP prior. Overall, our results provide an extension of GP update formulae to disintegrations as well as a purely functional formulation of the Bayesian assimilation/inversion process. This is leveraged in Chapter 4 to develop fast update formulae and can be used for further theoretical inquiries in Bayesian inversion.

Chapter 4: Even though GP priors have been shown to perform well on smaller-scale inverse problems, difficulties arise when one tries to apply them to large inverse problems (be it high dimensional or high resolution) and these get worse when one considers sequential data assimilation settings such as in Chevalier et al. (2014a). The main goal of Chapter 4 is to overcome these difficulties and to provide solutions for scaling GP priors to large-scale Bayesian inverse problems.

Specifically, this chapter focuses on a triple intersection of i) linear operator data, ii) large number of prediction points and iii) sequential data acquisition. There exists related works that focus on some of these items individually. For example, methods for extending Gaussian processes to large datasets (Hensman et al., 2013; Wang et al., 2019) or to a large number of prediction points (Wilson et al., 2020) gained a lot of attention over the last years. Also, much work has been devoted to extending GP regression to include linear constraints (Jidling et al., 2017) or integral observations (Hendriks et al., 2018; Jidling et al., 2019). On the sequential side, methods have been developed relying on infinite-dimensional state-space representations (Särkkä et al., 2013) and have also been extended to variational GPs (Hamelijnck et al., 2021). There are also works that focus on the three aspects at the same time (Solin et al., 2015). All these approaches rely on approximations of the covariance.

The topic of large-scale sequential assimilation of linear operator data has also been of central interest in the Kalman filter community. To the best of our knowledge, techniques employed in this framework usually rely on a low rank representation of the covariance matrix, obtained either via factorization (Kitanidis, 2015) or from an ensemble estimate (Mandel, 2006).

Our contribution in Chapter 4 is the elaboration of update methods for Gaussian processes that do not rely on a particular factorization of the covariance matrix, nor on an approximation scheme. This is achieved through the introduction of an implicit representation of the posterior covariance matrix that builds on theoretical results from Chapter 3. This new way of looking at the posterior covariance allows us to perform Bayesian inversion in large-scale settings, as well as applying sequential uncertainty reduction strategies in this setting. All our techniques are demonstrated on a real-world gravimetric inverse problem involving field data collected on the Stromboli volcano.

Chapter 5: After having shown in Chapter 4 how Bayesian inversion can be brought to bear on large-scale inverse problems, allowing for the inclusion of expert knowledge in the inversion process through the specification of the prior, we seek in Chapter 5 to extend this framework to priors that enable a more fine-grained expression of pre-existing knowledge about the inversion situation at hand. Indeed, traditional GP models used in Bayesian inversion have a limited degree of flexibility in the choice of the prior, most of which resides in the choice of the covariance kernel, allowing only for control of properties such as regularity and periodicity of the realizations of the model.

In traditional GP regression, more expressive priors have been developed under the framework of *universal kriging* (Matheron, 1969). These allow the user to encode their knowledge about the situation into linear combinations trend functions whose coefficients are learned from the data. In this chapter, we extend this framework to

Bayesian inversion by building on (Kitanidis, 1995) and demonstrate on real-world gravimetric inverse problems how the inclusion of trends in the prior enables users to incorporate their knowledge of the local geology in the inversion process.

Considering more flexible models creates new questions about model selection. In the second part of Chapter 5 we leverage fast k-fold cross-validation formulae (Ginsbourger and Schärer, 2021) to compare different trend models. We provide an overview of the new research question brought forth by the use of cross-validation in Bayesian inversion and present a heuristic study of possible cross-validation approaches for inversion, leaving rigorous theoretical groundwork for future inquiries.

Chapter 6: In this chapter, we present the theoretical side of developments that were achieved as part of a collaboration with NTNU Trondheim (Fossum et al., 2021b) focussing on sequential design strategies for estimating excursion sets of an unknown latent multivariate field. While sequential design of experiments strategies have been well studied in the case of scalar responses, not much has been done to extend these results to situations where the latent field of interest is multivariate, let alone to address the estimation of excursion sets. Though approaches such as co-kriging (see, e.g., Wackernagel, 2003) have been developed to learn multivariate latent functions from vector-valued observations and even if those estimates can be updated efficiently in the context of sequential data assimilation (Vargas-Guzmán and Jim Yeh, 1999), sequential strategies for estimating features of vector-valued random fields are still in their infancy. Le Gratiet et al. (2015) used co-Kriging based sequential designs to multi-fidelity computer codes, Poloczek et al. (2017) used related ideas for multi-information source optimization, but these did not consider excursion sets. To the best of our knowledge, the only works that mentions the possibility of stepwise uncertainty reduction strategies for excursion sets of multivariate functions is the PhD thesis (Stroh, 2018, p.82), yet only under the assumption of independent outputs.

Our goal in Chapter 6 is to develop sequential design strategies for the estimation of excursion sets of multivariate latent field using vector-valued observations. To that end, we start by providing a unified framework for multivariate GP regression from heterogeneous observations and then use it to extend uncertainty reduction criterion to vector-valued cases. Our approaches build on the general sequential design strategies from Ginsbourger (2018) and on the excursion volume estimation strategies from Bect et al. (2012). Using techniques developed in Chevalier et al. (2014a), we provide solutions to make the strategies computationally efficient and adapted to batch observations. When it comes to estimating the excursion sets themselves, rather than their volumes, we build upon the works of French and Sain (2013); Chevalier et al. (2013); Bolin and Lindgren (2015); Azzimonti et al. (2016). We note that our multivariate techniques could be extended to provide conservative estimates (Azzimonti et al., 2021) of multivariate excursion sets, tough this approach is not pursued in this work.
We demonstrate our strategies on a river plume mapping application (see Section 2.4.2) where the goal is to map the river-ocean interface at the mouth of a river, the problem being modeled as an excursion set estimation problem for a bivariate temperature-salinity field. Our strategies prove competitive against other pre-existing strategies in synthetic experiments and a real field test using an autonomous underwater vehicle (AUV) to map a river interface in Trondheim, Norway

was performed by an independent respearch group (Fossum et al., 2021b).

Chapter 2

Bayesian Inversion, Gaussian Processes and Applicative Examples

2.1 Inverse Problems and Deterministic Inversion

The core of this thesis centers around *inverse problems*. Here an inverse problem is the task of recovering some unknown function (or features thereof) $\rho \in \mathbb{F}$ in some abstract function space from the observation of some operator G mapping into a Banach space: $G : \mathbb{F} \to \mathbb{Y}$, applied to ρ. The operator G is called the **forward operator** and the Banach spaces $\mathbb{F}$ and $\mathbb{Y}$ are traditionally called the **model space** and **data space** respectively. The inverse problem is then the task of inverting the relation

$$y = G(\rho) + \boldsymbol{\epsilon}, \tag{2.1}$$

from the observed data y, where $\boldsymbol{\epsilon}$ is some $\mathbb{Y}$-valued random noise. The defining characteristic of inverse problems is their *ill-posedness* (Hadamard, 1902), meaning that the solution of the problem presents one of the following properties: (1) it might fail to exist, (2) it can be non-unique, or (3) it can be sensitive to perturbations in the data. Of all these properties, the third one has generated the most interest and is usually tamed using either some deterministic form of *regularization* (Tikhonov, 1963) or via probabilistic methods (Tarantola et al., 1982; Tarantola, 2005). We here quickly present some deterministic approaches, while probabilistic ones will be introduced in Section 2.2.2 and will be the focus of this thesis.

For the rest of this work, unless mentioned otherwise, we will always assume that the function ρ to be recovered is scalar-valued and at least continuous, so that we have $\rho \in C(D)$ for some domain D, where $C(D)$ denotes the space of real-valued functions on D. One of the first obvious paths towards solving the problem Eq. (2.1) is to discretize the domain and range of the forward operator, so that one is left with a linear algebra task

$$\bar{y} = \bar{G}\ \bar{\rho},\ \bar{G} \in \mathbb{R}^{q \times r}, \tag{2.2}$$

where the overbars are used to denote the vectors and matrices representing the discretized version of the problem. There are then, assuming $\bar{G}$ has maximal rank,

three cases according to the relation between the dimension of the data space q and the dimension of the model space r:

- If $q = r$: then a unique solution exists and is given by the inverse of the forward $\bar{\rho} = \bar{G}^{-1}\bar{y}$.
- If $q < r$: then the problem is underdetermined and the solution is not unique.
- If $q > r$: then the problem is overdetermined and the solution will fail to exist if the data does not lie withing the range of the forward.

In the overdetermined and underdetermined cases, a natural way to obtain a unique (approximate) solution is to look for *least-squares* solutions or *minimum norm* solutions. That is, in the overdetermined case one looks for the model vector that is closest (in Euclidean norm) to solving Eq. (2.2) and in the underdetermined case one looks for the solution with minimum Euclidean norm among all possible solutions. It turns out that all those cases can be subsumed by computing the solution in terms of the Moore-Penrose pseudoinverse $\bar{G}^{+}$ of the forward operator (Ben-Israel and Greville, 2003).

Theorem 1 ((Penrose, 1956))**.** *Let $\bar{G} \in \mathbb{R}^{q\times r}$ be a matrix with maximal rank. Then, for any $\bar{y} \in \mathbb{R}^q$, the vector $z_0 = \bar{G}^{+}\bar{y}$ enjoys the following properties:*

- *If $q \geq r$: $||\bar{G}z - \bar{y}|| \geq ||\bar{G}z_0 - \bar{y}||$, all $z \in \mathbb{R}^r$.*
- *If $q \leq r$: $||z_0|| \leq ||z||$ for all $z \in \mathbb{R}^r$ such that $\bar{G}z = \bar{y}$.*

Even if the above theorem offers a way to compute (approximate) solutions to inverse problems, the issue of sensitivity to the data still remains. Indeed, the computation of the pseudoinverse is numerically unstable for matrices with a large condition number, calling for approaches to *regularize* the solution. The most successful of these is *Tikhonov regularization*, where the idea is to solve the regularized problem:

$$\arg\min_{z\in\mathbb{R}^r} ||\bar{G}z - \bar{y}||^2 + \lambda||z - z_T||_T^2,$$

where z_T is some prespecified vector and $||\cdot||_T$ is some norm on the model space. There exists several algorithms to solve such regularized problems. Since deterministic inversion is not the focus of this thesis, we refer readers to (Hansen, 2010) for more details thereupon. Of significantly more interest to us is the probabilistic interpretation that can be given to Tikhonov regularization, which is explained in the next section.

2.2 Gaussian Processes and Probabilistic Inversion

Aside from the traditional deterministic inversion algorithms, a large literature has been devoted to developing probabilistic approaches to inverse problems, among which the most successful is Bayesian inversion (Tarantola et al., 1982; Stuart, 2010; Dashti and Stuart, 2016). The philosophy of Bayesian inversion is to consider the

solution $\rho : D \to \mathbb{R}$ to the inverse problem as a realization of a random function Z, where our prior knowledge about ρ or (expert knowledge, physical laws, ...) is encoded in the prior distribution of Z. One then uses the conditional distribution of Z under the observed data to approximate the true solution ρ (mathematical details are given in Section 2.2.2).

Bayesian inversion requires one to specify a prior on the set of real-valued functions on the domain D. The most popular class of functional priors is that of Gaussian processes (GP) (Rasmussen and Williams, 2006), owing to the existence of closed-form formulae for their conditional distribution under linear operator observations (Solak et al., 2003; Särkkä, 2011; Jidling et al., 2017; Mandelbaum, 1984; Tarieladze and Vakhania, 2007; Hairer et al., 2005; Owhadi and Scovel, 2015; Klebanov et al., 2021). The goal of the following sections is to provide an introduction to Gaussian processes and their practical use in Bayesian inversion, together with the mathematical theory underpinning the whole construction.

2.2.1 Mathematical Background and Definitions

We start by recalling some basic definitions. Our exposition will be mostly based on (Bovier, 2015) and on (Pavliotis, 2014). We also refer the reader to (Kallenberg, 2021) for a more rigorous exposition.

In the following, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. A couple $(\mathbb{F}, \mathcal{A})$ where the second letter is in calligraphic font will always denote a measurable space, the second member of the couple being the σ-algebra. When talking about measurable mappings, we will liberally use the notation

$$f : (\mathbb{F}, \mathcal{A}) \to (\mathbb{Y}, \mathcal{C})$$

allowing us to specify both the spaces and their respective σ-algebras at once.

We now proceed to define stochastic processes, which are, intuitively, collections of random variables indexed by an arbitrary index set D taking values in a measurable space E. The key concept needed to define stochastic processes is that of the cylindrical σ-algebra.

Definition 1 (Cylindrical σ-algebra)**.** Let $(E, \mathcal{E})$ be a measurable space and D an arbitrary index set. Consider the evaluation functionals defined by $\pi_x(f) := f(x)$ for any $f : D \to E$ and any $x \in D$. The, the cylindrical σ-algebra $\mathcal{E}^D$ is defined as the smallest σ-algebra on E^D making all pointwise evaluation functionals $\pi_x : E^D \to E$ measurable.

Stochastic processes are then defined as measurable mappings with respect to the cylindrical σ-algebra.

Definition 2 (Stochastic Process)**.** An E-valued stochastic process indexed by D is a measurable mapping

$$Z : (\Omega, \mathcal{F}) \to (E^D, \mathcal{E}^D)$$

Any stochastic process induces a measure on the space of functions from D to E, called the **law** of the process.

Definition 3 (Law of a Stochastic Process)**.** Let Z be an E-valued stochastic process indexed by D, then the law of Z is the image measure μ_Z of $\mathbb{P}$ under Z on $(E^D, \mathcal{E}^D)$ defined by

$$\mu_Z := \mathbb{P} \circ Z^{-1}.$$

Two useful by-products of the law of a stochastic process are its *mean function* and *covariance kernel*.

Definition 4. Let Z be an E-valued stochastic process indexed by D and assume that Z is *second order*, that is, for all $x \in D$, $\mathbb{E}\left[Z_x^2\right] < \infty$. Then the **mean function** $m : D \to E$ and **covariance kernel** $k : D \times D \to E$ of the process are defined as

$$\begin{aligned} m_x &= \mathbb{E}\left[Z_x\right] \\ k_{xx'} &= \mathbb{E}\left[\left(Z_x - m_x\right)\left(Z_{x'} - m_{x'}\right).\right] \end{aligned}$$

Note that for the sake of conciseness we use subscript notation for the arguments of the mean function and covariance kernel. Of all the possible types of stochastic processes, in this thesis we will be mostly interested in the class of *Gaussian processes*. Gaussian processes (GPs) are defined as stochastic processes whose finite-dimensional laws at any set of points are Gaussian.

Definition 5 (Gaussian Process)**.** A **Gaussian process** on a domain D is a real-valued stochastic process Z, such that for any finite set of points in the index space $x_1, .., x_n \in D$, the random vector $(Z_{x_1}, ..., Z_{x_n})$ is multivariate Gaussian distributed.

It can be shown that, for a GP, its mean function and covariance kernel completely determine the law of the process (this is a consequence of (Dudley, 2002, Theorem 12.1.3)). We will thus write $Z \sim \mathrm{Gp}(m, k)$ to indicate that Z is a Gaussian process with mean function m and covariance kernel k. Most of the regularity properties of the sample paths of a GP are encoded in its covariance kernel, including continuity and differentiability. While most of these regularity results are now classics in the GP community, we refer readers to (Steinwart, 2019) for the latest refinements in the study of sample path regularity.

This encoding of the regularity properties of the GP in the kernel allows one to tailor the process to the problem at hand by specifically tuning the kernel function of the GP. In general, any real valued function $m : D \to \mathbb{R}$ is a valid mean function for defining a GP on D, whereas the kernel function has to be a symmetric positive (semi-) definite function on $D \times D$ (Dudley, 2002, Theorem 12.1.3). In this thesis, we will mainly consider cases where the domain is a subset of Euclidean space $D \subset \mathbb{R}^d$ and covariance kernels are from one of the following families (Rasmussen and Williams, 2006):

- **exponential:** $k(x, x') = \sigma^2 e^{-||x-x'||/\lambda}$
- **squared exponential** $k(x, x') = \sigma^2 e^{-||x-x'||^2/2\lambda^2}$
- **Matérn** $k_\nu(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||x-x'||}{\lambda}\right)^\nu \mathcal{K}_\nu \left(\frac{\sqrt{2\nu}||x-x'||}{\lambda}\right)$,

where σ^2 and $\lambda, \nu > 0$ are real parameters, $||\cdot||$ is the Euclidean norm and $\mathcal{K}_\nu$ denotes the modified Bessel function of the second kind. These kernels are among the most used in the literature and lead to simple regularity properties of the sample paths. Informally, the sample paths of a GP with exponential kernel are continuous but not differentiable, those of the squared exponential kernel are infinitely differentiable and those of the Matérn kernel are k-times continuously differentiable for $k < \nu$. We note that this characterization, though often sufficient in practice, is far from rigorous. For a rigorous treatment using powers of reproducing kernel Hilbert spaces we refer to Section 4.4 of the review article (Kanagawa et al., 2018).

2.2.2 Bayesian Inversion

Gaussian processes have become a fundamental tool in the Bayesian approach to inverse problems, owing to the ease with which they allow one to define priors on function spaces. Before considering the use of GPs in an inversion setting, we first introduce Gaussian process regression, also known as kriging.

The goal in GP regression is to learn an unknown function $\rho : D \to \mathbb{R}$ from (noisy) evaluations of the function at a finite set of points $\boldsymbol{W} = (w_1, ..., w_r) \in D^r$. There exists several approaches to tackle this problem. We will here only focus on the ones based on random functions. These techniques first originated in geostatistics (Krige, 1951; Matheron, 1962) and a comprehensive historical account can be found in (Chilès and Desassis, 2018). We here adopt a Bayesian perspective (O'Hagan, 1978), rather than sticking to the traditional approach.

To approximate ρ, we assume it is a realization of some prespecified GP prior $Z \sim \mathrm{Gp}(m, k)$. Our goal is then to use the conditional law of Z under the available data to approximate ρ. To streamline equations, it helps to introduce notation that compactly represents concatenated quantities and matrices built from kernel evaluations. Thus, given two set of points $\boldsymbol{X} = (x_1, ..., x_m) \in D^m$ and $\boldsymbol{W} = (w_1, ..., w_r) \in D^r$, we will use $K_{\boldsymbol{XW}}$ to denote the $m \times r$ matrix obtained by evaluating the covariance function at all couples of points $K_{ij} = k(x_i, w_j)$. In a similar fashion, let $Z_{\boldsymbol{X}} \in \mathbb{R}^m$ denote the vector obtained by concatenating the values of the field at the different points. Similar notation will be used for the concatenated vector of mean function evaluated at several locations $m_{\boldsymbol{X}} \in D^m$. In general, boldface uppercase letter will be used to denote batches of points and concatenated quantities (usually datasets).

Using this compact notation, the data observation process is distributed according to:

$$\boldsymbol{Y} = Z_{\boldsymbol{W}} + \boldsymbol{\epsilon}, \tag{2.3}$$

where we have assumed that the observations are corrupted by additive centered Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$ with covariance matrix $\Delta \in \mathbb{R}^{r \times r}$. Then, assuming that one observes the values of ρ at the batch of points $\boldsymbol{W}$ and gets the data vector $\boldsymbol{y} = (y_1, ..., y_r) \in \mathbb{R}^r$, the conditional law of the process Z, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$ is Gaussian with mean function and covariance kernel given by:

$$\tilde{m}_{\boldsymbol{X}} = m_{\boldsymbol{X}} + K_{\boldsymbol{XW}} \left(K_{\boldsymbol{WW}} + \Delta\right)^{-1} \left(\boldsymbol{y} - m_{\boldsymbol{W}}\right), \tag{2.4}$$

$$\tilde{K}_{\boldsymbol{XX'}} = K_{\boldsymbol{XX'}} - K_{\boldsymbol{XW}} \left(K_{\boldsymbol{WW}} + \Delta\right)^{-1} K_{\boldsymbol{WX'}}. \tag{2.5}$$

In this setting (known mean function and covariance kernel), these equations agree with the simple kriging ones. In practice, the mean is usually inferred from the data (ordinary kriging, universal kriging) and the covariance kernel as well (MLE for parametric models, variogram fitting). Since GP regression under pointwise evaluation data is not the focus of this thesis, we refer the reader to (Rasmussen and Williams, 2006; Chilès and Delfiner, 2012) for more details and now turn to the case of linear operator data and inverse problems.

As explained in Section 2.1, a linear inverse problem can be viewed as a regression task with linear operator data observations. One possible approach is then to extend GP regression to such linear operator data in order to bring it to bear on inverse problems (Tarantola and Valette, 1982; Särkkä, 2011). Our exposition here

will follow the one in (Tarantola, 2005). For the rest of this section, assume that the forward operator is a linear operator between finite dimensional spaces $G : \mathbb{F} \to \mathbb{Y}$. Chapter 3 will be dedicated to infinite-dimensional formulations of Bayesian inversion, but in practice most problems are discretized before an attempt at a solution is made, so that finite-dimensional formulations are not much of a limitation. We also assume that the action of the forward only involves the values of the unknown function at a finite set of points $\boldsymbol{W} \in D^r$, so that the observation model can be written as:

$$\boldsymbol{Y} = GZ_{\boldsymbol{W}} + \boldsymbol{\epsilon}, \tag{2.6}$$

where we assume the data is corrupted by additive Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$. Under this observation model, assuming that the matrix $\left(GK_{\boldsymbol{WW}}G^T + \Delta\right)$ is invertible, the posterior law of Z, conditional on $\boldsymbol{Y} = \boldsymbol{y}$ is Gaussian with mean and covariance given by (Tarantola, 2005, Chapter 3):

$$\tilde{m}_{\boldsymbol{X}} = m_{\boldsymbol{X}} + K_{\boldsymbol{XW}}G^T \left(GK_{\boldsymbol{WW}}G^T + \Delta\right)^{-1} \left(\boldsymbol{y} - Gm_{\boldsymbol{W}}\right), \tag{2.7}$$

$$\tilde{K}_{\boldsymbol{XX'}} = K_{\boldsymbol{XX'}} - K_{\boldsymbol{XW}}G^T \left(GK_{\boldsymbol{WW}}G^T + \Delta\right)^{-1} GK_{\boldsymbol{WX'}}. \tag{2.8}$$

The posterior can then be used to approximate the unknown solution to the inverse problem. The Bayesian approach bears some connections to the traditional Tychonov regularization method, in that the posterior mean can be seen as the solution of a regularized problem with a regularization term that penalizes solutions away from the prior (Calvetti and Somersalo, 2018).

Although the Bayesian framework for inverse problems seems complete and straightforward to apply, in practice it can break down when one tries to scale it to large-scale problems. Chapter 4 of this thesis is dedicated to overcoming these difficulties. Moreover, the standard priors used in Bayesian inversion tend to lack expressiveness when it comes to the inclusion of expert knowledge. Chapter 5 tackles this issue by devlopping techniques for the use of more flexible priors. Finally, the theory behind the use of GP with linear operators observations generally leaves aside the details involved in constructing the conditional law in the infinite dimensional context. Chapter 3 is dedicated to providing a theoretical grounding in modern probability theory for Bayesian inversion, based on the theory of Gaussian measures, which we briefly introduce next.

2.2.3 Gaussian Measures

When trying to generalize Gaussian random variables beyond the usual scalar setting, one of the natural concepts that arises is that of *Gaussian measures*. Gaussian measures can in many ways be seen as a theoretical counterpart to GPs, the former being the tool of choice for theoretical inquiries while the latter is preferred for applied endeavours. While the two communities (process versus measure) tend to be mutually isolated, we believe that many fruitful developments can emerge from a tighter interplay of the two. Chapter 3 of this thesis is devoted to fostering such interplay, and we here introduce the basics in Gaussian measure theory that will be needed later.

The usual setting in which Gaussian measures are defined is that of Banach spaces. A Gaussian measure then being a measure upon a Banach space, satisfying

some Gaussianity properties. When thinking of the potential connection with GPs, one can, for all practical purposes, think of this Banach space as a space of functions in which sample paths of an "equivalent" GP would live (in a way that has yet to be defined). In the following, we will always assume that the Banach spaces under consideration are separable. Our exposition will mostly be based on (Bogachev, 1998). Other useful references include (Kuo, 1975) and also (Hairer, 2009) for a more introductory treatment.

Definition 6 (Gaussian Measure)**.** A **Gaussian measure** μ on a separable Banach space $\mathbb{F}$ is a Borel measure on $\mathbb{F}$ such that for any continuous linear functional $\ell \in \mathbb{F}^*$, the measure $\ell^{\#}\mu := \mu \circ \ell^{-1}$ on $\mathbb{R}$ is Gaussian.

The assumption of separability is not a very restricting one, since most usual function spaces, such as $L_p(D)$ spaces or Sobolev spaces $W^{k,2}(D)$ are separable (for finite p) and so is the space of continuous functions over a compact set (equipped with the sup-norm). The main reason for requiring separability is that it makes the two natural definitions of measurability on the space under consideration coincide, which prevents pathological cases from arising.

Theorem 2. *On a separable Banach space $\mathbb{F}$ the Borel σ-algebra and the cylindrical σ-algebra coincide.*

Proof. See Hairer (2009). □

Compared to GPs, when working with Gaussian measures, the notions of mean and covariance functions are respectively replaced by the *mean element* and *covariance operator*. Here we denote by $\mathbb{F}^*$ the (continuous) dual space of $\mathbb{F}$, and for any element $f \in \mathbb{F}$ and continuous linear form g^* we use the duality notation $\langle f, g^* \rangle = g^*(f)$.

Definition 7. Given a Gaussian measure μ on a Banach space $\mathbb{F}$, the mean of μ is the unique element $m_\mu \in \mathbb{F}$ such that:

$$\int_{\mathbb{F}} \langle f, g^* \rangle d\mu(f) = \langle m_\mu, g^* \rangle, \ \forall g^* \in \mathbb{F}^*. \tag{2.9}$$

The covariance operator of μ is the linear operator $C_\mu : \mathbb{F}^* \to \mathbb{F}$ defined by

$$\langle C_\mu g_1^*, g_2^* \rangle = \int_{\mathbb{F}} \left(\langle f, g_1^* \rangle - \langle m_\mu, g_1^* \rangle \right) \left(\langle f, g_2^* \rangle - \langle m_\mu, g_2^* \rangle \right) d\mu(f), \ \forall g_1^*, g_2^* \in \mathbb{F}^* \tag{2.10}$$

We note that there are some subtleties in the definition of the above notions. For example, at first sight the mean element should be an element of the bi-dual space $\mathbb{F}^{**}$ and the covariance operator should also map in that space. Nevertheless, in the separable setting, one can identify these with elements of $\mathbb{F}$ itself, thus greatly simplifying the theory. We will not focus on these details further and refer readers to (Vakhania et al., 1987, Chapter 3) for details on this identification.

2.3 Bayesian Set Estimation

Compared to usual "pointwise" estimation, where the goal is to learn a (possibly multivariate) unknown value, the problem of set estimation has received comparably less attention in the machine learning community. We here summarize some of the main techniques for set estimation in a Bayesian setting, focusing on the case of *excursion sets*.

Given an unknown function $\rho : D \to \mathbb{R}$ and some threshold T, the **excursion set of ρ above** T is the set

$$\Gamma^* = \{x \in D : \rho(x) \geq T\}.$$

When the unknown function of interest arises as part of a Bayesian inverse problem (see Section 2.2.2) with prior Z, there exists several approaches to approximate Γ^* using the posterior. For example, a naive estimate for Γ^* may be obtained using the **plug-in estimator**:

$$\hat{\Gamma}_{\text{plug-in}} := \{x \in D : \ \tilde{m}_x \geq T\},$$

where $\tilde{m}_x$ denotes the posterior mean function of the GP prior. In this work, we will focus on recently developed more sophisticated approaches to Bayesian excursion set estimation (Azzimonti et al., 2016; Chevalier et al., 2013) based on the theory of random sets (Molchanov, 2005). We here briefly recall some theory taken from the aforementioned source.

In the following, we focus on a Bayesian inversion setting as introduced in Section 2.2.2. Given some initial dataset, we denote by $\tilde{Z}$ a random field on D that is distributed according to the posterior distribution conditionally on the initial dataset. Then, the posterior distribution of the field gives rise to a **random closed set** (RACS):

$$\Gamma := \{x \in D : \ \tilde{Z}_x \geq T\}. \tag{2.11}$$

One can then consider the probability for any point in the domain to belong to that random set. This is captured by the **coverage function**:

$$\begin{aligned} p_\Gamma : & D \to [0,1] \\ & x \mapsto \mathbb{P}\left[x \in \Gamma\right]. \end{aligned}$$

The coverage function allows us to define a parametric family of set estimates for Γ, the **Vorob'ev quantiles**:

$$Q_\alpha := \{x \in D : p_\Gamma(x) \geq \alpha\}. \tag{2.12}$$

The family of quantiles Q_α gives us a way to estimate Γ by controlling the (pointwise) probability α that the members of our estimate lie in Γ. There exists several approaches for choosing α. One such possible approach is to choose it such that the volume of the resulting quantile is equal to the expected volume of the excursion set. This gives rise to the **Vorob'ev expectation**.

Definition 8. (Vorob'ev Expectation) The Vorob'ev expectation is the quantile Q_{α_V} with threshold α_V chosen such that

$$\mu(Q_\alpha) \leq \mathbb{E}[\nu(\Gamma)] \leq \mu(Q_{\alpha_V}), \ \forall \alpha > \alpha_V,$$

where $\nu(\cdot)$ is some fixed measure on the domain D.

In practice, the computation of the Vorob'ev expectation requires the computation of the expected excursion volume under the posterior. While direct computation of this quantity is cumbersome, under suitable conditions, Robbins's theorem relates the expected excursion volume to an integral of the coverage function:

$$\mathbb{E}[\nu(\Gamma)] = \int_D p_\Gamma(x) d\nu(x).$$

We refer the reader to Robbins (1944) and Molchanov (2005) for more details.

To illustrate the various Bayesian set estimation concepts introduced here, we apply them to a simple one-dimensional inverse problem, where one wants to estimate the excursion set above 1.0 of a function $f : [-1, 1] \to \mathbb{R}$ after 3 pointwise evaluations of the function have been observed (Figure 2.1).

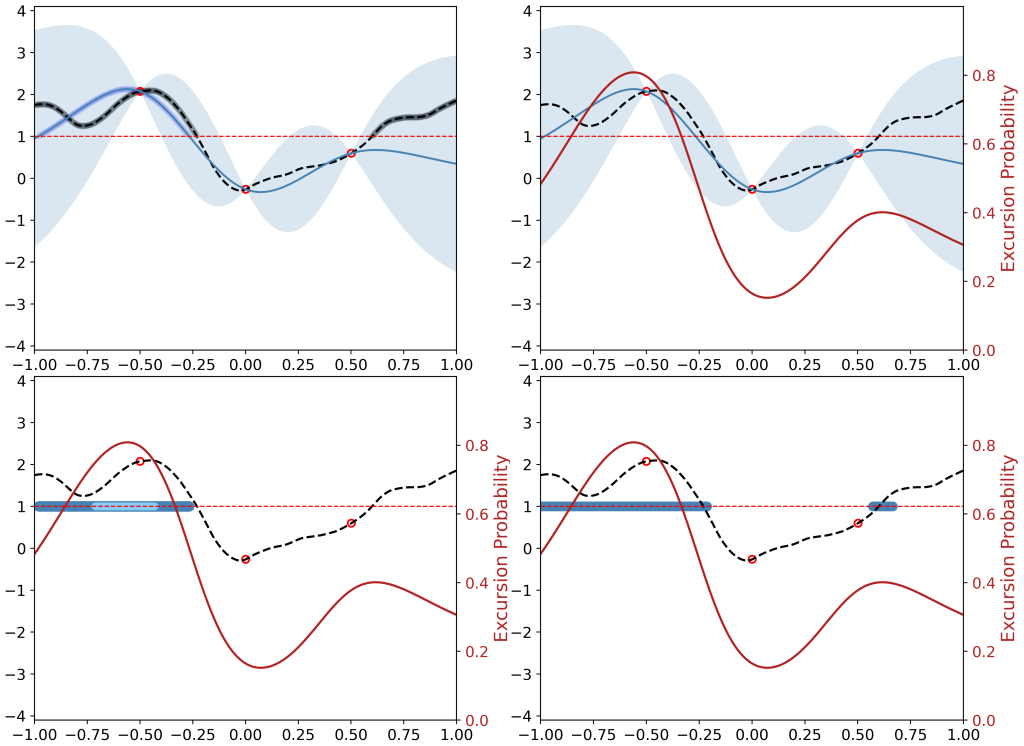


Figure 2.1: One-dimensional Bayesian set estimation example. Excursion threshold in red. True function in black, posterior mean and 2σ confidence regions in blue (conditionally on observations at the red dots). True excursion region is highlighted in black and plug-in estimate is highlighted in blue. Top right: coverage function (dark red). Bottom left: Vorob'ev quantiles at level $\alpha = 0.5$ (dark blue) and $\alpha = 0.75$ (light blue). Bottom right: Vorob'ev expectation (threshold $\alpha_V = 0.4$).

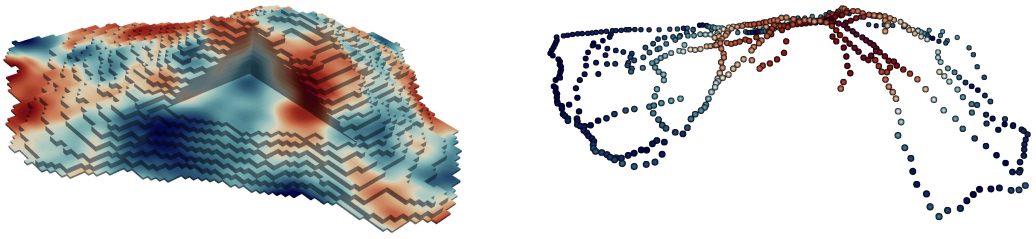
2.4 Applicative Examples

Since one of the core goals of this thesis is to bring more realistic GP models to the inverse problem community, we will strive to test all our methods on real-world problems. Compared to the synthetic examples that are sometimes used as benchmarks in the GP literature, inverse problems originating in the natural sciences tend

to carry their own load of contingencies that are usually unforeseen when developing algorithms for in-silico problems. In this regard, most of the chapters of this thesis are built with a target application in mind, that serves as a red thread for the development of our proposed techniques. Our two main applications are introduced next.

2.4.1 Gravimetric Inversion

One of the type of inverse problems that embodies the challenges that this thesis ambitions to tackle are gravimetric ones. In *gravimetric inverse problems*, the goal is to reconstruct the mass density distribution $\rho : D \to \mathbb{R}$ in some given underground domain D from observations of the vertical component of the gravitational field at points $u_1, ..., u_q$ on the surface of the domain.



(a) (b)

Figure 2.2: Overview of a generic gravimetric inverse problem: (a) underground mass density (realization from GP prior), (b) vertical intensity of the generated gravity field at selected locations.

Such gravimetric data are extensively used on volcanoes, and, in this thesis, we will focus on the Stromboli volcano as an applicative example, using gravimetric data gathered on the surface of the volcano during a field campaign in 2012 (Linde et al., 2014). For volcanoes, reconstructing the underground mass density field is useful for understanding geology, localizing ancient volcano conduits and present magma chambers, and for identifying regions of loose light-weight material that are prone to landslides that could in the case of volcanic islands generate tsunamis (Montesinos et al., 2006; Represas et al., 2012; Linde et al., 2017). Figure 2.2 displays the main components of the problem.

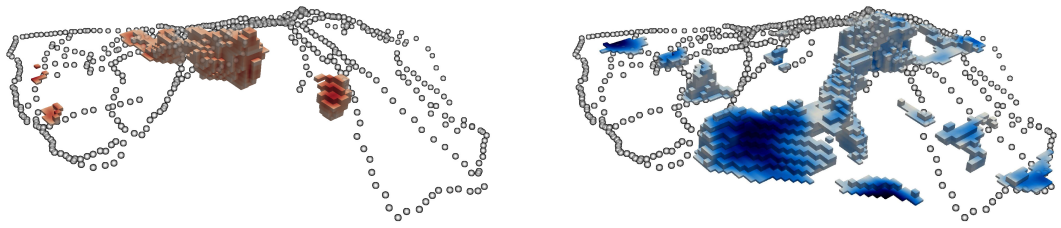
The observation operator describing gravity measurements is an integral one (see Section 4.6), which, after discretization, fits the Bayesian inversion framework of Section 2.2. After discretization on a finite grid of points $\boldsymbol{X} = (x_1, \dots, x_m)$, the forward version of the problem writes as:

$$\boldsymbol{Y} = \bar{G}\rho_{\boldsymbol{X}} + \boldsymbol{\epsilon}, \tag{2.13}$$

where the $q \times m$ matrix $\bar{G}$ represents the discretized version of the observation operator for the gravity field at $u_1, \dots, u_q$ and we assume i.i.d Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau^2 I_q)$. The posterior may then be computed using Eqs. (4.3) and (4.4).

This gravimetric inverse problem is an interesting example, in that it embodies some of the key challenges that arise when trying to scale the Bayesian inversion

techniques from Section 2.2 to real-world applications. First, the numerical resolution of such problems is fraught with memory overloads, owing to the domain being three-dimensional and often of large extent, resulting in large grid sizes. Second, the integral nature of the observation operator usually prevents the application of sparsification techniques that are meant to alleviate the computational difficulties linked with large grids. One of the contributions of this thesis is the development of new approaches to tackle these challenges and allow the solution of such large-scale Bayesian inverse problems. This will be the focus of Chapter 4.



(a) (b)

Figure 2.3: Set estimation in a generic gravimetric inverse problem (continuation of Fig. 2.2): (c) high density regions and (d) low density regions. Thresholds and color scales were chosen arbitrarily.

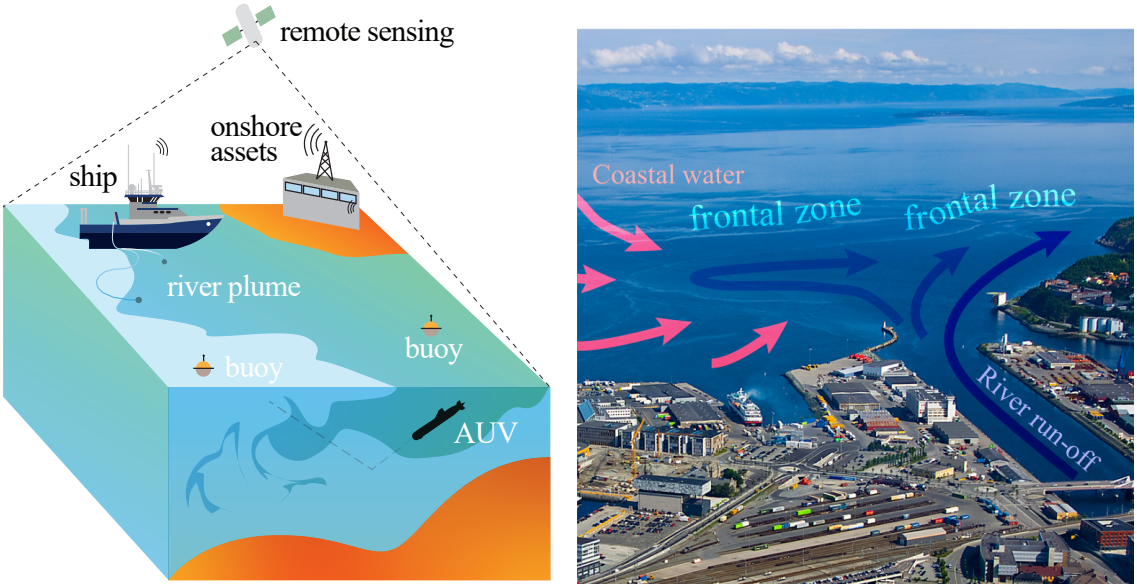
On top of the aforementioned characteristics, gravimetric inverse problems also provide a natural setting to demonstrate UQ and set estimation techniques in Bayesian inverse problems. Indeed, estimating excursion- and sojourn-sets, that is high- and low-density regions within the volcano is of interest for understanding the history and evolution of the volcano, these regions being typically linked to geological features of interest.

2.4.2 River Plume Mapping

The second applicative example considered in this thesis is that of mapping a river plume from observations of the temperature and salinity field inside a fjord. When a river enters a fjord, the mixing of river and ocean water does not happened straight away, and the inflow of cold freshwater creates a strong gradient in both temperature and salinity, forming a frontal region called *river plume* (see Fig. 2.4b).

River plumes host a range of complex bio-geophysical interactions, driven by an agglomeration of physical forcings (e.g. wind, topography, bathymetry, tidal influences, etc.) and incipient micro-biology driven by planktonic and coastal anthropogenic input, such as pollution and agricultural runoff transported into the ocean by the river. These interactions can result in a range of ecosystem-related phenomena such as blooms and plumes, with direct and indirect effects on society (Ryan et al., 2017), making river plumes a prototypical example of regions where the intermingling and climate change and anthropogenic impact can be studied and monitored.

This river plume mapping problem can be seen as a kind of inverse problem and



(a) Illustration of a range of ocean sensing opportunities.

(b) Frontal patterns off of the Nidelva river, Trondheim, Norway.

Figure 2.4: River plume mapping problem (Nidelva river). River-ocean interactions dynamically affect shape of rive plume (2.4b), calling for autonomous real-time mapping strategies (2.4a).

set estimation problem. Indeed, denoting by

$$\rho^{(1)}, \rho^{(2)} : D \to \mathbb{R}$$

the temperature and salinity field inside the fjord, then the river plume can be characterized as the region (sojourn set) of low temperature and salinity, or as the complement of the excursion set:

$$\Gamma^* := \{x \in D : \rho_x^{(1)} \geq T_1, \rho_x^{(2)} \geq T_2\},$$

where T_1, T_2 denote temperature and salinity thresholds. Note that we here chose to formulate the problem as an estimation of the excursion (ocean) rather than the sojourn (river) set, due to the fact that the techniques presented next are tailored for excursion sets (though they can in principle be extended to sojourn sets). The goal is then to estimate the region Γ^* from partial observations of the multivariate temperature-salinity field. Owing to these features, we will use this application to demonstrate multivariate extensions to traditional Bayesian inversion techniques.

Moreover, river-ocean interactions depend on variations in river discharge, tidal effects, coastal current and wind, leading to frequent distortions of the river plume boundary. This undermines any *static* mapping attempt and calls for dynamic and adaptive mapping strategies. In a Bayesian framework, this makes river plume estimation a natural testbed for multivariate extensions to sequential uncertainty reduction strategies, which we will develop in Chapter 6. In practice, the mapping strategies will be executed by an autonomous underwater vehicle (AUV) equipped with temperature and salinity sensors.

Chapter 3

Sequential Bayesian Inversion and Disintegrations of Gaussian Measures

This chapter reproduces the paper Travelletti and Ginsbourger (2022), co-authored with David Ginsbourger and submitted to the Electronic Journal of Statistics (DOI:10.48550/ARXIV.2207.13581).

3.1 Background

As explained in Section 2.2, Gaussian processes are able to assimilate pointwise observations and can also handle discretized operator observations. Recently however, the advent of indirect, functional data (tomographic data, derivative data (Solak et al., 2003; Ribaud, 2018)) that do not boil down to simple pointwise evaluations of the original latent function has sparked interest in extending GPs to different types of observations, such as integral observations (Hendriks et al., 2018; Jidling et al., 2019) or linear constraints (Jidling et al., 2017; Agrell, 2019). In this chapter, we aim at providing theoretical foundations to those approaches, focusing particularly on the question of what types of operator data can be assimilated using GPs and trying to provide a framework for sequential assimilation of operator data. To that end, we will formulate the assimilation process using the language of disintegrations of Gaussian measures. This will also allow us, in passing, to clarify some relations between Gaussian processes and Gaussian measures; which we hope will open new research venues at the intersection of Gaussian measures theory and Bayesian assimilation/inversion.

3.2 Introduction

Broadly speaking, all the methods for linear operator data assimilation with GPs aim at learning ρ from linear form data $\ell_i(\rho)$, where $\ell \to \mathbb{R}$ $(i = 1, \dots, q)$ are linear functionals on some Banach space $\mathbb{F}$ of functions on D. Just like under pointwise observations, working out conditional distributions boils down to applying conditioning formulae to finite-dimensional vectors, in that case to vectors of the

form $(Z_x, Z_{x'}, \ell_1(Z), \ldots, \ell_q(Z))$ $(x, x' \in D)$.

Compared to the basic case of pointwise observations, however, ensuring that the usual way of deriving conditional distributions does actually work under linear form data requires a bit of care. The usual approach in practice is to silently assume that the considered functionals of Z can be expressed as limits of linear combinations of pointwise field evaluations, so that everything will work as intended. In several cases, this condition might not be straightforward to verify, and things can get even worse when one considers observations described by linear operators between Banach spaces $G : \mathbb{F} \to \mathbb{Y}$, thus raising the question of what kind of operator data can be assimilated, or more precisely, of which properties an operator G needs to satisfy in order for the conditional law to be well-defined. While this question can be tricky to answer using the traditional Gaussian process framework, modern probability theory in Banach spaces offers a rigorous, generic approach to conditioning under linear operator using the language of disintegrations of measures, as we will clarify next.

Beyond establishing solid mathematical foundations for conditioning on linear operator data, another problem that has received much attention lately in the GP literature is that of efficiently performing sequential data assimilation (Attia et al., 2018; Huber, 2014; Solin et al., 2015). In such a framework, new data become available sequentially and predictions have to be recomputed along the way to incorporate the new information. To alleviate the computational burden associated to sequential learning, various *updating* scheme have been developed (Chevalier et al., 2014b; Emery, 2009; Gao et al., 1996; Barnes and Watson, 1992) which aim at expressing the contribution of the new data as an update to the current posterior.

In the present work, we focus on the intersection of the two aforementioned topics, that is, we concentrate on sequential assimilation of linear operator data. Our aim is to provide an abstract mathematical foundation for the above setting by formulating it in the language of disintegrations and to derive update formulae for disintegrations. In passing, we clarify the link between the traditional Gaussian process framework and the Gaussian measure language.

In this chapter, we will start by reviewing results from Rajput and Cambanis (1972) in order to prove equivalence of the Gaussian process and Gaussian measure approaches in various cases. We will also connect this with recent results on sample path properties of GP (Steinwart, 2019) to characterize situations under which GPs induce a Gaussian measure on some suitable space of functions.

We will then turn to disintegrations of Gaussian measures (Tarieladze and Vakhania, 2007), which we will extend to the non-centered and sequential case, thereby providing an extension of the usual kriging update formulae (Chevalier et al., 2014b) to disintegrations.

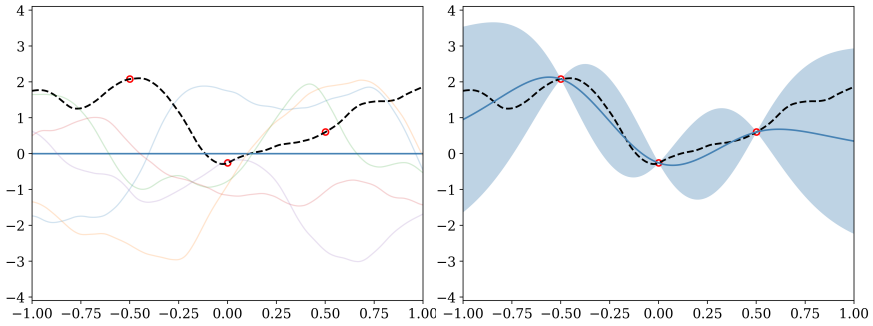
Those results offer prospects for theoretical inquiries in Bayesian optimization (Bect et al., 2019) as well as more applied uses, such as the formulation of discretization-independent algorithms in Bayesian inversion (Cotter et al., 2013). We also hope that our efforts to shed light on the Gaussian process - Gaussian measure equivalence will help bring benefits of the abstract language of disintegrations to the applied GP community. In Chapter 4 the results of the present chapter will be used to provide rigorous theoretical foundations to an implicit updating scheme for large covariance matrices.

Example. For the rest of this work, we will consider the task of learning an unknown

function ρ living in a separable Banach space $\mathbb{F}$ from data of the form $y_i = G_i(\rho),\ i = 1, \ldots, q$, where

$$G_i : \mathbb{F} \to \mathbb{Y},$$

are bounded linear operators into a separable Banach space $\mathbb{Y}$, we will call the G_i the *observation operators*. As a simple example of a problem falling into this setting, consider the task of learning a continuous function defined on the interval $[-1, 1]$ via different types of data: pointwise function values, integrals of the function, Fourier coefficients, etc. Figure 3.1 provides an illustration of solutions obtained under a Gaussian process prior. Note that the three different combinations of observations in Figure 3.1 can each be described by a linear operator $G : C([-1, 1]) \to \mathbb{R}^q$



(a) prior (b) pointwise

(c) pointwise + integral (d) pointwise + integral + Fourier

Figure 3.1: Conditional mean (blue) and 2σ credible intervals after inclusion of different types of data: (a) realizations independently sampled from the prior GP, (b) prediction based on pointwise data at 3 locations, (c) prediction based on pointwise data + integral over domain, (d) prediction based on pointwise data + integral over domain + first two Fourier coefficients. The true unknown function is shown in dashed black.

Note that this example can already serve to illustrate the theoretical difficulties associated with the conditional law under linear operator observations. Consider for example derivative observations of the form $y = \rho'(x_0),\ x_0 \in D$. The usual procedure when working with derivatives of GPs is to assume mean square differentiability of the process. But even then, results on the link between mean square differentiability of the process and almost sure differentiability of the paths (Cambanis, 1973; Scheuerer, 2010) require additional assumptions to ensure path differentiability, and in general the observation operator is not guaranteed to be bounded.

3.3 Gaussian Process-Measure Equivalence

As briefly explained in Section 2.2.3, when working with Gaussian priors over spaces of functions defined over an arbitrary domain D, two complementary approaches are often used: Gaussian processes and Gaussian measures. The goal of this section is to provide conditions under which the two approaches are equivalent.

When considering Gaussian processes with continuous trajectories over a compact metric space D, the Gaussian process and Gaussian measure points of view are known to be equivalent, with $\mathbb{F}$ being the Banach space of continuous functions $C(D)$ equipped with the sup norm. Indeed, one can show that a Gaussian measure on $C(D)$ defines an equivalent Gaussian process on D with continuous trajectories, and vice-versa. This allows one to work with Gaussian measures and Gaussian processes interchangeably on this Banach space. The equivalence is ensured by the following two theorems, which are multidimensional analogues of the one presented in Rajput and Cambanis (1972).

We first show that a Gaussian process on D with continuous sample paths induces a Gaussian measure on $C(D)$. Indeed, given such a Gaussian process Z, one may try to induce a measure on $C(D)$ by setting $\mu_Z := \mathbb{P} \circ \Phi^{-1}$, where $\Phi(\omega) := Z(\cdot; \omega) \in C(D)$. The next theorem guarantees that this indeed defines a Gaussian measure. This result is well known in the Gaussian measure literature (see e.g. Bogachev (1998)) and we provide a proof in the appendix for the sake of completeness.

Theorem 3. *Let $(\Omega, \mathcal{F}, \mathbb{P}; Z(\omega, x), x \in D)$ be a Gaussian process on a compact metric space D with continuous sample paths. Then the induced measure*

$$\mu_Z := \mathbb{P} \circ \Phi^{-1}$$

is well-defined (as a Borel measure) and Gaussian.

On the other hand, given a Gaussian measure μ on $C(D)$, the following theorem ensures that μ induces indeed a Gaussian process.

Theorem 4. *Let μ be a Gaussian measure on $C(D)$, for a compact metric space D. Then, letting $\Omega = C(D)$ and $\mathcal{F}$ be the Borel sigma algebra on $C(D)$, the collection of random variables*

$$Z_x : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})), \ \omega \mapsto \delta_x(\omega)$$

for all $x \in D$ defines a Gaussian process with paths in $C(D)$ which induces μ on $C(D)$.

Under this correspondence, the mean and covariance functions of the process may be obtained as special cases of the mean element and covariance operator of the corresponding measure by acting on them with pointwise evaluation functionals (which in this case belong to the continuous dual of the Banach space under consideration):

Lemma 1. *Let Z be a Gaussian process on a compact metric space D with continuous trajectories, and let μ be the corresponding induced measure on $C(D)$. Then the covariance operator and mean element of the measure are related to the mean and covariance function of the process via*

$$m_x = \mathbb{E}[Z_x] = \langle m_\mu, \delta_x \rangle, \tag{3.1}$$

$$k(x, x') = \mathbb{E}[Z_x Z_{x'}] - \mathbb{E}[Z_x]\,\mathbb{E}[Z_{x'}] = \langle C_\mu \delta_{x'}, \delta_x \rangle, \tag{3.2}$$

for all $x, x' \in D$.

These considerations allow us to work interchangeably with the two points of views. While in many practical circumstances the GP point of view is sufficient, Gaussian measures can be leveraged to provide rigorous updating of GPs under linear operator observations, as we will show in Section 3.4.

Remark 1. The correspondence between Gaussian processes and measures is not limited to the Banach space $C(D)$ of continuous functions over a compact metric space. Indeed Rajput and Cambanis (1972) also prove correspondence for L^p spaces and spaces of absolutely continuous functions. However, the proofs are done on a case by case basis.

Even if the Banach space $C(D)$ of continuous functions on a compact domain provides a basic setting for the Gaussian process - Gaussian measure equivalence, it often proves insufficient when one wants to use this correspondence to tackle conditioning under linear operator observations. For example, the differential operator d/dx is not even a well-defined operator on $C(D)$. For such operators, the natural domains to consider are Sobolev spaces. This shows that, in the Gaussian measure framework, when one wants to assimilate observations that are "finer" than simple pointwise evaluations, one has to go beyond the Banach space $C(D)$. This is what we will do in the following section by considering reproducing kernel Hilbert spaces.

The Reproducing Kernel Hilbert Space Case: The proofs of the process-measure equivalence theorems Theorems 3 and 4 in the Banach space of continuous functions over a compact domain rely on having a characterization of the dual space of the Banach space under consideration, and on being able to approximate elements of the dual via pointwise evaluations. Indeed, Gaussian measures on a Banach space are characterized by the Gaussianity of their linear functionals, whereas GPs are characterized by the Gaussianity of finite collections of pointwise evaluations, making the link between linear functionals and pointwise evaluations a crucial one in the correspondence.

The natural class of spaces where such a link exists is that of reproducing kernel Hilbert spaces (RKHS) (Aronszajn, 1950; Schwartz, 1964; Berlinet and Thomas-Agnan, 2004; Kanagawa et al., 2018). Indeed, one of the defining properties of RKHS is that their (continuous) dual contain the evaluation functionals, so that one can directly adapt the process-measure correspondence theorems. Note that the product measurability is still guaranteed by Theorem 12 since RKHS of functions over a compact metric space are contained in the Banach space of continuous functions provided that the reproducing kernel is continuous.

Theorem 5. *Let* $(\Omega, \mathcal{F}, \mathbb{P}; Z(\omega, x), x \in D)$ *be a Gaussian process with trajectories in a separable RKHS* $\mathcal{H}$ *of functions over a compact metric space* D*. Then the induced measure*

$$\mu_Z := \mathbb{P} \circ \Phi^{-1}$$

is well-defined (as a Borel measure) and Gaussian.

Theorem 6. *Let μ be a Gaussian measure on a separable RKHS $\mathcal{H}$ of functions over a compact metric space D. Then, letting $\Omega = \mathcal{H}$ and $\mathcal{F}$ be the Borel sigma algebra on $\mathcal{H}$, the collection of random variables*

$$Z_x : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})), \ \omega \mapsto \delta_x(\omega)$$

for all $x \in D$ is a Gaussian process with paths in $\mathcal{H}$ which induces μ on $\mathcal{H}$.

The question whether GP sample paths lie in an RKHS has been widely studied in the literature (Steinwart and Scovel, 2012; Steinwart, 2019). One of the most well-known results in this domain is a negative one, namely that for a GP with continuous covariance kernel and almost-sure sample paths, the probability that the trajectories lie within the RKHS associated to the kernel of the process is zero (Driscoll, 1973; Lukić and Beder, 2001). Recent works have aimed at finding "larger" RKHS that contain the paths of the process. It turns out that for a broad class of GPs, one can find an "interpolating" RKHS lying between the RKHS of the kernel of the process and $L^2(\nu)$ (for some measure ν) that contains the sample paths almost surely (Steinwart, 2019, Corollary 5.3).

We here only consider kernels that are bounded on the diagonal: $k(x, x) < \infty$, for all $x \in D$ (as is the case for all the usual kernels). Then, Steinwart and Scovel (2012, Lemma 5.1, Theorem 5.3) guarantees that the conditions required for the sample paths to be contained in powers of the base RKHS hold. Under these conditions, there are results that guarantee the existence of an RKHS containing the trajectories of the process with probability 1. The RKHS depends on the eigenvalues of the operator

$$T_k(f) := \int_D k(\cdot, x) f(x) d\nu(x), \ f \in L_2(\nu),$$

where ν is any finite Borel measure supported on D. The embedding RKHS is then constructed as a power $\mathcal{H}_k^\theta$ of the RKHS $\mathcal{H}_k$ of the kernel (Kanagawa et al., 2018, Definition 4.12).

Theorem 7. *[Kanagawa et al. (2018, Theorem 4.12), Steinwart (2019, Theorem 5.2)] Let Z be a Gaussian process over a compact domain $D \subset \mathbb{R}^d$ with covariance kernel k. Let also $(\lambda_i, \phi_i)_{i \in \mathbb{N}}$ be the eigensystem of the operator T_k. Then, provided $\sum_{i \in \mathbb{N}} \lambda_i^{1-\theta} < \infty$, there exists a version of Z whose sample paths lie in $\mathcal{H}_k^\theta$ with probability* 1.

In particular, for GPs with Gaussian kernels or Matérn kernels and sufficiently regular D, one can always find an RKHS that contains the sample paths of the GP with probability 1, as the following results from Kanagawa et al. (2018) guarantee:

Corollary 1 (Squared Exponential Random Fields, Kanagawa et al. (2018))**.** *If Z is a Gaussian random field with squared exponential kernel k over a compact domain $D \subset \mathbb{R}^d$ with Lipschitz boundary, then for any $0 < \theta < 1$ there exists a version of Z that lies in $\mathcal{H}_k^\theta$ with probability* 1.

Corollary 2 (Matérn Random Fields and Sobolev Spaces, Kanagawa et al. (2018))**.** *When Z is a Mátern Gaussian random field with Matérn kernel $k_{\alpha,\lambda}^{\mathrm{Mat}}$ of order α and lengthscale λ over a domain $D \subset \mathbb{R}^d$ with Lipschitz boundary, then (Kanagawa et al.,*

2018, Corollary 4.15) guarantees that there exists a version of Z that lies in $\mathcal{H}_{k^{\mathrm{Mat}}_{\alpha',\lambda'}}$ with probability 1 for all $\alpha', \lambda' > 0$ satisfying $\alpha > \alpha' + d/2$, provided that D satisfies an interior cone condition (see (Kanagawa et al., 2018, Definition 4.14)).

Wrapping everything together, we can formulate a sufficient condition for a Gaussian process to induce a Gaussian measure on its space of trajectories:

Corollary 3. *Let $(\Omega, \mathcal{F}, \mathbb{P}; Z(\omega, x), x \in D)$ be a Gaussian process on a compact metric space D with covariance kernel k that is continuous and bounded on the diagonal. Then there exists $0 < \theta \leq 1$ such that Z induces a Gaussian measure on $\mathcal{H}_k^\theta$.*

Remark 2. Note that the construction of the power of a RKHS depends on the choice of the measure ν. This is not a significant handicap since the goal of Corollary 3 is to show that under given conditions on a GP one can always induce a measure from it. Nevertheless, recent results (Karvonen, 2021) provide constructions of RKHS containing the sample paths that do not depend on a given measure and are "smaller" than constructions involving powers of RKHS. These constructions are mostly useful in providing more fine-grained descriptions of sample path properties for infinitely smooth kernels (Karvonen, 2021, Chapter 2). We refer the interested reader to the aforementioned literature for more details.

Remark 3. In practice, when working with derivative-type observations, it is often preferable to have simple conditions on the covariance kernel that enforce the paths to live in some Sobolev space that makes the observation operator under consideration a bounded one. Useful results to that end can be found in (Scheuerer, 2010). In particular, it is shown that continuity on the diagonal of the generalized mixed derivatives of the covariance kernel up to order k ensures that the sample paths lie in the local Sobolev space $W^{k,2}_{loc}(D)$ of order k almost-surely (Scheuerer, 2010, Theorem 1).

3.4 Disintegration of Gaussian Measures under Operator Observations

Now that we have discussed the equivalence of the process and the measure approaches, we consider the posterior in the Gaussian measure formulation of conditioning. In this setting, conditional laws are defined using the language of *disintegrations* of measures. The treatment presented here will follow that in Tarieladze and Vakhania (2007) and extend some of the theorems therein.

In the following, we will let $\mathbb{F}$ be a separable Banach space of functions over an arbitrary domain D such that the measure-processes correspondence introduced in Section 3.3 holds, and use μ to denote a Gaussian measure on $\mathbb{F}$ and Z for a corresponding associated Gaussian process on D. Again $G : \mathbb{F} \to \mathbb{Y}$ will denote a bounded linear operator.

Definition 9. Given measurable spaces $(\mathbb{F}, \mathcal{A})$ and $(\mathbb{Y}, \mathcal{C})$, a probability measure μ on $\mathbb{F}$ and a measurable mapping $G : \mathbb{F} \to \mathbb{Y}$, a disintegration of μ with respect to G is a mapping $\tilde{\mu} : \mathcal{A} \times \mathbb{Y} \to [0, 1]$ satisfying the following properties:

1. (**measurability**) For each $y \in \mathbb{Y}$ the set function $\tilde{\mu}(\cdot, y)$ is a probability measure on $\mathbb{F}$ and for each $A \in \mathcal{A}$ the function $\tilde{\mu}(A, \cdot)$ is $\mathcal{C}$-measurable.

2. (**concentration on the fiber**) There exists $\mathbb{Y}_0 \in \mathcal{C}$ with $\mu \circ G^{-1}(\mathbb{Y}_0) = 1$ such that for all $y \in \mathbb{Y}_0$ we have $\{y\} \in \mathcal{C}$ and for each $y \in \mathbb{Y}_0$, the probability measure $\tilde{\mu}(\cdot, y)$ is concentrated on the fiber $G^{-1}(\{y\})$ that is:

$$\tilde{\mu}\left(G^{-1}(\{y\}), y\right) = 1.$$

3. (**mixing**) The measure μ may be written as a mixture of the family $(\tilde{\mu}(\cdot, y))_{y \in \mathbb{Y}}$ with respect to the mixing measure $\mu \circ G^{-1}$:

$$\mu(A) = \int_{\mathbb{Y}} \tilde{\mu}(A, y)\, d\left(\mu \circ G^{-1}\right)(y), \ \forall A \in \mathcal{A}.$$

We will use the notation $\mu_{|G=y}(\cdot) := \tilde{\mu}(\cdot, y)$ for the *disintegrating measure*.

The computation of the posterior, in the Gaussian measure formulation, then amounts to computing a disintegration of the prior with respect to the observation operator. The existence of the disintegration is guaranteed by Theorem 3.11 in Tarieladze and Vakhania (2007), which we will here generalize to non-centered measures. Explicit formulae for the posterior mean and covariance can be obtained through the use of *representing sequences*, which we quickly introduce before presenting the disintegration theorem.

Definition 10 (Tarieladze and Vakhania (2007))**.** Given a Banach space $\mathbb{F}$ and a symmetric positive operator $R : \mathbb{F}^* \to \mathbb{F}$, a family $(f_i^*)_{i \in I}$ of elements of $\mathbb{F}^*$ is called R-representing if the following two conditions hold:

- R-orthogonality: $\langle Rf_i^*, f_j^* \rangle = \delta_{ij}$,
- spanning property: $\sum_{i \in I} \langle Rf_i^*, f^* \rangle^2 = \langle Rf^*, f^* \rangle, \ \forall f^* \in \mathbb{F}^*$.

Theorem 8. *Let $\mathbb{F}$, $\mathbb{Y}$ be real separable Banach spaces and μ be a Gaussian measure on the Borel σ-algebra $\mathcal{B}(\mathbb{F})$ with mean element $m_\mu \in \mathbb{F}$ and covariance operator $C_\mu : \mathbb{F}^* \to \mathbb{F}$. Let also $G : \mathbb{F} \to \mathbb{Y}$ be a bounded linear operator. Then, provided that the operator $C_\nu := GC_\mu G^* : \mathbb{Y}^* \to \mathbb{Y}$ has finite rank q, there exists a continuous affine map $\tilde{m}_\mu : \mathbb{Y} \to \mathbb{F}$, a symmetric positive operator $\tilde{C}_\mu : \mathbb{F}^* \to \mathbb{F}$ and a disintegration $\left(\mu_{|G=y}\right)_{y \in \mathbb{Y}}$ of μ with respect to G such that for each $y \in \mathbb{Y}$ the measure $\mu_{|G=y}$ is Gaussian with mean element $\tilde{m}_\mu(y)$ and covariance operator $\tilde{C}_\mu$. Furthermore, for any C_ν-representing sequence y_i^*, $i = 1, ..., q$, these mean and covariance are equal to*

$$\tilde{m}_\mu(y) = m_\mu + \sum_{i=1}^{q} \langle y - Gm_\mu, y_i^* \rangle\, C_\mu G^* y_i^* \tag{3.3}$$

$$\tilde{C}_\mu = C_\mu - \sum_{i=1}^{q} \langle C_\mu G^* y_i^*, \cdot \rangle\, C_\mu G^* y_i^*. \tag{3.4}$$

The mean element also satisfies $G\tilde{m}_\mu(y) = y$ for all $y \in \mathbb{Y}_0 := Gm_\mu + C_\nu(\mathbb{Y}^*)$.

For further developments, it will be useful to introduce the pushforward measure $\nu := G_\sharp \mu$, which is a Gaussian measure on $\mathbb{Y}$ with mean element Gm_μ and covariance operator C_ν.

Remark 4. In the case where $\mathbb{F}$ is a finite-dimensional Hilbert space of dimension q, one can explicitly compute an R-representing sequence by defining $f_i^* := R^{-1/2}e_i,\ i = 1, ..., q$ where $e_i,\ i = 1, ..., q$ is an orthonormal basis of $\mathbb{F}$ (see Section 3.7 for a proof). This fact will be used to link the posterior provided by Theorem 8 to the usual formulae for Gaussian processes in the case of finite-dimensional data.

Using Lemma 1 we can translate the disintegration provided by Theorem 8 to the language of Gaussian processes in the case where $\mathbb{F}$ is the Banach space $C(D)$ of continuous functions over a compact metric space D:

Corollary 4. *Let Z be a Gaussian process on some domain D with trajectories in a space $\mathbb{F}$ such that either of the equivalence theorems Theorem 3 or Theorem 5 hold. Furthermore, let $G : \mathbb{F} \to \mathbb{Y}$ be a linear bounded operator into a real separable Banach space $\mathbb{Y}$. Denote by C_μ the covariance operator of the measure associated to the process Z. Provided the operator $C_\nu := GC_\mu G^*$ has finite rank q, then, for all $y \in \mathbb{Y}$ the conditional law of Z given $GZ = y$ is Gaussian with mean and covariance function given by, for all $x, x' \in D$:*

$$\tilde{m}_x(y) = \langle \tilde{m}_\mu(y), \delta_x \rangle = m_x + \sum_{i=1}^{q} \langle y - Gm., y_i^* \rangle (C_\mu G^* y_i^*)|_x,$$

$$\tilde{k}(x, x') = \langle \tilde{C}_\mu \delta_{x'}, \delta_x \rangle = k(x, x') - \sum_{i=1}^{q} (C_\mu G^* y_i^*)|_{x'}\ (C_\mu G^* y_i^*)|_x,$$

where m_x denotes the mean function of Z and $Gm.$ denotes application of the operator G to the mean function seen as an element of $\mathbb{F}$ and $(y_i^*)_{i=1,...,q}$ is any C_ν-representing sequence.

Link to Finite-Dimensional Case: When G maps into a finite-dimensional Euclidean space and $\mathbb{F} = C(D)$ for some compact metric space D, then one can explicitly compute representing sequences and duality pairings, allowing the conditional mean and covariance in Corollary 4 to be entirely written in terms of the prior mean and covariance function of the process, making the link to the Gaussian process conditioning formulae as found for example in Tarantola and Valette (1982). Indeed, since the dual of $C(D)$ is the space of Radon measures on D, any bounded linear operator $G : C(D) \to \mathbb{R}^q$ may be written as a collection of integral operators $GZ = \left(\int_D Z_x d\lambda_i(x)\right)_{i=1,...,q}$ where the λ_i's are Radon measures on D. This special form allows us to compute closed-from expressions for the conditional mean and covariance.

Corollary 5. *Consider the situation of Corollary 4 and let $G : \mathbb{F} \to \mathbb{R}^q$. Then the conditional law of Z given $GZ = y$ is Gaussian with mean and covariance function given by, for all $x, x' \in D$:*

$$\tilde{m}_x(y) = m_x - K_{xG} K_{GG}^{-1} (y - Gm.), \tag{3.5}$$

$$\tilde{k}(x, x') = k(x, x') - K_{xG} K_{GG}^{-1} K_{x'G}^T \tag{3.6}$$

where we have defined the following vectors and matrices:

$$K_{xG} := (G_i k(\cdot, x))_{i=1,\dots,q}^T \in \mathbb{R}^{1\times q}, \tag{3.7}$$

$$K_{GG} := (G_i (G_j k(\cdot,\cdot)))_{i,j=1,\dots,q} \in \mathbb{R}^{q\times q}, \tag{3.8}$$

where $k(\cdot,\cdot)$ denotes the covariance function of Z. This corollary provides a Gaussian measure-based justification to previously used formulae (Särkkä, 2011; Jidling et al., 2018; Purisha et al., 2019; Longi et al., 2020).

The above corollary provides rigorous formulae for the conditional law under linear operator observations when the GP has trajectories that lie either in $C(D)$ or in some RKHS.

Sequential Disintegrations and Update: We now turn to the situation where several stages of conditioning are performed sequentially. Let again $\mathbb{F}$ be a real separable Banach space and consider two bounded linear operators $G_1 : \mathbb{F} \to \mathbb{Y}_1$ and $G_2 : \mathbb{F} \to \mathbb{Y}_2$, where $\mathbb{Y}_1$ and $\mathbb{Y}_2$ are also real separable Banach spaces. Then, if one views these operators as defining two stages of observations, there are two ways in which one can compute the posterior.

- On the one hand, one can compute it in two steps by first computing the disintegration of μ under G_1 and then, for each $y_1 \in \mathbb{Y}_1$, compute the disintegration of $\mu_{|G_1=y_1}$ under G_2.
- On the other hand, one can compute it in one go by considering the disintegration of μ with respect to the *bundled* operator $G : \mathbb{F} \to \mathbb{Y}_1 \bigoplus \mathbb{Y}_2,\ f \mapsto (G_1(f), G_2(f))$. From now on, we will denote this operator by $G_1 \oplus G_2$.

We show that these two approaches yield the same disintegration, as guaranteed by the following theorem.

Theorem 9. *Let $\mathbb{F}, \mathbb{Y}_1, \mathbb{Y}_2$ be real separable Banach spaces, μ be a Gaussian measure on $\mathcal{B}(\mathbb{F})$ with mean element m_μ and covariance operator $C_\mu : \mathbb{F}^* \to \mathbb{F}$. Also let $G_1 : \mathbb{F} \to \mathbb{Y}_1$ and $G_2 : \mathbb{F} \to \mathbb{Y}_2$ be bounded linear operators. Suppose that both the operators $G_1 C_\mu G_1^*$ and $G_2 C_\mu G_2^*$ have finite rank q_1 and q_2, respectively. Then*

$$\mu_{|(G_1,G_2)=(y_1,y_2)} = \left(\mu_{|G_1=y_1}\right)_{|G_2=y_2},$$

where the equality holds for almost all $(y_1, y_2) \in \mathbb{Y}_1 \bigoplus \mathbb{Y}_2$ with respect to the push-forward measure $(G_1 \oplus G_2)_\sharp \mu$ on $\mathbb{Y}_1 \bigoplus \mathbb{Y}_2$.

This theorem can be viewed as a measure-theoretic counterpart to the update formulae for GPs. Since both disintegrating measures are equal, it follows that their moments are equal too, we can thus characterize sequential disintegration in terms of mean element and covariance operator. Indeed, for the special case of GPs with trajectories in the Banach space of continuous functions on a compact domain with finite-dimensional data, we can provide explicit update formulate, this yields, using Corollary 5:

Corollary 6. *Let Z be a Gaussian process on a compact metric space D with continuous trajectories. Consider two observation operators $G_1 : C(D) \to \mathbb{R}^{q_1}$, $(G_1 Z)_i = \int_D Z_x d\lambda_i^{(1)}$ and $G_2 : C(D) \to \mathbb{R}^{q_2}$, $(G_2 Z)_i = \int_D Z_x d\lambda_i^{(2)}$. Denote by $m.$ and $k(\cdot,\cdot)$ the mean and covariance function of Z. Then, for any $y = (y_1, y_2) \in \mathbb{R}^{q_1+q_2}$ and any $x, x' \in D$, we have:*

$$\tilde{m}_x(y) = m_x + K_{xG_1} K_{G_1G_1}^{-1} (y_1 - G_1 m.) + K_{xG_2} \left(\tilde{K}^{(1)}_{G_2G_2}\right)^{-1} \left(y_2 - G_2 \tilde{m}^{(1)}_{\cdot}\right),$$
$$\tilde{k}(x,x') = k(x,x') - K_{xG_1} K_{G_1G_1}^{-1} K_{x'G_1}^T - \tilde{K}^{(1)}_{xG_2} \left(\tilde{K}^{(1)}_{G_2G_2}\right)^{-1} \left(\tilde{K}^{(1)}_{x'G_2}\right)^T,$$

where $G := (G_1, G_2)$ and $\tilde{m}^{(1)}$ denotes the conditional mean of Z given $G_1 Z = y_1$ as given by Corollary 5. Also $\tilde{K}^{(1)}_{G_2G_2}$ and $\tilde{K}^{(1)}_{xG_2}$ denote the same matrices as in Eqs. (3.7) *and* (3.8) *with the prior covariance $k(\cdot,\cdot)$ replaced by the conditional covariance of Z given $G_1 Z$.*

Infinite Rank Data: For the sake of completeness, we also consider sequential conditioning in the presence of 'infinite rank data'. That is, we want to adapt Theorem 8 and its corollaries, as well as Theorem 9 to the case where $C_\nu := GC_\mu G^* : \mathbb{Y}^* \to \mathbb{Y}$ does not have finite rank. Thanks to (Tarieladze and Vakhania, 2007, Lemma 3.5) we are still able to find a C_ν-representing sequence and (Tarieladze and Vakhania, 2007, Lemma 3.4) guarantees the convergence of the series defining the covariance operator. The main difference compared to the finite rank case is that we can only define the disintegration on a full measure subspace of the data:

Theorem 10. *Let $\mathbb{F}$, $\mathbb{Y}$, μ, G, ν and C_ν be as in Theorem 8 and assume that C_ν has infinite rank. Then there exists a subspace $\mathbb{Y}_0$ of $\mathbb{Y}$ with $\nu(\mathbb{Y}_0) = 1$ and a disintegration $\left(\mu_{|G=y}\right)_{y\in\mathbb{Y}_0}$ of μ with respect to G such that for each $y \in \mathbb{Y}_0$ the measure $\mu_{|G=y}$ is Gaussian with mean element and covariance operator:*

$$\tilde{m}_\mu(y) = m_\mu + \sum_{i=1}^{\infty} \langle y - Gm_\mu, y_i^* \rangle \, C_\mu G^* y_i^* \tag{3.9}$$
$$\tilde{C}_\mu = C_\mu - \sum_{i=1}^{\infty} \langle C_\mu G^* y_i^*, \cdot \rangle \, C_\mu G^* y_i^*, \tag{3.10}$$

where $(y_i^*)_{i\in\mathbb{N}}$ is any C_ν-representing sequence. Furthermore, the map $\tilde{m}_\mu : \mathbb{Y}_0 \to \mathbb{F}$ is continuous and affine and the mean element satisfies $G\tilde{m}_\mu(y) = y$ for all $y \in \mathbb{Y}_0 := Gm_\mu + C_\nu(\mathbb{Y}^*)$.

Concerning the transitivity of disintegrations in the infinite rank data setting, one sees that Theorem 9 holds with only slight modifications. Indeed, the only necessary adaptation is that one should restrict the joint disintegration to the direct sum of the subspaces where the individual disintegrations are defined, but since those are of full measure, the conclusion of the theorem still holds.

Theorem 11. *Let $\mathbb{F}, \mathbb{Y}_1, \mathbb{Y}_2$ be real separable Banach spaces, μ be a Gaussian measure on $\mathcal{B}(X)$ with mean element m_μ and covariance operator $C_\mu : \mathbb{F}^* \to \mathbb{F}$. Also let $G_1 : \mathbb{F} \to \mathbb{Y}_1$ and $G_2 : \mathbb{F} \to \mathbb{Y}_2$ be bounded linear operators. Then there exists a subspace $\mathbb{Y}_0 := (\mathbb{Y}_0^{(1)}, \mathbb{Y}_0^{(2)}) \subset \mathbb{Y}$ such that $\nu(\mathbb{Y}_0) = 1$ and for all $(y_1, y_2) \in (\mathbb{Y}_1^{(0)}, \mathbb{Y}_2^{(0)})$ we have:*

$$\mu_{|(G_1,G_2)=(y_1,y_2)} = \left(\mu_{|G_1=y_1}\right)_{|G_2=y_2}.$$

This theorem provides a rigorous basis for Gaussian process update in the case of infinite rank data. We stress that assimilation of such data can be theoretically challenging when using the standard Gaussian process framework, which relies on linear combinations of pointwise field evaluations to define conditional laws. We believe the above showcases the convenience of the measure-disintegration framework and how it can handle such type of data more naturally. We hope this can serve as a basis for further contributions.

As a final byproduct, one can write update formulae for sequential conditioning (disintegration) of Gaussian measures in terms of their moments. Denoting by $m_\mu^{(1)}(y_1)$ and $C_\mu^{(1)}$ the mean element and covariance operator of the disintegrating measure $\mu_{|G_1=y_1}$ and by $m_\mu^{(1\oplus 2)}(y_1,y_2)$, respectively $C_\mu^{(1\oplus 2)}$ those of the disintegration measure $\mu_{|(G_1,G_2)=(y_1,y_2)}$ one obtains the following corollary.

Corollary 7. *Consider the same setting as Theorem 11 and let $(y_i^{*(2)})_{i=1,\dots,p_2}$ be any $G_2C_\mu^{(1)}G_2^*$-representing sequence. Then the mean element and covariance operator of the disintegrating measure $\mu_{|(G_1,G_2)=(y_1,y_2)}$ can be written in terms of the moments of the intermediate disintegrating measure $\mu_{|G_1=y_1}$ as:*

$$m_\mu^{(1\oplus 2)}(y_1,y_2) = m_\mu^{(1)}(y_1) + \sum_{i=1}^{\infty} \left\langle y_2 - G_2 m_\mu^{(1)}(y_1), y_i^{(2)*} \right\rangle C_\mu^{(1)} G_2^* y_i^{(2)*}$$

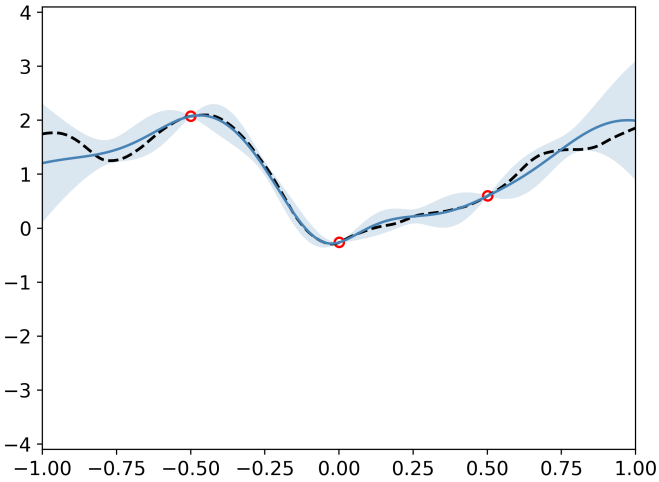
$$C_\mu^{(1\oplus 2)} = C_\mu^{(1)} - \sum_{i=1}^{\infty} \left\langle C_\mu^{(1)} G_2^* y_i^{(2)*}, \cdot \right\rangle C_\mu^{(1)} G_2^* y_i^{(2)*},$$

where the equalities hold for almost all $(y_1,y_2) \in \mathbb{Y}_1 \bigoplus \mathbb{Y}_2$ with respect to $\mu \circ (G_1,G_2)^{-1}$.

Note that this corollary provides an extension to Gaussian measures and operator observations of the well-known kriging update formulae (Chevalier et al., 2014b) and can be viewed as subsuming various Gaussian conditioning update formulae under a rigorous and abstract theoretical framework.

Example (continued)**.** We now come back to the example from the introduction to demonstrate the machinery developed in the two preceding sections. Assume that we want to add derivative observation at $x = 0$.
First, in order to apply the disintegration theorems, we need to make sure that the observation operator under consideration is a bounded operator on a Banach space in which the path of the prior lie with probability one. In this example, the prior that was used was a Matérn 5/2 GP with lengthscale parameter $\lambda = 0.4$. According to Corollary 2, the path of the prior almost surely lie in the Sobolev space $H_2([-1,1])$, so taking $\mathbb{F} = H_2([-1,1])$ ensures that the observation operators are bounded (integral and Fourier observations are bounded since the domain is compact and the paths continuous).

Now, $H_2([-1,1])$ is a RKHS and thus by Theorem 5 the Gaussian measure - Gaussian process correspondence is applicable. Furthermore, the 7 observations (3 pointwise + 1 integral + 2 Fourier + 1 derivative) considered can be described by a bounded operator between separable Banach spaces $G : H_2([-1,1]) \to \mathbb{R}^7$, so that the disintegration framework from Section 3.4 can be used. Finally, using the updated formulae (Corollary 7) one can express the posterior mean and covariance after inclusion of the derivative observation as an update of the one after assimilation of the previous observations:



(a) pointwise + integral + Fourier + derivative

Figure 3.2: Continuation of the introductory example with addition of derivative observation at $x = 0$.

$$\tilde{m}^{(7)}_{x_1}(y_7) = \tilde{m}^{(6)}_{x_1}(y_1, ..., y_6) + \frac{d}{dx'}\tilde{k}^{(6)}(x_1, x')|_{x'=0} \left(\frac{d}{dx}\frac{d}{dx'}\tilde{k}^{(6)}(x, x')|_{x,x'=0} \right)^{-1} (y_7 - \frac{d}{dx}\tilde{m}^{(6)}_x(y_1, ..., y_6)|_{x=0})$$

$$\tilde{k}^{(7)}(x_1, x_2) = \tilde{k}^{(6)}(x, x) - \frac{d}{dx}\tilde{k}^{(6)}(x_1, x)|_{x=0} \left(\frac{d}{dx}\frac{d}{dx'}\tilde{k}^{(6)}(x, x')|_{x,x'=0} \right)^{-1} \frac{d}{dx}\tilde{k}^{(6)}(x, x_2)|_{x=0}$$

where $\tilde{m}^{(6)}_{x_1}(y_1, ..., y_6)$ and $\tilde{k}^{(6)}(x_1, x_2)$ denote the mean and covariance function after inclusion of the first 6 observations. Note that the correspondence between the mean element and covariance operator of the induced measure and the mean and covariance function of the process (Lemma 1) can be used since the pointwise evaluation functionals belong to the dual of $H_2([-1, 1])$. This example demonstrates how the Gaussian measure framework can be used to provide a thorough theoretical grounding to previously known techniques (Solak et al., 2003; Ribaud, 2018; Agrell, 2019).

3.5 Conclusion

By bridging recent results about GP sample path properties with the framework of Gaussian measures, we provide a formulation of sequential data assimilation of linear operator data under Gaussian models in the language of disintegrations of measures. We show equivalence of the Gaussian process and Gaussian measure approaches and generalize the GP update formulae to disintegrations. While providing a purely functional formulation of the assimilation process, the framework of disintegrations also allows for a more rigorous abstract treatment of the conditional law. This can be leveraged to provide fast update formulae for GP under linear operator observations (Travelletti et al., 2023) and we hope it can serve as foundations for further theoretical inquiries and practical developments in probabilistic function modelling.

3.6 Appendix A: Proofs of Equivalence of Gaussian Process and Gaussian measure

We here briefly recall the theorems and definitions needed to prove our main results, and present the proofs. For the functional analysis background, we refer the reader to Folland (2013) and to Tarieladze and Vakhania (2007); Vakhania et al. (1987) for the background about Gaussian measures. The theorems for equivalence between Gaussian processes and Gaussian measures are adapted from Rajput and Cambanis (1972), while the one for conditioning / disintegration of Gaussian measures are adapted from Tarieladze and Vakhania (2007).

Most of this chapter will be concerned with random variables taking values in the space of continuous function $C(D)$, where D is a compact metric space. When endowed with the sup-norm, $C(D)$ turns into a Banach space. This space enjoys two useful properties:

1. $C(D)$ is separable, and as a consequence, the Borel σ-algebra and the cylindrical σ-algebra on $C(D)$ agree.

2. The dual space $C(D)^*$ is the space of Radon measures on D and (by Riesz-Markov-Kakutani (Rudin, 1974)) for all $\ell \in C(D)^* : \exists \lambda$ Radon measure on D such that
$$\forall f \in C(D) : \ell(f) = \int f d\lambda.$$

In order to prove Theorem 3 and Theorem 4, we first recall a classic approximation result for continuous real-valued functions on compact metric spaces that will be useful for proving measurability properties and Gaussianity of the measure induced by a GP. For reference, see (Folland, 2013, Theorem 2.10).

Lemma 2. *Let D be a compact metric space and $f : D \to \mathbb{R}$ be continuous. Then, there exists a sequence of simple functions f_n converging to f uniformly on D. For each n, the approximating function can be written as:*

$$f_n = \sum_{k=0}^{K(n)} f\left(t_k^{(n)}\right) \mathbb{1}_{A_k^{(n)}}, \tag{3.11}$$

where $K(n) \in \mathbb{N}$, $t_k^{(n)} \in D$ and the $A_k^{(n)}$'s are Borel measurable sets for all k.

We now show that, for stochastic processes on compact metric spaces, having continuous sample paths is enough to ensure product measurability.

Theorem 12. *Let $(\Omega, \mathcal{F}, \mathbb{P}; Z(x;\omega), x \in D)$ be a stochastic process on a compact metric space D with continuous sample paths. Then it is measurable as a mapping $(D \times \Omega,\ \mathcal{B}(D) \times \mathcal{F}) \to (\mathbb{R},\ \mathcal{B}(\mathbb{R}))$ (product measurable).*

Proof. This is a direct consequence of Gowrisankaran (1972, Theorem 2). □

We now have all the ingredients to prove the main theorems about equivalence of process and measure.

Proof. (Theorem 3) By Theorem 12, the only thing left to prove is that for all $\ell \in C(D)^*$ the real random variable $\ell \circ \Phi$ is Gaussian.
By the Riesz-Markov representation theorem, there exists a Radon measure λ on D representing ℓ. Now, for each $\omega \in \Omega$, we use Lemma 2 to get a uniform approximation $Z_n(\cdot;\omega) \to Z(\cdot;\omega)$ as in Equation (3.11). We then have:

$$\begin{aligned}\ell \circ \Phi(\omega) &= \ell\left(\lim_{n\to\infty} Z_n(\cdot;\omega)\right) = \lim_{n\to\infty} \int \sum_{k=0}^{K(n)} Z\left(t_k^{(n)};\omega\right) \mathbb{1}_{A_k^{(n)}} d\lambda \\ &= \lim_{n\to\infty} \sum_{k=0}^{K(n)} Z\left(t_k^{(n)};\omega\right) \lambda\left(A_k^{(n)}\right).\end{aligned}$$

Now, as a convergent series of Gaussian random variables, the above is Gaussian (use characteristic functions and Lévy convergence theorem). □

We now turn to the proof of Theorem 4.

Proof. (Theorem 4) Let $\Omega = C(D)$ and $\mathcal{F}$ be the Borel sigma algebra on $C(D)$ and define a collection of random variables

$$Z_x : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})), \ \omega \mapsto \delta_x(\omega)$$

for all $x \in D$. Since for all $x \in D$, the pointwise evaluation functionals δ_x belong to the dual of $C(D)$, we have that Z_x is a Gaussian real random variable for all $x \in D$. Now, for $x_1, ..., x_n \in D$, any linear combination of the components of the vector $(Z_{x_1}, ..., Z_{x_n})$ may be written an element of $C(D)^*$, and will hence be Gaussian distributed by Gaussianity of the measure. This shows that Z is a Gaussian process on D. □

From Theorems 3 and 4 and a simple change of variable, we have that, if Z is the process induced by a Gaussian measure μ on $C(D)$, then for any $x \in D$, we have

$$\langle m_\mu, \delta_x \rangle = \int_{C(D)} \delta_x(f) d\mu(f) = \int_\Omega \delta_x(Z(\cdot,\omega)) d\mathbb{P}(\omega) = \mathbb{E}[Z_x] \tag{3.12}$$

and the same is true if Z is a GP on D with trajectories in $C(D)$ and μ is the measure induced by the process. This allows us to translate everything from process to measure and back without needing to worry about the details. Finally, using the fact that the pointwise evaluation functionals belong to the dual we may also prove Lemma 1 about the correspondence between mean element and covariance operator of the induced measure and mean and covariance function of the process.

Proof. (Lemma 1) For $x, x' \in D$, let:

$$\begin{aligned}\langle \delta_x, C_\mu \delta_{x'} \rangle &= \int_{C(D)} \left(f(x') - \langle m_\mu, \delta_{x'} \rangle\right)\left(f(x) - \langle m_\mu, \delta_x \rangle\right) d\mu(f) \\ &= \mathbb{E}[Z_x Z_{x'}] - \mathbb{E}[Z_x]\,\mathbb{E}[Z_{x'}].\end{aligned}$$

where the last equality is a consequence of Equation (3.12). □

The extension of Theorem 3 and Theorem 4 to processes and measures on RKHS is straightforward. Indeed, the measure-to-process correspondence follows directly from the fact that the evaluation functionals belong to the dual of the RKHS. For the process-to-measure correspondence, the crucial property is the Gaussianity of linear functionals of the field, which in a RKHS $\mathcal{H}$ is automatically satisfied since any linear functional can be expressed as an infinite linear combination of reproducing kernel values, which in turn act as evaluation functionals:

$$\langle \ell, Z \rangle = \left\langle \sum_{i=1}^{\infty} a_i k(x_i, \cdot), Z \right\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} a_i Z_{x_i},$$

which, as a convergent sum of Gaussian random variables, is Gaussian.

3.7 Appendix B: Conditioning, Disintegration and Link to Finite-Dimensional Formulation

We now turn to the proof of Theorem 8.

Proof. (Theorem 8) To prove the theorem, we have to adapt the proof of Tarieladze and Vakhania (2007)[Theorem 3.11] to the non-centered case. Compared to the original theorem, the conditional covariance operator $\tilde{C}_\mu$ hasn't changed, whereas the conditional mean $\tilde{m}_\mu(y)$ clearly still defines a continuous mapping satisfying $G\tilde{m}_\mu(y) = y$ for all y in the range of C_ν. Hence, for all $y \in \mathbb{Y}$, we can still use Tarieladze and Vakhania (2007)[Lemma 3.8] to define $\mu_{|G=y}$ as a Gaussian measure having mean element $\tilde{m}_\mu(y)$ and covariance operator $\tilde{C}_\mu$. What is left to check is that it satisfies the conditions in Definition 9 to be a disintegration of μ with respect to G.

In the following, let $y \in \mathbb{Y}$ and $A \in \mathcal{A}$ be arbitrary.

- The measurability of the mapping $y \mapsto \mu_{|G=y}(A)$ for fixed A holds since, compared to the centered case, the conditional mean $\tilde{m}_\mu(y)$ is only translated by an element that does not depend on y.

- Define $\mathbb{Y}_0 := Gm_\mu + C_\nu(\mathbb{Y}^*)$. We have $\mu \circ G^{-1}(\mathbb{Y}_0) = 1$ by Tarieladze and Vakhania (2007)[Lemma 3.3] and Tarieladze and Vakhania (2007)[Corollary 3.7]. Following the exact same reasoning as in the proof of Tarieladze and Vakhania (2007)[Theorem 3.11] we have that $\mu_y(G^{-1}(y)) = 1$.

- By Tarieladze and Vakhania (2007)[Proposition 3.2], the last thing we have to check is that

$$\hat{\mu}(g^*) = \int_{\mathbb{Y}} \hat{\mu}_{|G=y}(g^*)\, d\nu(y), \ \forall g^* \in \mathbb{F}^*,$$

  where $\hat{\mu}(\cdot)$ denotes the characteristic functional of μ (see Tarieladze and Vakhania (2007)[Section 3.2]. Compared to the original proof, only the mean element is changed, so for the sake of simplicity we only consider the steps

of the proof that differ from the original ones.
We have that

$$\int_{\mathbb{Y}} \exp\left[i\left\langle \tilde{m}_\mu(y), g^*\right\rangle\right] d\nu(y) = \exp\left[i\left\langle m_\mu, g^*\right\rangle\right] \cdot \int_{\mathbb{Y}} \exp\left[i\left\langle \sum_{i=1}^{n} \left\langle y - Gm_\mu, y_i^*\right\rangle C_\mu G^* y_i^*, g^*\right\rangle\right] d\nu(y),$$

which, after a change of variable $y \mapsto y - Gm_\mu$ can be seen to be the characteristic function of a centered Gaussian measure with covariance C_μ by following the same argument as in the original proof (the same argument is presented in more detail in the proof of the next theorem).

□

We can now turn to the proof of our central result Theorem 9 about the transitivity of disintegrations.

Proof. (Theorem 9) Since, by construction, $\mu_{|(G_1,G_2)}$ is a disintegration of μ with respect to G, by uniqueness of disintegrations (see Remark 3.12 in Tarieladze and Vakhania (2007)), we only have to prove that the family

$$\left(\left(\mu_{|G_1=y_1}\right)_{|G_2=y_2}\right)_{(y_1,y_2)\in\mathbb{Y}_1 \bigoplus \mathbb{Y}_2}$$

defines a disintegration of μ with respect to $G_1 \oplus G_2$. First a word of caution: there exist no *canonical* norm on the direct sum of Banach spaces. However, there are several norms on the direct sum that induce the product topology (see, for example, Exercise 1.30 in Bühler and Salamon (2018)). We here assume that $\mathbb{Y}_1 \bigoplus \mathbb{Y}_2$ has been endowed with any of these. Then, the Borel σ-algebra on the direct sum is given by the product of the Borel σ-algebras of the components (see p.244 of Billingsley (1999)).

After one step of disintegration, one obtains the (family of) disintegrating measure $\mu_{|G_1=y_1}$ which is Gaussian and whose mean element and covariance operator we denote by $m_\mu^{(1)}(y_1)$ and $C_\mu^{(1)}$. Before proceding further, we introduce the Gaussian measures $\nu_1 := (G_1)_\sharp\mu$ and $\nu_{2,y_1} := (G_2)_\sharp(\mu_{|G_1=y_1})$ on $\mathbb{Y}_1$, respectively $\mathbb{Y}_2$. These measures are Gaussian, with mean element and covariance operator $G_1\mu$ and $C_{\nu_1} = G_1 C_\mu G_1^*$, respectively $G_2 m_\mu^{(1)}(y_1)$ and $C_{\nu_2} = G_2 C_\mu^{(1)} G_2^*$. Note that the assumptions of the theorem guarantee that C_{ν_2} has finite rank p_2, for some p_2. Now, by construction, for any $(y_1, y_2) \in Y_1 \bigoplus Y_2$, the measure $\left(\mu_{|G_1=y_1}\right)_{|G_2=y_2}$ is a Gaussian measure with mean element

$$m_\mu^{(1,2)} := m_\mu^{(1)}(y_1) + \sum_{i=1}^{q_2} \left\langle y_2 - G_2 m_\mu^{(1)}(y_1), y_i^{(2)*}\right\rangle C_\mu^{(1)} G_2^* y_i^{(2)*},$$

and covariance operator

$$C_\mu^{(1,2)} := C_\mu^{(1)} - \sum_{i=1}^{q_2} \left\langle C_\mu^{(1)} G_2^* y_i^{(2)*}, \cdot\right\rangle C_\mu^{(1)} G_2^* y_i^{(2)*},$$

where $(y_i^{(2)*})_{i=1,\dots,q_2}$ is any C_{ν_2}-representing sequence. Since for all $y_1 \in \mathbb{Y}_1$ the measure $\mu_{|G_1=y_1}$ is Gaussian, we have by Theorem 8 that $\left(\mu_{|G_1=y_1}\right)_{|G_2=y_2}$ is Gaussian.

To ease the notation for the coming proofs, we isolate the update components stemming from G_1, respectively G_2 in the conditional covariance, by rewriting it as:

$$C_\mu^{(1,2)} = C_\mu - R_1 - R_2, \tag{3.13}$$

where $C_\mu^{(1)} = C_\mu - R_1$, so that $C_\mu^{(1,2)} = C_\mu^{(1)} - R_2$. We now check that the family of measures constructed above satisfies the conditions of Definition 9 for it to be a disintegration of μ with respect to $G_1 \oplus G_2$. We begin by checking the *measurability* and *concentration on the fiber* properties:

- For fixed A, the mapping $(y_1, y_2) \mapsto \left(\mu_{|G_1=y_1}\right)_{G_2=y_2}(A)$ is an addition of a $\mathcal{B}(\mathbb{Y}_1)$-measurable mapping with a $\mathcal{B}(\mathbb{Y}_2)$-measurable mapping, and, as such, measurable with respect to the product σ-algebra.

- Let $\mathbb{Y} := \mathbb{Y}_1 \bigoplus \mathbb{Y}_2$ and note that $\mathbb{Y}^* = \mathbb{Y}_1^* \bigoplus \mathbb{Y}_2^*$ (dual of direct sum is the direct sum of the duals). Then define $\mathbb{Y}_0 = Gm_\mu + GC_\mu G^*\left(\mathbb{Y}_1^* \bigoplus \mathbb{Y}_2^*\right)$. Note that the Gaussian measure $\mu \circ G^{-1}$ has mean Gm_μ and covariance operator $GC_\mu G^*$, hence $\mu \circ G^{-1}(\mathbb{Y}_0) = 1$ by Tarieladze and Vakhania (2007)[Lemma 3.3].

  For any $(y_1, y_2) \in \mathbb{Y}_0$ we have that the Gaussian measure $\left(\mu_{|G_1=y_1}\right)_{|G_2=y_2} \circ G^{-1}$ has covariance operator $G\tilde{C}_\mu^{(1,2)}G^*$. Computing the operator componentwise, we have that:

$$G_2\tilde{C}_\mu^{(1,2)}G_2^* = G_2\tilde{C}_\mu^{(1)}G_2^* - \sum_{i=1}^{q_2}\left\langle \tilde{C}_\mu^{(1)}G_2^*y_i^{(2)*}, G_2^*\right\rangle G_2\tilde{C}_\mu^{(1)}G_2^*y_i^{(2)*} = 0,$$

  where the last equality follows from Tarieladze and Vakhania (2007)[Lemma 3.4, (c)] since $y_i^{(2)*}$ is a $G_2\tilde{C}_\mu^{(1)}G_2^*$-representing sequence. An analogous computation for the other components shows that they all vanish.

Finally, for the *mixing property*, we proceed as in the last proof by showing that the characteristic functional of the original measure can be written as a mixing of the characteristic functionals of the disintegrating measure, i.e. we show that, for any $g^* \in \mathbb{F}^*$:

$$\hat{\mu}(g^*) = \int_{\mathbb{Y}_1 \bigoplus \mathbb{Y}_2} \hat{\mu}_{y_1,y_2}^{(1,2)}(g^*)\, d\left(\mu \circ (G_1 \oplus G_2)^{-1}\right)(y_1, y_2), \tag{3.14}$$

where we use the compact notation $\mu_{y_1,y_2}^{(1,2)} := \left(\mu_{|G_1=y_1}\right)_{|G_2=y_2}$. Before proceeding any further, we want to rewrite the integral over the direct sum as a double integral. To that end, we use the following result.

Lemma 3. *For any $\mu \circ (G_1 \oplus G_2)^{-1}$-integrable function $h : \mathbb{Y}_1 \times \mathbb{Y}_2 \to \mathbb{R}$, we have:*

$$\int_{\mathbb{Y}_1 \bigoplus \mathbb{Y}_2} h(y_1, y_2) d\left(\mu \circ (G_1 \oplus G_2)^{-1}\right)(y_1, y_2) = \int_{\mathbb{Y}_1}\int_{\mathbb{Y}_2} h(y_1, y_2) d\nu_{2,y_1}(y_2) d\nu_1(y_1).$$

Using the above integration lemma and the general form of the characteristic function of a Gaussian measure, we can rewrite the right-hand side of Eq. (3.14) as:

$$\int_{\mathbb{Y}_1}\int_{\mathbb{Y}_2}\exp\left[i\left\langle m_\mu^{(1,2)}(y_1,y_2),g^*\right\rangle-\frac{1}{2}\left\langle C_\mu^{(1,2)}g^*,g^*\right\rangle\right]d\nu_{2,y_1}(y_2)d\nu_1(y_1). \tag{3.15}$$

By factoring the exponential, the covariance term can be taken out of the integral. The integral over the mean term can then be expanded using the formulae for the conditional mean and one has to compute:

$$\int_{\mathbb{Y}_1}\int_{\mathbb{Y}_2}\exp\left[i\left\langle m_\mu^{(1)}(y_1)+\sum_{i=1}^{p_2}\left\langle y_2-G_2m_\mu^{(1)}(y_1),y_i^{(2)*}\right\rangle C_\mu^{(1)}G_2^*y_i^{(2)*},g^*\right\rangle\right]d\nu_{2,y_1}(y_2)d\nu_1(y_1)$$

Now, by considering the transformation $T:y_2\mapsto y_2-G_2m_\mu^{(1)}(y_1)$ and noticing that $T_\sharp\nu_{2,y_1}$ is a Gaussian measure with mean 0 and covariance C_{ν_2} we can change variables and obtain:

$$\int_{\mathbb{Y}_1}\int_{\mathbb{Y}_2}\exp\left[i\left\langle m_\mu^{(1)}(y_1)+\sum_{i=1}^{p_2}\left\langle y_2,y_i^{(2)*}\right\rangle C_\mu^{(1)}G_2^*y_i^{(2)*},g^*\right\rangle\right]d(T_\sharp\nu_{2,y_1})(y_2)d\nu_1(y_1).$$

Defining the mapping: $M_2:\mathbb{Y}_2\to\mathbb{F}$, $M_2(y_2):=\sum_{i=1}^{p_2}\left\langle y_2,y_i^{(2)*}\right\rangle C_\mu^{(1)}G_2^*y_i^{(2)*}$, we can rewrite the above term as:

$$\int_{\mathbb{Y}_1}\int_{\mathbb{Y}_2}\exp\left[i\left\langle m_\mu^{(1)}(y_1),g^*\right\rangle+i\left\langle y_2,M_2^*(g^*)\right\rangle\right]d(T_\sharp\nu_{2,y_1})(y_2)d\nu_1(y_1).$$

The integral over $\mathbb{Y}_2$ can be computed as a characteristic function of a Gaussian measure, yielding:

$$\int_{\mathbb{Y}_1}\exp\left[i\left\langle m_\mu^{(1)}(y_1),g^*\right\rangle\right]\left(\widehat{T_\sharp\nu_{2,y_1}}\right)\left(M_2^*\left(g^*\right)\right)d\nu_1(y_1),$$

and since $T_\sharp\nu_{2,y_1}$ is a Gaussian measure with mean 0 and covariance operator C_{ν_2}, we can compute the characteristic function:

$$\begin{aligned}\left(\widehat{T_\sharp\nu_{2,y_1}}\right)\left(M_2^*\left(g^*\right)\right)&=\exp\left[-\frac{1}{2}\left\langle C_{\nu_2}M_2^*\left(g^*\right),M_2^*\left(g^*\right)\right\rangle\right]\\&=\exp\left[-\frac{1}{2}\left\langle C_{\nu_2}\sum_{i=1}^{p_2}y_i^{(2)*}\left\langle C_\mu^{(1)}G_2^*y_i^{(2)*},g^*\right\rangle,\sum_{j=1}^{p_2}y_j^{(2)*}\left\langle C_\mu^{(1)}G_2^*y_j^{(2)*},g^*\right\rangle\right\rangle\right]\\&=\exp\left[-\frac{1}{2}\sum_{i=1}^{p_2}\left\langle C_\mu^{(1)}G_2^*y_i^{(2)*},g^*\right\rangle^2\right]=\exp\left[-\frac{1}{2}\sum_{i=1}^{p_2}\left\langle R_2g^*,g^*\right\rangle\right],\end{aligned}$$

where the penultimate equality follows from the C_{ν_2}-orthogonality of the $y_i^{(2)*}$'s, and we have used the decomposition of the conditional covariance operator Eq. (3.13). Coming back to Eq. (3.15), we are left with an integral over $\mathbb{Y}_1$:

$$\int_{\mathbb{Y}_1}\exp\left[i\left\langle m_\mu^{(1)}\left(y_1\right),g^*\right\rangle-\frac{1}{2}\left\langle C_\mu^{(1)}g^*,g^*\right\rangle\right]d\nu_1(y_1).$$

Performing the same calculation as before on the first term of the integral gets rid of the $\langle R_1 g^*, g^* \rangle$ part and we are left with:

$$\exp\left[i \langle m_\mu, g^* \rangle - \frac{1}{2} \langle C_\mu g^*, g^* \rangle \right],$$

which is the characteristic function of μ. This completes the proof. □

Link to Finite Dimensional case When the inversion data is *finite-dimensional*, that is the observation operator G maps into $\mathbb{R}^n$ and $\mathbb{R}^n$ is considered as a Banach space with respect to the 2-norm. One can then canonically identify $\mathbb{R}^n$ with its dual using the dot product: $v \mapsto \langle v, \cdot \rangle$. In the following, when elements of $\mathbb{R}^n$ are involved, the duality bracket $\langle \cdot, \cdot \rangle$ will denote the dot product, also, $e_i, i = 1, ..., n$ will be used to denote the canonical basis of $\mathbb{R}^n$. We now prove that $y_i := C_\nu^{-1/2} e_i,\ i = 1, ..., n$ forms a C_ν-representing sequence.

Proof. (Remark 4) First of all, the y_i form a C_ν-orthonormal family since

$$\langle C_\nu y_i, y_j \rangle = \langle C_\nu^{1/2} y_i, C_\nu^{1/2} y_j \rangle = \langle e_i, e_j \rangle = \delta_{ij},$$

where the first equality follows by self-adjointness of C_ν. Also remember that since here we are working over $\mathbb{R}^n$, the duality bracket denotes the dot product and $\mathbb{R}^n$ is identified with its dual. Finally, according to Tarieladze and Vakhania (2007)[Lemma 3.4], the last thing we have to show is that for any $v \in \mathbb{R}^n$: $C_\nu v = \sum_{i=1}^n \langle C_\nu y_i, v \rangle C_\nu y_i$. Note that since C_ν is a positive self-adjoint operator, the y_i's form a basis of $\mathbb{R}^n$, and we can thus write $v = \sum_{i=1}^n v_i y_i$ for some component v_i. Then

$$\sum_{i=1}^n \langle C_\nu y_i, v \rangle C_\nu y_i = \sum_{i,j=1}^n \langle C_\nu y_i, v_j y_j \rangle C_\nu y_i = \sum_{i=1}^n v_i C_\nu y_i = C_\nu v$$

□

Proof. (Corollary 5) As before, let $y_i := C_\nu^{-1/2} e_i,\ i = 1, ..., n$. In order to get closed-form formulae for the posterior under such operators, we need to be able to compute the action of the adjoint G^*. We begin by recalling the definition of the adjoint of a linear operator $T : \mathbb{F} \to \mathbb{Y}$ between Banach spaces:

$$\begin{aligned} T^* : &\mathbb{Y}^* \to \mathbb{F}^* \\ &y^* \mapsto (f \mapsto \langle y^*, Tf \rangle). \end{aligned}$$

Now if we consider a (bounded) linear form $G_j : \mathbb{F} :\to \mathbb{R}$, then its adjoint is given by:

$$\begin{aligned} G_j^* : &\mathbb{R} \to \mathbb{F}^* \\ &a \mapsto (f \mapsto a \cdot G_j f). \end{aligned}$$

So the adjoint of the observation operator may be written as:

$$\begin{aligned} G^* : &\mathbb{R}^n \to \mathbb{F}^* \\ &(a_1, ..., a_n) \mapsto (f \mapsto a_1 \cdot G_1 f + ... + a_n \cdot G_n f). \end{aligned}$$

There is one last computation that we need to perform before getting the mean and covariance:

$$\left\langle C_\mu G^* y^{(i)}, \delta_x \right\rangle = \left\langle C_\mu \delta_x, G^* y^{(i)} \right\rangle = y^{(i)} \cdot G\left(C_\mu \delta_x\right) = y^{(i)} \cdot Gk(\cdot, x) = y^{(i)} \cdot K_{xG}.$$

Putting everything together we are now able to express the covariance operator:

$$\begin{aligned}
\tilde{k}(x, x') &= k(x, x') - \sum_{i=1}^{n} y^{(i)} \cdot K_{xG} y^{(i)} \cdot K_{x'G} \\
&= k(x, x') - \sum_{i=1}^{n} K_{xG}^T y^{(i)} \left(y^{(i)}\right)^T K_{x'G} \\
&= k(x, x') - \sum_{i=1}^{n} K_{xG}^T C_\nu^{-1/2} e_i e_i^T C_\nu^{-1/2} K_{x'G} \\
&= k(x, x') - K_{xG}^T K_{GG}^{-1} K_{x'G}.
\end{aligned}$$

Where we have used the fact that $\sum_{i=1}^{n} e_i e_i^T = \boldsymbol{I}_n$ and that:

$$e_i \cdot GC_\mu G^* e_j = G_i(G_j k(\cdot, \cdot)).$$

Note that this last step requires one to explicitly compute the action of the G_i's on the covariance operator C_μ. This can be done in the case where $\mathbb{F} = C(D)$ since the individual components on the observation operator can be written as integrals with respect to Radon measures $G_i f = \int_D f(x) d\lambda_i(x)$ or in the case where $\mathbb{F}$ is a RKHS, since then the components can be written as infinite linear combinations of pointwise evaluation functionals $G_i f = \sum_{k=1}^{\infty} a_k^{(i)} f(x_k^{(i)})$. Computing the action on the covariance operator in the general case is not trivial. The mean can be obtained through a similar argument. □

Proofs for Infinite Rank Data

Proof. (Theorem 10) As before, compared to the centered case, only the conditional mean changes. Thanks to (Tarieladze and Vakhania, 2007, Lemma 3.5) we can still select a countably infinite C_ν representing sequence $(y_i)_{i \in \mathbb{N}}$. Now define, for all $n \in \mathbb{N}$:

$$\tilde{m}_\mu^{(n)}(y) = m_\mu + \sum_{i=1}^{n} \left\langle y - Gm_\mu, y_i^* \right\rangle C_\mu G^* y_i^*. \tag{3.16}$$

Furthermore, define the spaces: $\mathbb{Y}_2 := \{y \in \mathbb{Y} : \ \tilde{m}_\mu^{(n)}(y) \text{ converges}\}$, and $\mathbb{Y}_3 := \{y \in \mathbb{Y} : \ \lim_{n\to\infty} ||y - \sum_{i=1}^{n} \langle y - Gm_\mu, y_i^* \rangle C_\nu y_i^*|| = 0\}$. We begin by showing that these subspaces of $\mathbb{Y}$ have full measure.

Claim: $\nu(Y_2) = 1$.

Proof. Our goal is to show that the random element $\tilde{m}_\mu^n$ converges ν-almost surely in $\mathbb{F}$. First, define $\xi_i := \langle y - Gm_\mu, y_i^* \rangle C_\mu G^* y_i^*$. Thanks to C_ν-orthonormality, the y_i^* are independent Gaussian random variables, and hence the ξ_i too. Hence, by Ito-Nisio (Vakhania et al., 1987, Theorem 5.2.4), we get ν-almost-sure convergence

provided we can show that there exists a random probability measure μ' on $\mathbb{F}$ such that the joint characteristic function converges to the characteristic function of μ':

$$\prod_{i=1}^{n} \hat{\mathbb{P}}_{\xi_i}(f) \to \hat{\mu'}(f), \text{ all } f \in \mathbb{F}^*.$$

By independence of the ξ_i, we have, for $f \in \mathbb{F}^*$:

$$\begin{aligned}\prod_{i=1}^{n} \hat{\mathbb{P}}_{\xi_i}(f) &= \int_{\mathbb{Y}} \exp\left[i \left\langle f, \sum_{i=1}^{n} \langle y - Gm_\mu, y_i^* \rangle C_\mu G^* y_i^* \right\rangle \right] d\nu(y) \\ &= \int_{\mathbb{Y}} \exp\left[i \left\langle y', \sum_{i=1}^{n} y_i^* \langle f, C_\mu G^* y_i^* \rangle \right\rangle \right] d\nu'(y'),\end{aligned}$$

where we have performed a change of variable $y' := y - Gm_\mu$ and hence ν' is a centered Gaussian measure with covariance operator C_ν. Now, using the characteristic function of Gaussian measures, the above is equal to:

$$\begin{aligned}\hat{\nu}\left(\sum_{i=1}^{n} y_i^* \langle f, C_\mu G^* y_i^* \rangle \right) &= \exp\left[-\frac{1}{2} \sum_{i,j=1}^{n} \langle f, C_\mu G^* y_i^* \rangle \left\langle f, C_\mu G^* y_j^* \right\rangle \left\langle C_\nu y_i^*, y_j^* \right\rangle \right] \\ &= \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \langle f, C_\mu G^* y_i^* \rangle^2 \right],\end{aligned}$$

where the last equality follows from C_ν-orthonormality of the representing sequence. We thus have:

$$\lim_{n\to\infty} \prod_{i=1}^{n} \hat{\mathbb{P}}_{\xi_i}(f) = \exp\left[-\frac{1}{2} \langle R_1 f, f \rangle \right], \tag{3.17}$$

where $R_1 := \lim_{n\to\infty} \sum_{i=1}^{\infty} \langle C_\mu G^* y_i^*, \bullet \rangle C_\mu G^* y_i^*$ is a Gaussian covariance by (Tarieladze and Vakhania, 2007, Lemma 3.4 and Proposition 3.9). The Claim follows from the fact that for any Gaussian covariance, there exists a Gaussian measure having that covariance as covariance operator (Tarieladze and Vakhania, 2007, Lemma 3.8). □

Claim: $\nu(\mathbb{Y}_3) = 1$.

Proof. Note that if $y - Gm_\mu$ can be written as $C_\nu y^*$ for some $y^* \in \mathbb{Y}^*$, then it immediately follows, by (Tarieladze and Vakhania, 2007, Lemma 3.4), that:

$$\sum_{i=1}^{\infty} \langle y - Gm_\mu, y_i^* \rangle C_\nu y_i^* = \sum_{i=1}^{\infty} \langle y^*, C_\nu y_i^* \rangle C_\nu y_i^* = C_\nu y^* = y - Gm_\mu.$$

Now, the subspace whose elements can be written as above is exactly the Cameron-Martin space $C_\nu(\mathbb{Y}^*)$. While this is a ν-null space, it is a well-known fact that its closure in $\mathbb{Y}$ has full measure, so that there exists a subset of full measure whose elements can be approximated by elements of $C_\nu(\mathbb{Y}^*)$ and thus the defining property of $\mathbb{Y}_3$ holds on a set of full measure. □

Now, we define $\mathbb{Y}_0 := \mathbb{Y}_2 \cap \mathbb{Y}_3$. We construct a disintegration $\left(\mu_{|G=y}\right)_{y\in\mathbb{Y}_0}$ as in the finite rank case, but now restricting to the subspace $\mathbb{Y}_0$ where the conditional mean is defined. What is left to check is that it satisfies the three defining properties of disintegrations (Definition 9). Property 1 holds as in the finite rank case. For Property 2, we notice that, for any $y \in \mathbb{Y}_0$:

$$G\tilde{m}_\mu(y) = \lim_{n\to\infty} \tilde{m}_\mu^{(n)}(y) = Gm_\mu - \sum_{i=1}^{\infty} \langle Gm_\mu, y_i^* \rangle C_\nu y_i^* + \sum_{i=1}^{\infty} \langle y, y_i^* \rangle C_\nu y_i^* = y,$$

since Gm_μ is the mean of ν and thus belongs to the Cameron-Martin space. Finally, for Property 3, thanks to (Tarieladze and Vakhania, 2007, Proposition 3.2), we only have to show that the characteristic function of μ writes as a mixing of the characteristic functions of the conditionals, i.e. that:

$$\hat{\mu}(f) = \int_{\mathbb{Y}} \hat{\mu}_{|G=y}(f) d\nu(y), \text{ all } f \in \mathbb{F}^*.$$

Now, for $y \in \mathbb{Y}_0$, we have that $\mu_{|G=y}$ is Gaussian, with mean $\tilde{m}_\mu(y)$ and covariance operator $C_\mu - R_1$. Hence, we have:

$$\begin{aligned}\int_{\mathbb{Y}} \hat{\mu}_{|G=y}(f) d\nu(y) &= \int_{\mathbb{Y}} \exp\left[i \langle \tilde{m}_\mu(y), f\rangle - \frac{1}{2}\langle C_\mu, f\rangle + \frac{1}{2}\langle R_1 f, f\rangle \right] d\nu(y) \\ &= \exp\left[i\langle m_\mu, f\rangle - \frac{1}{2}\langle C_\mu f, f\rangle \right] = \hat{\mu}(f),\end{aligned}$$

where the second-to-last equality follow from Equation (3.17). This completes the proof in the infinite rank case. □

3.8 Appendix C: Explicit Update Formulae for Mean Element and Covariance Operator

For the sake of completeness, we here provide detailed update formulae for the mean element and covariance operator, as a direct consequence of Theorem 9.

Corollary 8. *Consider the setting of Theorem 9 and let $(y_i^{*(12)})_{i=1,\dots,p_{12}}$ be a $GC_\mu G^*$-representing sequence, $(y_i^{*(1)})_{i=1,\dots,p_1}$ be a $G_1 C_\mu G_1^*$-representing sequence and $(y_i^{*(2)})_{i=1,\dots,p_2}$ be a $G_2 C_\mu^{(1)} G_2^*$-representing sequence. Then we have:*

$$\begin{aligned}C_\mu - \sum_{i=1}^{p_{12}} \left\langle C_\mu G^* y_i^{(12)}, \bullet \right\rangle C_\mu G^* y_i^{*(12)} &= C_\mu - \sum_{i=1}^{p_1} \left\langle C_\mu G^* y_i^{*(1)}, \bullet \right\rangle C_\mu G_1^* y_i^{*(1)} - \sum_{j=1}^{p_2} \left\langle C_\mu G_2^* y_j^{*(2)}, \bullet \right\rangle C_\mu G_2^* y_j^{*(2)} \\ &\quad + \sum_{j=1}^{p_2}\sum_{i=1}^{p_1} \left\langle C_\mu G_2^* y_j^{*(2)}, \bullet \right\rangle \left\langle C_\mu G_1^* y_i^{*(1)}, G_2^* y_j^{*(2)} \right\rangle C_\mu G^* y_i^{*(1)} \\ &\quad - \sum_{j=1}^{p_2}\sum_{i=1}^{p_1}\sum_{k=1}^{p_1} \left\langle C_\mu G_1^* y_i^{*(1)}, G_2^* y_j^{*(2)} \right\rangle \left\langle C_\mu G_1^* y_i^{*(1)}, \bullet \right\rangle \left\langle C_\mu G_1^* y_i^{*(1)}, G_2^* y_j^{*(2)} \right\rangle C_\mu G_1^* y_i^{*(1)},\end{aligned}$$

and the equality is independent of the choice of the representing sequences.

As for the mean element, we have:

$$
\begin{aligned}
m_\mu + \sum_{i=1}^{q} \left\langle y - Gm_\mu, y_i^{*(12)} \right\rangle C_\mu G^* y_i^{*(12)} = m_\mu + \sum_{i=1}^{q_1} \left\langle y_1 - G_1 m_\mu, y_i^{*(1)} \right\rangle C_\mu G_1^* y_i^{*(1)} \\
+ \sum_{j=1}^{q_2} \left\langle y_2, y_j^{*(2)} \right\rangle C_\mu G_2^* y_j^{*(2)} - \sum_{j=1}^{q_2} \sum_{i=1}^{q_1} \left\langle y_2, y_j^{*(2)} \right\rangle \left\langle C_\mu G_1^* y_i^{*(1)}, G_2^* y_j^{*(2)} \right\rangle C_\mu G_1^* y_i^{*(1)} \\
- \sum_{j=1}^{q_2} \left\langle G_2 m_\mu, y_j^{*(2)} \right\rangle C_\mu G_2^* y_j^{*(2)} \\
- \sum_{j=1}^{q_2} \sum_{i=1}^{q_1} \left\langle G_2 C_\mu G_1^* y_i^{*(1)}, y_j^{*(2)} \right\rangle \left\langle y_1 - G_1 m_\mu, y_i^{*(1)} \right\rangle C_\mu G_2^* y_j^{*(2)} \\
+ \sum_{j=1}^{q_2} \sum_{k=1}^{q_1} \left\langle G_2 m_\mu, y_j^{*(2)} \right\rangle \left\langle C_\mu G_1^* y_k^{*(1)}, G_2^* y_j^{*(2)} \right\rangle C_\mu G_1^* y_k^{*(1)} \\
+ \sum_{i=1}^{q_2} \sum_{i=j}^{q_1} \sum_{k=1}^{q_1} \left\langle G_2 C_\mu G_1^* y_i^{*(1)}, y_j^{*(2)} \right\rangle \left\langle y_1 - G_1 m_\mu, y_i^{*(1)} \right\rangle \left\langle C_\mu G_1^* y_k^{*(1)}, G_2^* y_j^{*(2)} \right\rangle C_\mu G_1^* y_k^{*(1)}.
\end{aligned}
$$

Chapter 4

Implicit Covariance Representation for Fast Update in Large-scale Bayesian Inversion

This chapter reproduces the paper Travelletti et al. (2023), co-authored with David Ginsbourger and Niklas Linde and published in the SIAM Journal of Uncertainty Quantification
(DOI:10.48550/ARXIV.2109.03457).

4.1 Introduction

In this chapter, we use the theoretical background on sequential disintegrations of Gaussian measures developed in Chapter 3 to devise a new framework for sequentially updating Gaussian processes that allow them to be brought to bear on large-scale linear Bayesian inverse problems.

We focus on a subcategory of Bayesian inverse problems that are situated at a triple intersection, namely, we consider situations in which As stated above, we focus on the case where: (1) the number of prediction points is large, (2) the data has to be assimilated sequentially, and (3) it comes in the form of integral operators observations. Integral operators are harder to handle than pointwise observations since, when discretized on a grid (which is the usual inversion approach), they turn into a matrix with entries that are non-zero for most grid points, preventing the use of techniques that leverage sparse matrices. This situation is typical of Bayesian large-scale inverse problems because those are often solved on a discrete grid, forcing one to consider a large number of prediction points when inverting at high resolution; besides, the linear operators found in inverse problems are often of integral form (e.g. gravity, magnetics).

Our main contribution to overcome the above difficulties is the introduction of an implicit representation of the posterior covariance matrix that only requires storage of low rank intermediate matrices and allows individual elements to be accessed on-the-fly, without ever storing the full matrix. Our method relies on an extension of the *kriging update formulae* (Chevalier et al., 2014b; Emery, 2009; Gao et al., 1996; Barnes and Watson, 1992) to linear operator observations. As a minor contribution, we also provide a technique for computing posterior means on fine discretizations using a chunking technique and explain how to perform posterior simulations in the

considered setting. The developed implicit representation allows for fast updates of posterior covariances under linear operator observations on very large grids. This is particularly useful when computing sequential data acquisition plans for inverse problems, which we demonstrate by computing sequential experimental designs for excursion set learning in gravimetric inversion. We find that our method provides significant computational time savings over brute-force conditioning and scales to problem sizes that are too large to handle using state-of-the-art techniques.

This whole chapter will use the Stromboli gravimetric inversion example from Section 2.4.1 as a red thread for the development of new techniques and as an applicative testbed for these. To the best of our knowledge, this is the first time that sequential experimental design for set estimation is considered in such a setting.

4.2 Background: Sequential Bayesian Data Assimilation and Related Challenges

In this section, we focus on the challenges that arise when using Gaussian process priors to solve Bayesian inverse problems with linear operators observations. Most of these problems trace their roots back to the discretization of the problem. Indeed, even tough one could formulate everything in an infinite dimensional setting (see Chapter 3) and should discretize as late as possible (Stuart, 2010), in practice there is always some form of discretization involved, be it through quadrature methods (Hansen, 2010) or through basis expansion (Wagner et al., 2021). It turns out that, regardless of the type of discretization used, one quickly encounters computational bottlenecks arising from memory limitations when trying to scale inversion to real-world problems. We here focus on inverse problems discretized on a grid, but stress that the computational difficulties described next also plague spectral approaches. For the rest of this section, we consider the same setting as in Section 2.2.2.

Let $\boldsymbol{W} = (w_1, ..., w_r) \in D^r$ be a given set of discretization points, we consider observation operators that (after discretization) may be written as linear combinations of Dirac delta functionals

$$G : C(D) \to \mathbb{R}^q,\ G = \left(\sum_{j=1}^{r} g_{ij} \delta_{w_j} \right)_i,\ i = 1, ..., q, \tag{4.1}$$

with arbitrary coefficients $g_{ij} \in \mathbb{R}$. When working with such discrete operators it is more convenient to use matrices, we will thus use $\bar{G}$ to denote the $q \times r$ matrix with elements g_{ij}. Then, assuming we have a GP prior $Z \sim \mathrm{Gp}(m, k)$ on D and data of the form:

$$\boldsymbol{Y} = \bar{G} Z_{\boldsymbol{W}} + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Delta)$ is some observational noise, we can compute the conditional law of the process, conditionally on the data using Corollary 5. Given a batch of prediction points $\boldsymbol{X} = (x_1, ..., x_m)$, the conditional mean and covariance, conditional

on $\boldsymbol{Y} = \boldsymbol{y}$ are then given by:

$$\tilde{m}_{\boldsymbol{X}} = m_{\boldsymbol{X}} + K_{\boldsymbol{XW}}\bar{G}^T\left(\bar{G}K_{\boldsymbol{WW}}\bar{G}^T + \Delta\right)^{-1}\left(\boldsymbol{y} - \bar{G}m_{\boldsymbol{W}}\right), \tag{4.3}$$

$$\tilde{K}_{\boldsymbol{XX'}} = K_{\boldsymbol{XX'}} - K_{\boldsymbol{XW}}\bar{G}^T\left(\bar{G}K_{\boldsymbol{WW}}\bar{G}^T + \Delta\right)^{-1}K_{\boldsymbol{WX'}}. \tag{4.4}$$

Even if Eqs. (4.3) and (4.4) only involve basic matrix operations, their computational cost depends heavily on the number of prediction points $\boldsymbol{X} = (x_1, ..., x_m)$ and on the number of discretization points $\boldsymbol{W} = (w_1, ..., w_r)$, making their application to real-world inverse problems a non-trivial task. Indeed, when both m and r are big, there are two main difficulties that hamper the computation of the conditional distribution:

- the $r \times r$ matrix $K_{\boldsymbol{WW}}$ may never be built in memory due to its size, and
- the $m \times m$ posterior covariance $K_{\boldsymbol{XX}}$ may be too large to store.

While the first of these difficulties can be solved by performing the product $K_{\boldsymbol{WW}}G^T$ in chunks, as described in Section 4.3, the second one only becomes of particular interest in sequential settings. Indeed, in practice, data often becomes available in stages and one is interested in updating the posterior from stages to stages. In such a setting, a set of observations described by a (discretized) operator G_i is made at each stage (from now on we only consider observation operators in matrix form and drop underbars). One then observes a realization $\boldsymbol{y}_i$ of

$$\boldsymbol{Y}_i = G_i Z_{\boldsymbol{W}_i} + \boldsymbol{\epsilon}_i, \tag{4.5}$$

where $\boldsymbol{W}_i$ is some set of points in D and $\boldsymbol{\epsilon}_i$ is some centered Gaussian distributed noise with covariance matrix Δ_i. Then, in order to avoid a full recomputation of the posterior, one can use Corollary 6 to obtain the posterior mean and covariance after each stage by performing a low rank update of their counterparts at the previous stage:

Theorem 13. *Let $Z \sim \mathrm{Gp}(m, k)$ and let $m^{(n)}$ and $K^{(n)}$ denote the conditional mean and covariance function conditional on the data $\{\boldsymbol{Y}_i = \boldsymbol{y}_i : i = 1, ..., n\}$ with $\boldsymbol{Y}_i$ defined as in Eq. (4.5), where $n \geq 1$ and $m^{(0)}$ and $K^{(0)}$ are used to denote the prior mean and covariance. Then:*

$$m^{(n)}_{\boldsymbol{X}} = m^{(n-1)}_{\boldsymbol{X}} + \lambda_n(\boldsymbol{X})^T\left(\boldsymbol{y}_n - G_n\, m^{(n-1)}_{\boldsymbol{W}_n}\right),$$
$$K^{(n)}_{\boldsymbol{XX'}} = K^{(n-1)}_{\boldsymbol{XX'}} - \lambda_n(\boldsymbol{X})^T S_n \lambda_n(\boldsymbol{X'}),$$

with $\lambda_n(\boldsymbol{X})$, S_n defined as:

$$\lambda_n(\boldsymbol{X}) = S_n^{-1} G_n K^{(n-1)}_{\boldsymbol{W}_n\boldsymbol{X}},$$
$$S_n = G_n K^{(n-1)}_{\boldsymbol{W}_n\boldsymbol{W}_n} G_n^T + \Delta_n.$$

This is in essence an extension of Chevalier et al. (2014b); Emery (2009); Gao et al. (1996); Barnes and Watson (1992) to linear operator observations. At each stage n, these formulae require computation of the $q_n \times m$ matrix $\lambda_n(\boldsymbol{X})$, which involves a $q_n \times q_n$ matrix inversion, where q_n is the dimension of the operator G_n describing the current dataset to be included. This allows computational savings

by reusing already computed quantities, avoiding inverting the full dataset at each stage, which would require a q_{tot}^2 matrix inversion, where $q_{tot} = \sum_{i=1}^{n} q_i$.

In order for these update equations to bring computational savings, one has to be able to store the past covariances $K^{(n-1)}_{\boldsymbol{W}_n \boldsymbol{W}_n}$ (Chevalier et al., 2015). This makes their application to large-scale sequential Bayesian inverse problems difficult, since the covariance matrix on the full discretization may become too large for storage above a certain number of discretization points. The next section presents our main contributions to overcome this limitation. They rely on an implicit representation of the posterior covariance that allows the computational savings offered by the kriging update formulae to be brought to bear on large scale inverse problems.

4.3 Implicit Covariance Representation and Update

We consider the same sequential data assimilation setup as in the previous section, and for the sake of simplicity we assume that $\boldsymbol{W}_1, ..., \boldsymbol{W}_n = \boldsymbol{X}$ and use the lighter notation $m^{(i)} := m^{(i)}_{\boldsymbol{X}}$ and $K^{(i)} := K^{(i)}_{\boldsymbol{X}\boldsymbol{X}}$. The setting we are interested in here is the one where $\boldsymbol{X}$ is so large that the covariance matrix gets bigger than the available computer memory.

Our key insight is that instead of building the full posterior covariance $K^{(n)}$ at each stage n, one can just maintain a routine that computes the product of the current posterior covariance with any other low rank matrix. More precisely, at each stage n, we provide a routine CovMul_n (Algorithm 1), that allows to compute the product of the current covariance matrix with any *thin* matrix $A \in \mathbb{R}^{m \times a}$, $a \ll m$:

$$\mathit{CovMul}_n : A \mapsto K^{(i)} A,$$

where *thin* is to be understood as small enough so that the result of the multiplication can fit in memory.

This representation of the posterior covariance was inspired by the covariance operator of Gaussian measures. Indeed, if we denote by $C_{\mu^{(n)}}$ the covariance operator of the Gaussian measure associated to the posterior distribution of the GP at stage n, then

$$\left(K^{(n)} A\right)_{ij} = \sum_{k=1}^{m} \left\langle C_{\mu^{(n)}} \delta_{x_i}, \delta_{x_k} \right\rangle A_{kj}.$$

Hence, the procedure CovMul_n may be thought of as computing the action of the covariance operator of the Gaussian measure associated to the posterior on the Dirac delta functionals at the discretization points.

This motivates us to think in terms of an *updatable covariance* object, where the inclusion of new observations (the updating) amounts to redefining a right-multiplication routine. It turns out that by grouping terms appropriately in Theorem 13 such a routine may be defined by only storing low rank matrices at each data acquisition stage.

Lemma 4. *For any $n \in \mathbb{N}$ and any $m \times a$ matrix A:*

$$K^{(n)}A = K^{(0)}A - \sum_{i=1}^{n} \bar{K}_i R_i^{-1} \bar{K}_i^T A,$$

with intermediate matrices $\bar{K}_i$ and R_i^{-1} defined as:

$$\bar{K}_i := K^{(i-1)} G_i^T,$$
$$R_i^{-1} := \left(G_i K^{(i-1)} G_i^T + \Delta_i \right)^{-1}.$$

Hence, in order to compute products with the posterior covariance at stage n, one only has to store n matrices $\bar{K}_i$, each of size $m \times q_i$ and n matrices R_i^{-1} of size $q_i \times q_i$, where q_i is the number of observations made at stage i (i.e. the number of lines in G_i). In turn, each of these objects is defined by multiplications with the covariance matrix at previous stages, so that one may recursively update the multiplication procedure CovMul_n. Algorithms 1, 2, and 3 may be used for multiplication with the current covariance matrix, update of the representation and update of the posterior mean.

Algorithm 1 *Covariance Right Multiplication Procedure $CovMul_n$*

Require:
Precomputed matrices $\bar{K}_i,\ R_i^{-1}$, $i = 1, \ldots, n$.
Prior multiplication routine CovMul_0.
Input matrix A.
Ensure: $K^{(n)}A$.
procedure CovMul_n(A)
Compute $K^{(0)}A = \text{CovMul}_0(A)$.
Return $K^{(0)}A - \sum_{i=1}^{n} \bar{K}_i R_i^{-1} \bar{K}_i^T A$.

Algorithm 2 *Updating intermediate quantities at conditioning stage n*

Require:
Last multiplication routine CovMul_{n-1}.
Measurement matrix G_n, noise variance τ^2.
Ensure: Step n intermediate matrices $\bar{K}_n$ and R_n
procedure Update_n
Compute $\bar{K}_n = \text{CovMul}_{n-1} G_n^T$.
Compute $R_n^{-1} = \left(G_n \bar{K}_n + \Delta_n \right)^{-1}$.

Prior Covariance Multiplication Routine and Chunking: To use Algorithm 1, one should be able to compute products with the prior covariance matrix $K^{(0)}$. To achieve this, we use the same technique as in (Wang et al., 2019), performing products in chunks. We note that our implicit representation framework is able to handle updates, whereas (Wang et al., 2019) only consider a single step of assimilation.
We start by chunking the set of grid points into n_c subsets $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_{n_c})$, where each $\boldsymbol{X}_i$ contains a subset of the points. Without loss of generality, we assume all subsets to have the same size m_c. We may then write the product as

$$K^{(0)}_{\boldsymbol{X}\boldsymbol{X}} A = \left(K^{(0)}_{\boldsymbol{X}_1 \boldsymbol{X}} A, \ldots, K^{(0)}_{\boldsymbol{X}_{n_c} \boldsymbol{X}} A \right)^T.$$

Algorithm 3 *Computation of conditional mean at step n*

Require:
Previous conditional mean $m^{(n-1)}$.
Current data $\boldsymbol{y}_n$ and forward G_n.
Intermediate matrices $\bar{K}_n$ and R_n^{-1}.
Ensure: Step n conditional mean $m^{(n)}$.
procedure MeanUpdate$_n$
Return $m^{(n-1)} + \bar{K}_n R_n^{-1} \left(\boldsymbol{y}_n - G_n m^{(n-1)}\right)$.

Each of the subproducts may then be performed separately and the results gathered together at the end. The individual products then involve matrices of size $m_c \times m$ and $m \times a$. One can then choose the number of chunks so that these matrices can fit in memory. Each block $K^{(0)}_{\boldsymbol{X}_i \boldsymbol{X}}$ may be built on-demand provided $K^{(0)}_{\boldsymbol{X}\boldsymbol{X}}$ is defined through a given covariance function.
This ability of the prior covariance to be built quickly on-demand is key to our method. The fact that the prior covariance matrix does not need to be stored allows us to handle larger-than-memory posterior covariances by expressing products with it as a multiplication with the prior and a sum of multiplications with lower rank matrices.

Remark 5 (Choice of Chunk Size)**.** Thanks to chunking, the product may be computed in parallel, allowing for significant performance improvements in the presence of multiple computing devices (CPUs, GPUs, ...). In that case, the chunk size should be chosen as large as possible to limit data transfers, but small enough so that the subproducts may fit on the devices.

Computational Cost: For the sake of comparison, assume that all n datasets have the same size q_c and let $q = nq_c$ denote the total data size. The cost of computing products with the current posterior covariance matrix at some intermediate stage is given by:

Lemma 5 (Multiplication Cost)**.** *Let A be an $m \times a$ matrix. Then, the cost of computing $K_n A$ at some stage n using Algorithms 1 and 2 is $\mathcal{O}\left(m^2 a + n(mq_c a + q_c^2 a)\right)$.*

Using this recursively, we can then compute the cost of creating the implicit representation of the posterior covariance matrix at stage n:

Lemma 6 (Implicit Representation Cost)**.** *To leading order in m and q, the cost of defining CovMul$_n$ is $\mathcal{O}\left(m^2 q + mq^2 + q^2 q_c\right)$. This is also the cost of computing $m^{(n)}$.*

This can then be compared with a non-sequential approach where all datasets would be concatenated into a single dataset of dimension q. More precisely, define the $q \times m$ matrix $\underline{\mathrm{G}}$ and the q-dimensional vector $\boldsymbol{y}$ as the concatenations of all the measurements and data vectors into a single operator, respectively vector. Then computing the posterior mean using Eq. (4.3) with those new observation operators and data vector the cost is, to leading order in q and m:

$$\mathcal{O}\left(m^2 q + mq^2 + q^3\right).$$

In this light, we can now sum up the two main advantages of the proposed sequential approach:

- the cubic cost $\mathcal{O}(q^3)$ arising from the inversion of the data covariances is decreased to $\mathcal{O}(q^2 q_c)$ in the sequential approach
- if a new set of observations has to be included, then the direct approach will require the $\mathcal{O}(m^2 q)$ computation of the product $K\underline{\mathrm{G}}^T$, which can become prohibitively expensive when the number of prediction points is large, whereas the sequential approach will only require a marginal computation of $\mathcal{O}(m^2 q_c)$.

Aside from the computational cost, our implicit representation also provides significant memory savings compared to an *explicit* approach where the full posterior covariance matrix would be stored. The storage requirement for the implicit-representation as a function of the number of discretization points m is shown in Figure 4.1.

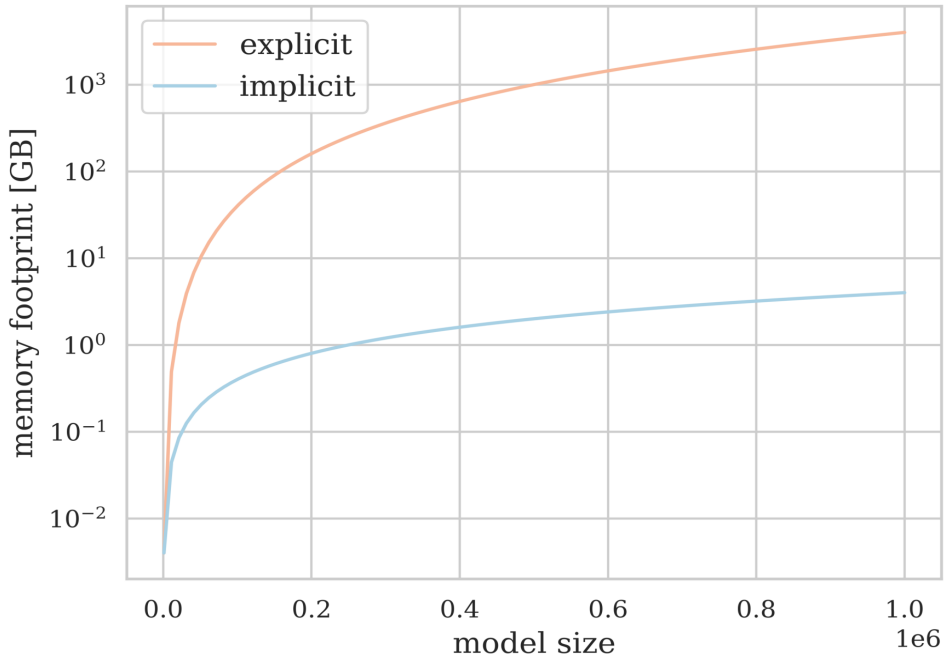


Figure 4.1: Memory footprint of the posterior covariance matrix as a function of discretization size for explicit and implicit representation.

4.4 Application: Scaling Gaussian Processes to Large-Scale Inverse Problems

In this section, we demonstrate how the implicit representation of the posterior covariance introduced in Section 4.3 allows scaling Gaussian processes to situations that are too large to handle using more traditional techniques, such as frequently arises in large-scale inverse problems. We will focus our exposition on the gravimetric inverse problem example presented in Section 2.4.1, demonstrating how our implicit representation allows training prior hyperparameters, sample from the posterior on large grids and finally how it allows us to address a state-of-the-art sequential experimental design problem for excursion set recovery.

4.4.1 Hyperparameter Optimization

When using Gaussian process priors to solve inverse problems, one has to select the hyperparameters of the prior. There exists different approaches for optimizing hyperparameters. We here only consider maximum likelihood estimation (MLE).

We restrict ourselves to GP priors that have a constant prior mean $m_0 \in \mathbb{R}$ and a covariance kernel k that depends on a prior variance parameter σ_0^2 and other correlation parameters $\boldsymbol{\theta}_0 \in \mathbb{R}^t$:

$$k(x, x') = \sigma_0^2 r(x, x'; \boldsymbol{\theta}_0), \tag{4.6}$$

where $r(., .; \boldsymbol{\theta}_0)$ is a correlation function, such that $r(x, x; \boldsymbol{\theta}_0) = 1, \forall x \in D$. The maximum likelihood estimator for the hyperparameters may then be obtained by minimizing the negative marginal log likelihood (nmll) of the data, which in the discretized setting of Section 4.2 may be written as (Rasmussen and Williams, 2006):

$$\begin{aligned} \mathcal{L}(m_0, \sigma_0, \boldsymbol{\theta}_0; \boldsymbol{y}) &= \frac{1}{2} \log\det R + \frac{1}{2}\Big(\boldsymbol{y} - G m_{\boldsymbol{X}}\Big)^T R^{-1}\Big(\boldsymbol{y} - G m_{\boldsymbol{X}}\Big) + \frac{n}{2}\log 2\pi, \\ R &:= \Big(G K_{\boldsymbol{X}\boldsymbol{X}} G^T + \Delta\Big). \end{aligned} \tag{4.7}$$

Since only the quadratic term depends on m_0, we can adapt concentration identities (Park and Baek, 2001) to write the optimal m_0 as a function of the other hyperparameters:

$$\hat{m}_0^{MLE}(\sigma_0, \boldsymbol{\theta}_0) = \Big(\mathbf{1}_m^T G^T R^{-1} G \mathbf{1}_m\Big)^{-1} y^T R^{-1} G \mathbf{1}_m, \tag{4.8}$$

where $\mathbf{1}_m$ denotes the m-dimensional column vector containing only 1's. Here we always assume R to be invertible. The remaining task is then to minimize the concentrated nmll:

$$(\sigma_0, \boldsymbol{\theta}_0) \mapsto \mathcal{L}\left(\hat{m}_0^{MLE}(\sigma_0, \boldsymbol{\theta}_0), \sigma_0, \boldsymbol{\theta}_0\right).$$

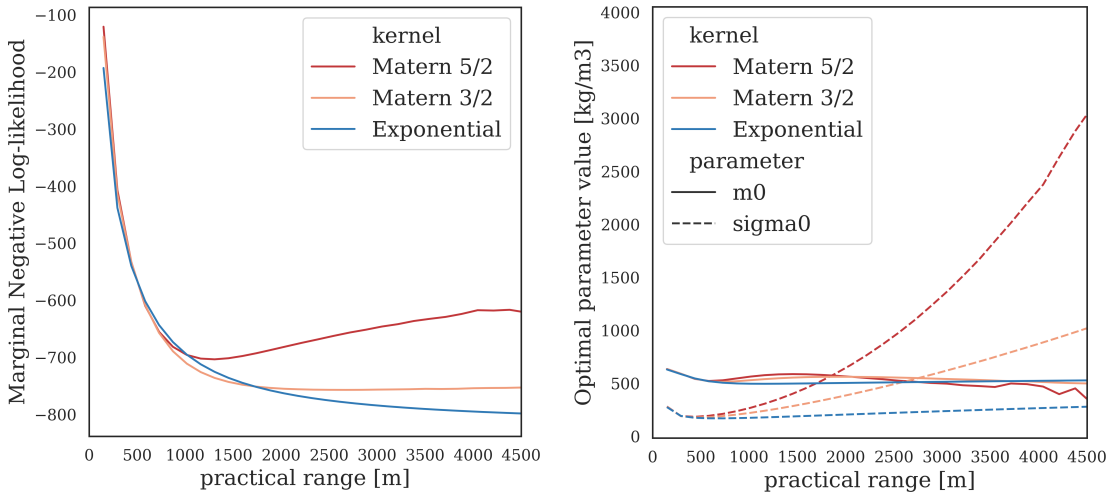
Note that the main computational challenge in the minimization of Eq. (4.7) comes from the presence of the $m \times m$ matrix $K_{\boldsymbol{X}\boldsymbol{X}}$. In the following, we will only consider the case of kernels that depend on a single length scale parameter: $\boldsymbol{\theta}_0 = \lambda_0 \in \mathbb{R}$, though the procedure described below can in principle be adapted for multidimensional $\boldsymbol{\theta}_0$.

In practice, for kernels of the form Eq. (4.6) the prior variance σ_0^2 may be factored out of the covariance matrix (for known noise variance), so that only the prior length scale λ_0 appears in this large matrix. One then optimizes these parameters separately, using chunking to compute matrix products. Since σ_0 only appears in an $q \times q$ matrix which does not need to be chunked (the data size q being moderate in real applications), one can use automatic differentiation libraries such as Paszke et al. (2019) to optimize it by gradient descent. On the other hand, there is no way to factor out λ_0 out of the large matrix $K_{\boldsymbol{X}\boldsymbol{X}}$, so we resort to a brute force approach by specifying a finite search space for it. To summarize, we proceed here in the following way:

(i) (brute force search) Discretize the search space for the length scale by only allowing $\lambda_0 \in \Lambda_0$, where Λ_0 is a discrete set (usually equally spaced values on a reasonable search interval);

(ii) (gradient descent) For each possible value of λ_0, minimize the (concentrated) $\mathcal{L}$ over the remaining free parameter σ_0 by gradient descent.

We ran the above approach on the Stromboli dataset with standard stationary kernels (Matérn 3/2, Matérn 5/2, exponential). In agreement with Linde et al. (2014), the observational noise is i.i.d. Gaussian distributed with standard deviation is 0.1 $[mGal]$. The optimization results for different values of the length scale parameter are shown in Fig. 4.2. The best estimates of the parameter values for each kernel are shown in Table 4.1. The table also shows the practical range $\bar{\lambda}$ which is defined as the distance at which the covariance falls to 5% of its original value.



(a) (b)

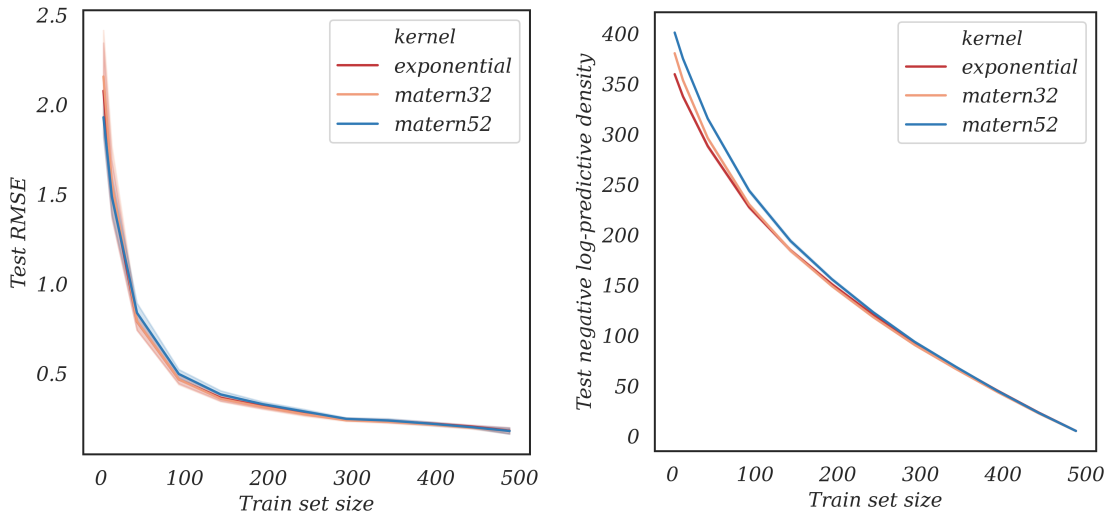
Figure 4.2: (a) Concentrated negative marginal log-likelihood and (b) optimal hyperparameter values for different length scale parameters λ_0.

We assess the robustness of each kernel by predicting a set of left out observations using the other remaining observations. Fig. 4.3 displays RMSE and negative log predictive density for different proportion of train/test splits.

	Hyperparameters				Metrics	
Kernel	λ	$\bar{\lambda}$	m_0	σ_0	$\mathcal{L}$	Train RMSE
Exponential	1925.0	5766.8	535.4	308.9	-804.4	0.060
Matérn 3/2	**651.6**	**1952.0**	**2139.1**	**284.65**	**-1283.5**	**0.071**
Matérn 5/2	441.1	1321..3	2120.9	349.5	-1247.6	0.073

Table 4.1: Optimal hyperparameters (Stromboli dataset) for different kernels.

Note thatthe above procedure is more of a quality assurance than a rigorous statistical evaluation of the model, since all datapoints were already used in the fitting of the hyperparameters. Due to known pathologies of the exponential kernel (MLE



(a) (b)

Figure 4.3: (a) Root mean squared error and (b) negative log predictive density on test set for the different models (with optimal hyperparameters). The full dataset contains 501 observations.

for length scale parameter going to infinity), we choose to use the Matérn 3/2 model for the experiments of Section 4.4.2 and Section 4.4.3. The maximum likelihood estimator of the prior hyperparameters for this model are $\hat{m}_0^{MLE} = 2139.1\ [kg/m^3]$, $\hat{\sigma}_0^{MLE} = 284.65\ [kg/m^3]$ and $\hat{\lambda}_0^{MLE} = 651.6\ [m]$.

4.4.2 Posterior Sampling

Our implicit representation also allows for efficient sampling from the posterior by using the *residual kriging* algorithm (Chilès and Delfiner, 2012; de Fouquet, 1994), which we here adapt to linear operator observations. Note that in order to sample a Gaussian process at m sampling points, one needs to generate m correlated Gaussian random variables, which involves covariance matrices of size m^2, leading to the same computational bottlenecks as described in Section 4.2. On the other hand, the residual kriging algorithm generates realizations from the posterior by updating realizations of the prior, as we explain next.

As before, suppose we have a GP Z defined on some compact Euclidean domain D and assume Z has continuous sample paths almost surely. Furthermore, say we have q observations described by linear operators $\ell_1, ..., \ell_q \in C(D)^*$. Then the conditional expectation of Z conditional on the σ-algebra $\Sigma := \sigma\left(\ell_1\left(Z\right), ..., \ell_q\left(Z\right)\right)$ is an orthogonal projection (in the L^2-sense (Williams, 1991)) of Z onto Σ. This orthogonality can be used to decompose the conditional law of Z conditional on Σ into a conditional mean plus a residual. Indeed, if we let Z' be another GP with the same distribution as Z and let $\Sigma' := \sigma\left(\ell_1\left(Z'\right), ..., \ell_q\left(Z'\right)\right)$, then we have the following equality in distribution:

$$Z_x \mid \Sigma = \mathbb{E}\left[Z_x \mid \Sigma\right] + \left(Z'_x - \mathbb{E}\left[Z'_x \mid \Sigma'\right]\right), \text{ all } x \in D. \tag{4.9}$$

Compared to direct sampling of the posterior, the above approach involves two main

operations: sampling from the prior and conditioning under operator data. When the covariance kernel is stationary and belongs to one of the usual families (Gaussian, Matérn), methods exist to sample from the prior on large grids (Mantoglou and Wilson, 1982); whereas the conditioning part may be performed using our implicit representation.

Remark 6. Note that in a sequential setting as in Section 4.2, the residual kriging algorithm may be used to maintain an ensemble of realizations from the posterior distribution by updating a fixed set of prior realizations at every step in the spirit of Chevalier et al. (2015).

4.4.3 Sequential Experimental Design for Excursion Set Recovery

As a last example of application where our implicit update method provides substantial savings, we consider a sequential data collection task involving an inverse problem. Though sequential design criteria for inverse problems have already been considered in the literature (Attia et al., 2018), most of them only focus on selecting observations to improve the reconstruction of the unknown parameter field, or some linear functional thereof.

We here consider a different setting. In light of recent progress in excursion set estimation (Azzimonti et al., 2016; Chevalier et al., 2013), we instead focus on the task of recovering an excursion set of the unknown parameter field ρ, that is, we want to learn the unknown set $\Gamma^* := \{x \in D : \rho(x) \geq T\}$, where T is some threshold. In the present context of Stromboli, high density areas are related to dykes (previous feeding conduits of the volcano), while low density values are related to deposits formed by paroxysmal explosive phreato-magmatic events (Linde et al., 2014). To the best of our knowledge, such sequential experimental design problems for excursion set learning in inverse problems have not been considered elsewhere in the literature.

Remark 7. For the sake of simplicity, we focus only on excursion sets above some threshold, but all the techniques presented here may be readily generalized to generalized excursion sets of the form $\Gamma^* := \{x \in D : \rho(x) \in I\}$ where I is any finite union of intervals on the extended real line.

We here consider a sequential setting, where observations are made one at a time and at each stage we have to select which observation to make next in order to optimally reduce the uncertainty on our estimate of Γ^*. Building upon Picheny et al. (2010); Bect et al. (2012); Azzimonti et al. (2021); Chevalier et al. (2014a), there exists several families of criteria to select the next observations. Here, we restrict ourselves to a variant of the weighted IMSE criterion (Picheny et al., 2010). The investigation of other state-of-the-art criteria is left for future work. We note in passing that most Bayesian sequential design criteria involve posterior covariances and hence tend to become intractable for large-scale problems. Moreover in a sequential setting, fast updates of the posterior covariance are crucial. Those characteristics make the problem considered here particularly suited for the implicit update framework introduced in Section 4.3.

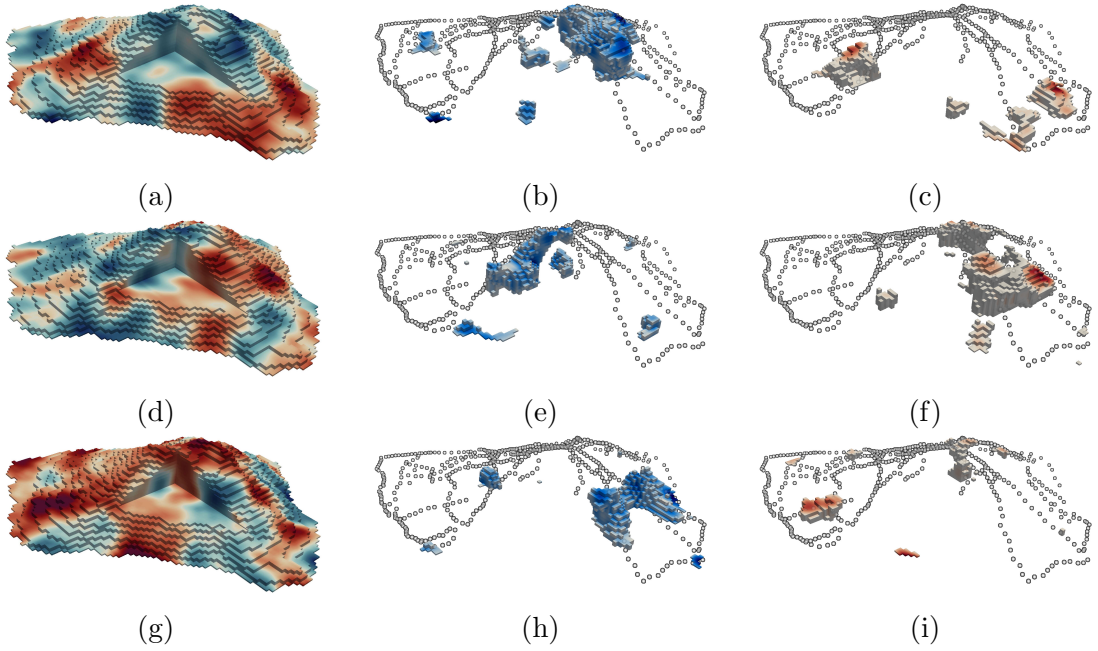


Figure 4.4: Realizations from Matérn 3/2 GP prior (hyperparameters taken from Table 4.1) with corresponding excursion sets: (left to right) Underground mass density field (arbitrary color scale), high density regions and low density regions, thresholds: 2600 $[kg/m^3]$ and 1700 $[kg/m^3]$.

The weighted IMSE criterion selects next observations by maximizing the variance reduction they will provide at each location, weighted by the probability for that location to belong to the excursion set Γ^*. Assuming that n data collection stages have already been performed and using the notation of Section 4.2, the variant that we are considering here selects the next observation location by maximizing the *weighted integrated variance reduction* (wIVR):

$$\text{wIVR}^n(u) = \int_D \left(K_{xx}^{(n)} - K_{xx}^{(n+1)}\left[G_u\right] \right) p_n(x) dx, \tag{4.10}$$

where u is some potential observation location, $K^{(n+1)}$ denotes the conditional covariance after including a gravimetric observation made at u (this quantity is independent of the observed data) and G_u is the forward operator (matrix) corresponding to this observation. Also, here p_n denotes the *coverage function* at stage n (we refer the reader to Section 2.3 for more details on Bayesian set estimation). After discretization, applying Theorem 13 turns this criterion into:

$$\sum_{x \in \boldsymbol{X}} K_{x\boldsymbol{X}}^{(n)} G_u^T \left(G_u K_{\boldsymbol{X}\boldsymbol{X}}^{(n)} G_u^T + \Delta \right)^{-1} G_u K_{\boldsymbol{X}x}^{(n)} p_n(x), \tag{4.11}$$

where we have assumed that all measurements are affected by $\mathcal{N}(0, \Delta)$ distributed noise.

Note that for large-scale problems, the wIVR criterion in the form given in Eq. (4.11) becomes intractable for traditional methods because of the presence of the full posterior covariance matrix $K_{\boldsymbol{X}\boldsymbol{X}}^{(n)}$ in the parenthesis. The implicit representation presented in Section 4.3 can be used to overcome this difficulty. Indeed,

the criterion can be evaluated using the posterior covariance multiplication routine Lemma 4 (where the small dimension q is now equal to the number of candidate observations considered at a time, here 1 but batch acquisition scenarios could also be tackled). New observations can be seamlessly integrated along the way by updating the representation using Algorithm 2.

Experiments and Results: We now study how the wIVR criterion can help to reduce the uncertainty on excursion sets within the Stromboli volcano. We here focus on recovering the volume of the excursion set instead of its precise location. To the best of our knowledge, in the existing literature such sequential design criteria for excursion set recovery have only been applied to small-scale inverse problems and have not been scaled to larger, more realistic problems where the dimensions at play prevent direct access to the posterior covariance matrix.

In the following experiments, we use the Stromboli volcano inverse problem and work with a discretization into cubic cells of 50 $[m]$ side length. We use a Matérn 3/2 GP prior with hyperparameters trained on real data (Table 4.1) to generate semi-realistic ground truths for the experiments. We then simulate numerically the data collection process by computing the response that results from the considered ground truth and adding random observational noise. When computing sequential designs for excursion set estimation, the threshold that defines the excursion set can have a large impact on the accuracy of the estimate. Indeed, different thresholds will produce excursion sets of different sizes, which may be easier or harder to estimate depending on the set estimator used. For the present problem, Fig. 4.5 shows the distribution of the excursion volume under the considered prior for different excursion thresholds.

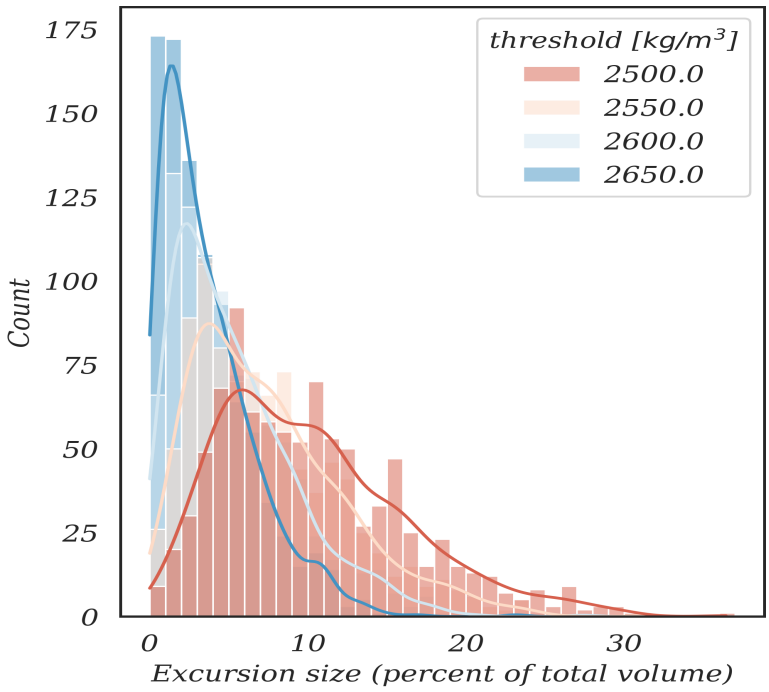
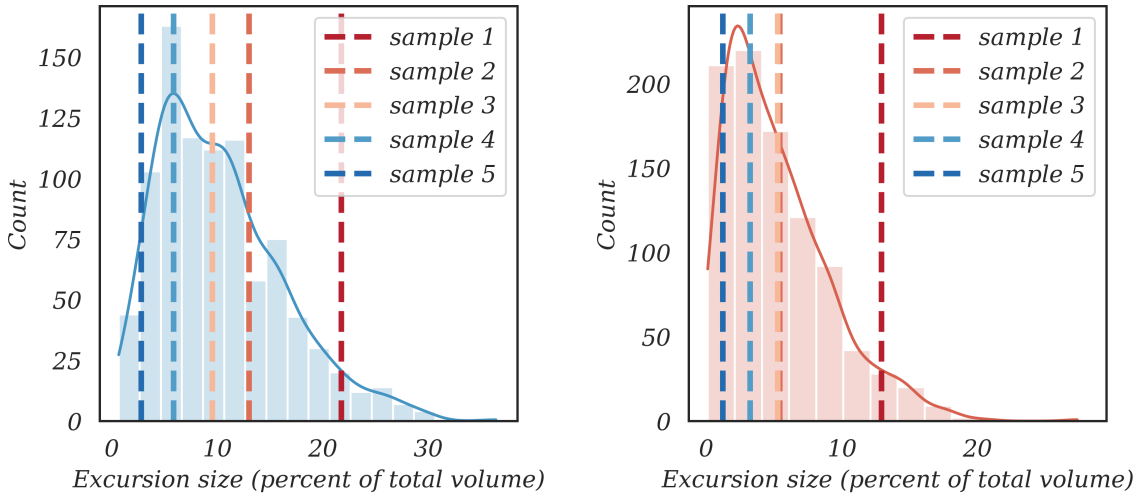


Figure 4.5: Distribution of excursion set volume under the prior for different thresholds. Size is expressed as a percentage of the volume of the inversion domain.

It turns out that the estimator used in our experiments (Vorob'ev expectation) behaves differently depending on the size of the excursion set to estimate. Indeed, the Vorob'ev expectation tends to produce a smoothed version of the true excursion set, which in our situation results in a higher fraction of false positives for larger sets. Thus, we consider two scenarios: a *large* scenario where the generated excursion sets have a mean size of 10% of the total inversion volume and a *small* scenario where the excursion sets have a mean size of 5% of the total inversion volume. One should note that those percentages are in broad accordance with the usual size of excursion sets that are of interest in geology. The chosen thresholds are 2500 $[kg/m^3]$ for the *large* excursions and 2600 $[kg/m^3]$ for the *small* ones.



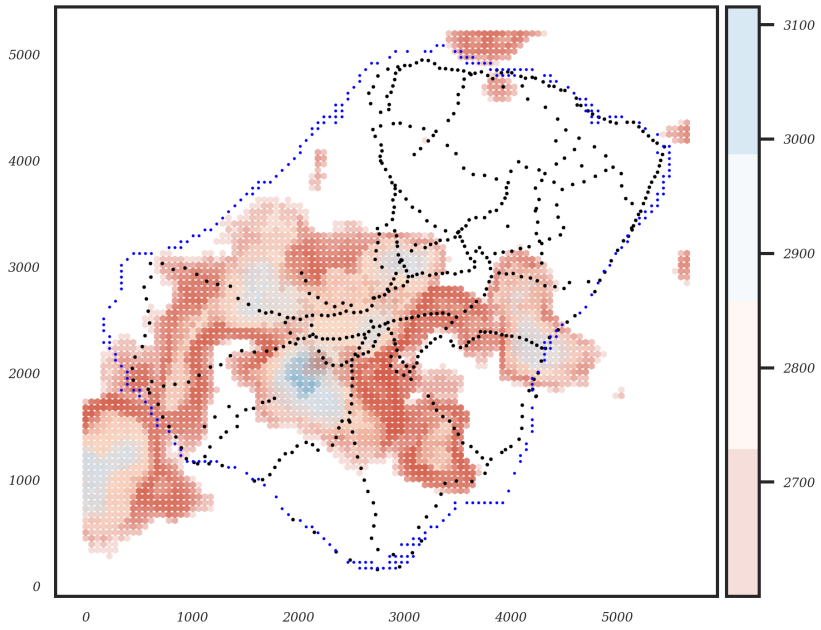
(a) (*large* scenario) threshold: 2500 $[kg/m^3]$

(b) (*small* scenario) threshold: 2600$[kg/m^3]$

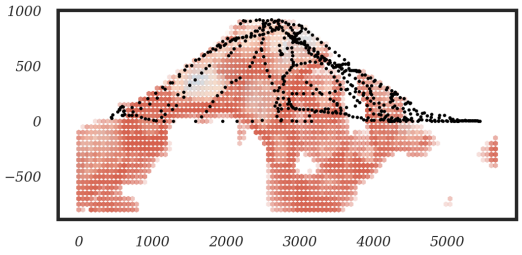
Figure 4.6: Distribution of excursion volume (with kernel density estimate) under the prior for the two considered thresholds, together with excursion volumes for each ground truth.

The experiments are run on five different ground truths, which are samples from a Matérn 3/2 GP prior (see previous paragraphs). The samples were selected such that their excursion set for the *large* scenario have volumes that correspond to the 5%, 27.5%, 50%, 72.5% and 95% quantiles of the prior excursion volume distribution for the corresponding threshold. Fig. 4.6 shows the prior excursion volume distribution together with the volumes of the five different samples used for the experiments. Fig. 4.7 shows a profile of the excursion set (small scenario) for one of the five samples used in the experiments. The data collection location from the 2012 field campaign (Linde et al., 2014) are denoted by black dots. The island boundary is denoted by blue dots. Note that, for the sake of realism, in the experiments we only allow data collection at locations that are situated on the island (data acquired on a boat would have larger errors); meaning that parts of the excursion set that are outside the island will be harder to recover.

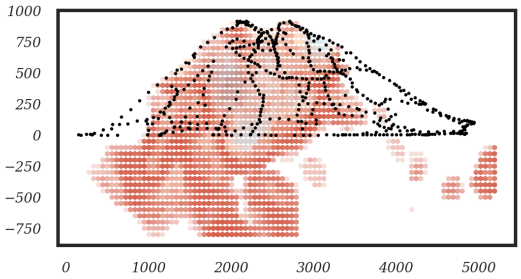
Experiments are run by starting at a fixed starting point on the volcano surface, and then sequentially choosing the next observation locations on the volcano surface according to the wIVR criterion. Datapoints are collected one at a time. We here only consider myopic optimization, that is, at each stage, we select the next



(a) xy projection



(b) xz projection



(c) yz projection

Figure 4.7: Projection of the excursion set (small scenario) for the first ground truth. Island boundary denoted in blue, observation location from previous field campaign denoted by black dots. Distances are displayed in [m] and density in [kg/m^3].

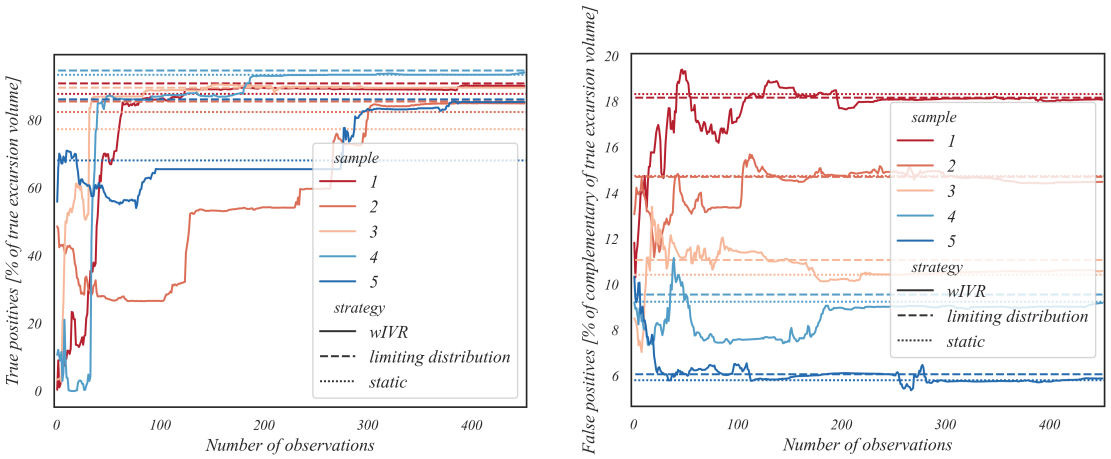
observation site u_{n+1} according to:

$$u_{n+1} = \arg\min_{u \in \boldsymbol{U}_c} \mathrm{wIVR}^n(u),$$

where ties are broken arbitrarily. Here $\boldsymbol{U}_c$ is a set of candidates among which to pick the next observation location. In our experiments, we fix $\boldsymbol{U}_c$ to consist of all surface points within a ball of radius 150 meters around the last observation location. Results are summarized in Figs. 4.8 and 4.9, which shows the evolution of the fraction of true positives and false positives as a function of the number of observations gathered.

We see that in the *large* scenario (Fig. 4.8) the wIVR criterion is able to correctly detect 70 to 80% of the excursion set (in volume) for each ground truth after 450 observations. For the *small* scenario (Fig. 4.9) the amount of true positives reached after 450 observations is similar, though two ground truths are harder to detect.

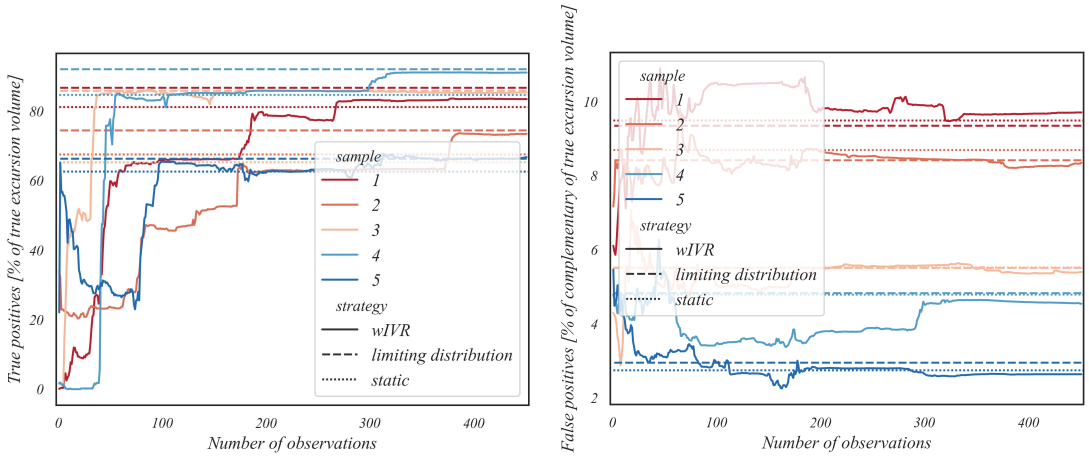
Note that in Figs. 4.8 and 4.9 the fraction of false negatives is expressed as a



(a) True positives (b) False positives

Figure 4.8: Evolution of true and false positives for the *large* scenario as a function of the number of observations.

percentage of the volume of the complementary of the true excursion set $D \setminus \Gamma^*$. We see that the average percentage of false positives after 450 observations tends to lie between 5 and 15%, with smaller excursion sets yielding fewer false positives. While the Vorob'ev expectation is not designed to minimize the amount of false positives, there exists *conservative set estimators* (Azzimonti et al., 2021) that specialize on this task. We identify the extension of such estimators to inverse problems as a promising venue for new research.

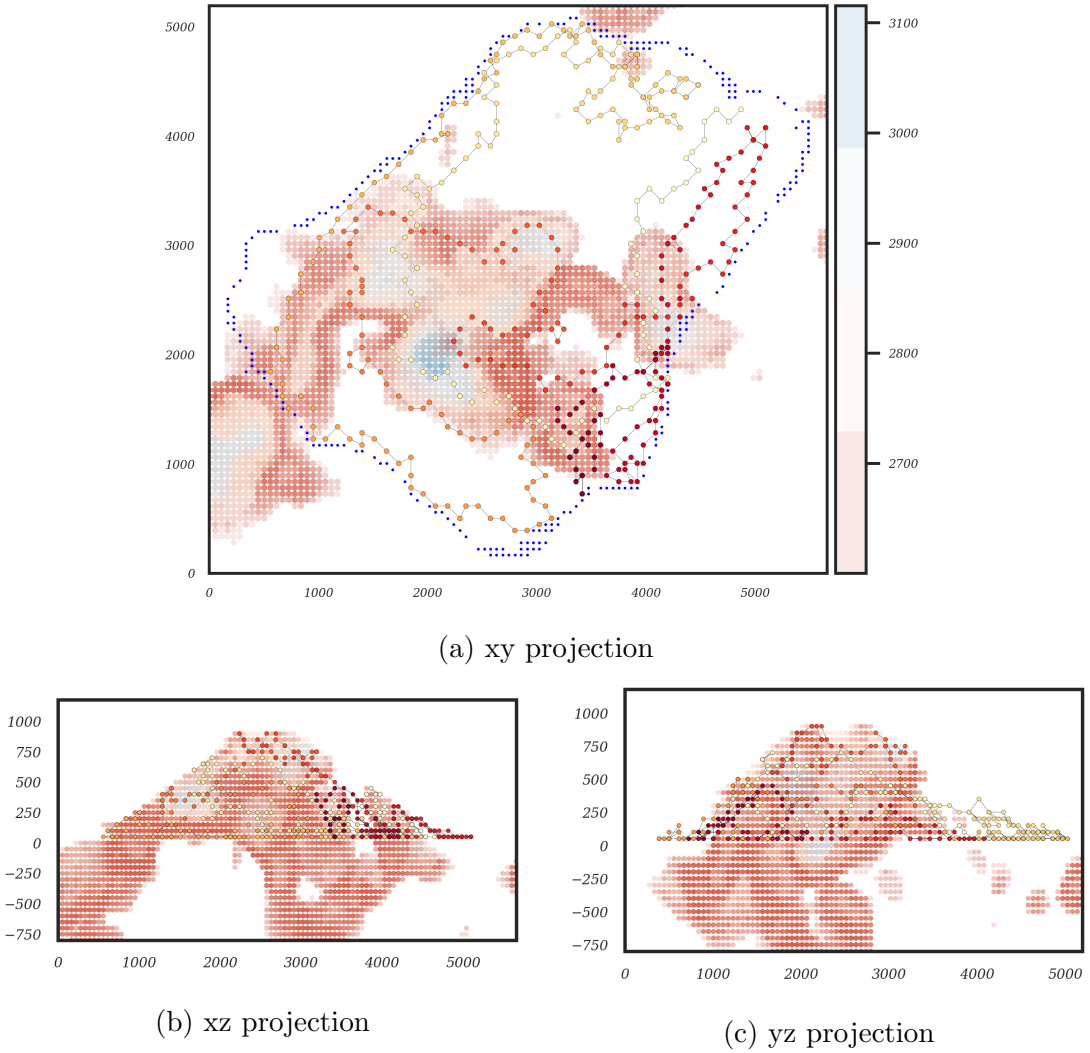


(a) True positives (b) False positives

Figure 4.9: Evolution of true and false positives for the *small* scenario as a function of the number of observations.

In both figures we also plot the fraction of true positives and false positives that result from the data collection plan that was used in Linde et al. (2014). Here only the situation at the end of the data collection process is shown. We see that for some of the ground truths the wIVR criterion is able to outperform static designs by around 10%. Note that there are ground truths where it performs similarly to a static design. We believe this is due to the fact that for certain ground truths

most of the information about the excursion set can be gathered by spreading the observations across the volcano, which is the case for the static design that also considers where it is practical and safe to measure.



(a) xy projection

(b) xz projection

(c) yz projection

Figure 4.10: Projection of the true excursion set (small scenario) and visited locations (wIVR strategy) for the first ground truth. Island boundary is shown in blue. Distances are displayed in [m] and density in [kg/m^3].

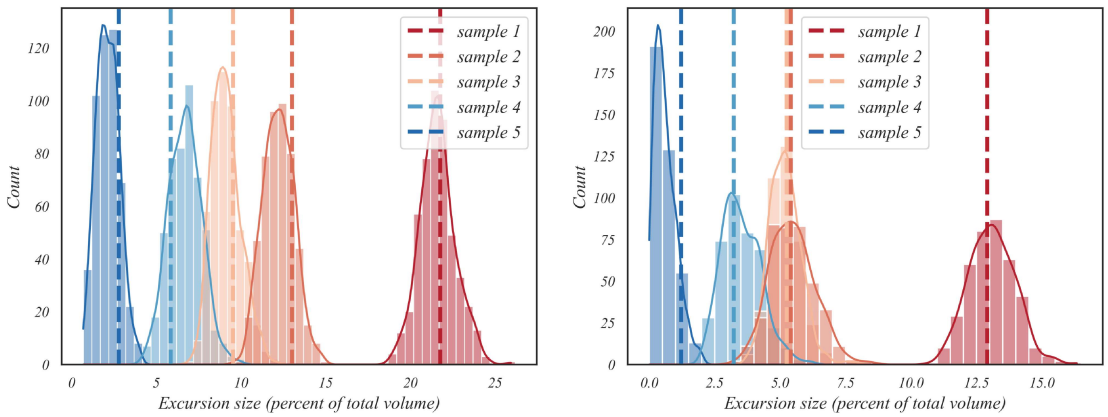
Limiting Distribution: The dashed horizontal lines in Figs. 4.8 and 4.9 show the detection percentage that can be achieved using the *limiting distribution*. We define the *limiting distribution* as the posterior distribution one were to obtain if one had gathered data at all allowed locations (everywhere on the volcano surface). This distribution may be approximated by gathering data at all points of a given (fine-grained) discretization of the surface. In general, this is hard to compute since it requires ingestion of a very large amount of data, but thanks to our implicit representation (Section 4.3) we can get access to this object, thereby, allowing new forms of uncertainty quantification.

In a sense, the limiting distribution represents *the best we can hope for* when covering the volcano with this type of measurements (gravimetric). It gives a measure

of the residual uncertainty inherent to the type of observations used (gravimetric). Indeed, it is known that a given density field is not identifiable from gravimetric data alone (see Blakely (1995) for example). Even if gravity data will never allow for a perfect reconstruction of the excursion set, we can use the limiting distribution to compare the performance of different sequential design criteria and strategies. It also provides a mean of quantifying the remaining uncertainty under the chosen class of models. A sensible performance metric is then the number of observations that a given criterion needs to approach the minimal level of residual uncertainty which is given by the limiting distributions.

As a last remark, we stress that the above results and the corresponding reconstruction qualities are tied to an estimator, in our case the Vorob'ev expectation. If one were to use another estimator for the excursion set, those results could change significantly.

Posterior Volume Distribution: Thanks to our extension of the residual kriging algorithm to inverse problems (see Section 4.4.2), we are able to sample from the posterior at the end of the data collection process. This opens new venues for uncertainty quantification in inverse problems. For example, we can use sampling to estimate the posterior distribution of the excursion volume and estimate the residual uncertainty on the size of the excursion set.



(a) (*large* scenario) threshold: 2500 $[kg/m^3]$ (b) (*small* scenario) threshold: 2600 $[kg/m^3]$

Figure 4.11: Empirical posterior distribution (after 450 observations) of the excursion volume for each ground truth. True volumes are denoted by vertical lines.

Fig. 4.11 shows the empirical posterior distribution of the excursion volume for each of the ground truths considered in the preceding experiments. When compared to the prior distribution, Fig. 4.6, one sees that the wIVR criterion is capable of significantly reducing the uncertainty on the excursion volume. This shows that though the location of the excursion set can only be recovered with limited accuracy, as shown in Figs. 4.8 and 4.9, the excursion volume can be estimated quite well. This is surprising given that the criterion used (wIVR) is a very crude one and was not designed for that task. On the other hand, there exist more refined criteria, like the so-called *SUR strategies* (sequential uncertainty reduction) (Chevalier et al., 2014a; Bect et al., 2019), among which some were specifically engineered to reduce the uncertainty on the excursion volume (Bect et al., 2012). Even though those

criteria are more computationally challenging than the wIVR one, especially in the considered framework, we identify their application to large Bayesian inverse problem as a promising avenue for future research.

4.5 Conclusion and Perspectives

Leveraging the new results about sequential disintegrations of Gaussian measures developed in Chapter 3, we have introduced an implicit almost matrix free representation of the posterior covariance of a GP and have demonstrated fast update of the posterior covariance on large grids under general linear functional observations. Our method allows streamline updating and fast extraction of posterior covariance information even when the matrices are larger than the available computing memory. Using our novel implicit representation, we have shown how targeted design criteria for excursion set recovery may be extended to inverse problems discretized on large grids. We also demonstrated UQ on such problems using posterior sampling via residual kriging. Our results suggest that using the considered design criteria allows reaching close-to-minimal levels of residual uncertainty using a moderate number of observations and also exhibit significant reduction of uncertainty on the excursion volume. The GP priors used in this work are meant as a proof of concept and future work should address the pitfalls of such priors, such as lack of positiveness of the realizations and lack of expressivity. Other promising research avenues include extensions to more sophisticated estimators such as conservative estimates Azzimonti et al. (2021). On the dynamic programming side, extending the myopic optimization of the criterion to finite horizon optimization in order to provide optimized data collection trajectories is an obvious next step which could have significant impact on the geophysics community. Also, including location dependent observation costs such as accessibility in the design criterion could help provide more realistic observation plans. These last two topics are briefly touched upon in the conclusion of this thesis.

4.6 Appendix A: Forward operator for Gravimetric Inversion

Given some subsurface density $\rho D \to \mathbb{R}$ inside a domain $D \subset \mathbb{R}^3$ and some location u outside the domain, the vertical component of the gravitational field at u is given by:

$$\mathfrak{G}_u[\rho] = \int_D \rho(x) g(x,u) dx, \tag{4.12}$$

with Green kernel

$$g(x,u) = \frac{x^{(3)} - u^{(3)}}{\|x-u\|^3}, \tag{4.13}$$

where $x^{(3)}$ denotes the vertical component of x.

We discretize the domain D into m identical cubic cells $D = \cup_{i=1}^m D_i$ with centroids $\boldsymbol{X} = (X_1, \ldots, X_m)$ and assume the mass density to be constant over each

cell, so the field ρ may be approximated by the vector $\rho\boldsymbol{X}$. The vertical component of the gravitational field at u is then given by:

$$\int_{\cup_{i=1}^m D_i} g(x,u)\rho(x)dx \approx \sum_{i=1}^m \left(\int_{D_i} g(x,u)dx\right)\rho X_i := G_u\rho\boldsymbol{X}.$$

Integrals of Green kernels over cuboids may be computed using the *Banerjee formula* (Banerjee and Das Gupta, 1977).

Theorem (Banerjee)**.** *The vertical gravity field at points (x_0, y_0, z_0) generated by a prism with corners $(x_h, x_l,\ y_h, y_l, ...)$ of uniform mass density ρ is given by:*

$$\begin{aligned} g_z = \frac{1}{2}\gamma_N\rho\Bigg[& x\ \log\Big(\frac{\sqrt{x^2+y^2+z^2}+y}{\sqrt{x^2+y^2+z^2}-y}\Big) \\ & + y\ \log\Big(\frac{\sqrt{x^2+y^2+z^2}+x}{\sqrt{x^2+y^2+z^2}-x}\Big) \\ & - 2z\ \arctan\Big(\frac{xy}{z\sqrt{x^2+y^2+z^2}}\Big)\Bigg]\Bigg|_{x_l-x_0}^{x_h-x_0}\Bigg|_{y_l-y_0}^{y_h-y_0}\Bigg|_{z_l-z_0}^{z_h-z_0} \end{aligned}$$

4.7 Appendix B: Proofs

Proof. (Lemma 4) We proceed by induction. The case $n = 1$ follows from Eq. (4.4). The induction step is directly given by Theorem 13. □

Proof. (Lemma 5) The product is computed using Algorithm 1. It involves multiplication of A with the prior covariance, which costs $\mathcal{O}(m^2 a)$ and multiplication with all the previous intermediate matrices, which contribute $\mathcal{O}(mq_c a)$ and $\mathcal{O}(q_c^2 a)$ respectively, at each stage. □

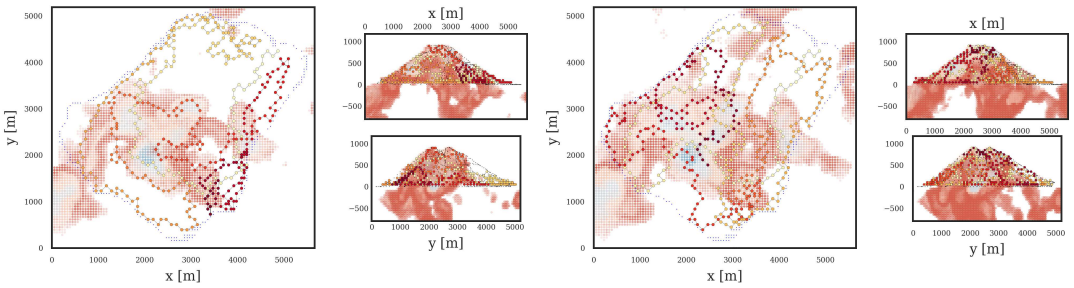
Proof. (Lemma 6) The cost of computing the i-th pushforward $\bar{K}_i$ is $\mathcal{O}(m^2 q_c + i(mq_c^2 + q_c^3))$. Summing this cost for all stages $i = 1, \ldots, n$ then gives $\mathcal{O}(m^2 Q + mQ^2 + q^2 q_c)$. To that cost, one should add the cost of computing R_i^{-1}, which costs $\mathcal{O}(q_c^3)$ at each stage, yielding a $\mathcal{O}(Qq_c^2)$ contribution to the total cost, which is dominated by $Q^2 q_c$ since $q_c < Q$. □

4.8 Appendix C: Supplementary Experimental Results

We here include more detailed analysis of the results of Section 4.4.3 that do not fit in the main text.

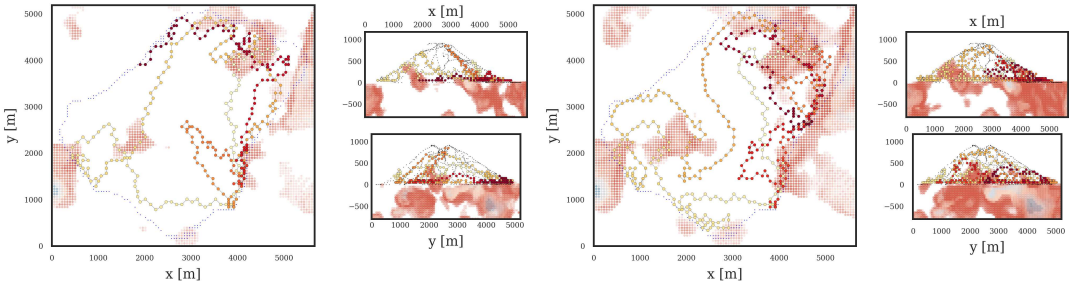
Figs. 4.8 and 4.9 showed that there are differences in detection performance for the different ground truths. These can be better understood by plotting the actual location of the excursion set for each of the ground truths as well as the observation locations chosen by the wIVR criterion, as done in Fig. 4.12. One sees that the (comparatively) poor performance shown by Fig. 4.9 for *Sample 2* in the *small* scenario may be explained by the fact that, for this ground truth, the excursion set

is located mostly outside the accessible data collection zone (island surface), so that the strategy is never able to collect data directly above the excursion.



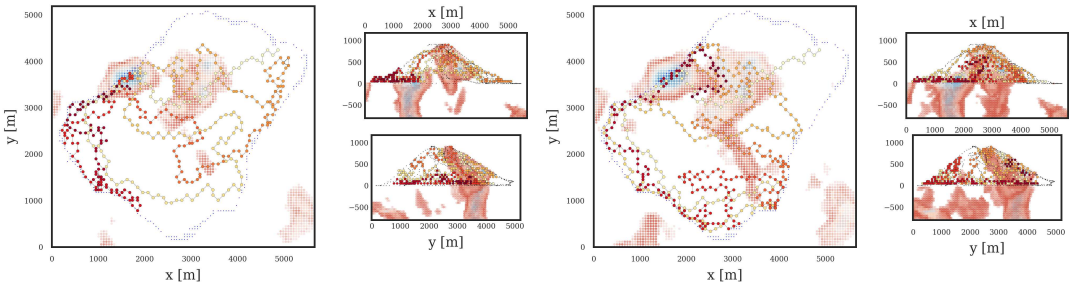
(a) sample 1: small scenario

(b) sample 1: large scenario



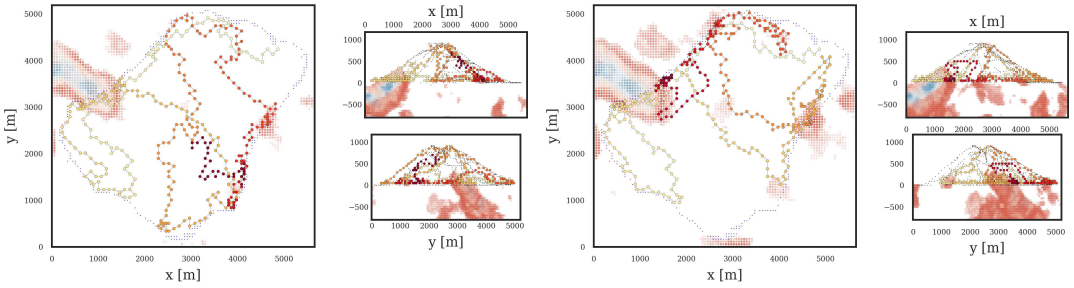
(c) sample 2: small scenario

(d) sample 2: large scenario



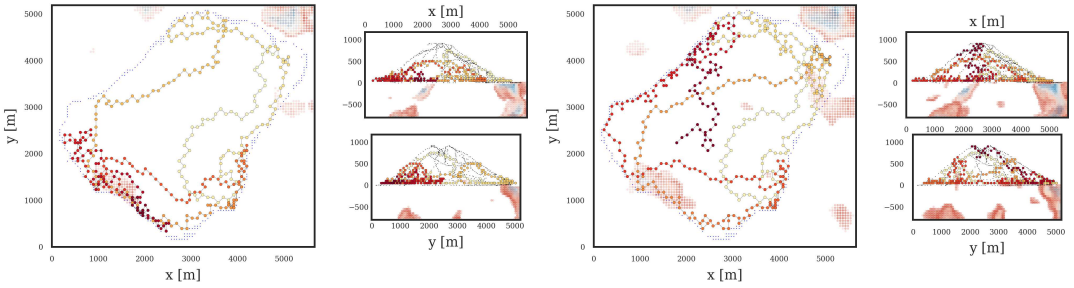
(e) sample 3: small scenario

(f) sample 3: large scenario



(g) sample 4: small scenario

(h) sample 4: large scenario



(i) sample 5: small scenario

(j) sample 5: large scenario

Figure 4.12: True excursion set and visited locations (wIVR strategy). Island boundary is shown in blue.

Chapter 5

Universal Inversion: Bayesian Inversion with Trends

5.1 Introduction

After having introduced techniques to apply Gaussian processes to large-scale inverse problems, a natural next step is to try to extend other methods from geostatistics to be brought to bear on such problems. Among these methods, the first obvious candidate is *universal kriging* (Matheron, 1969).

The idea behind universal kriging is to fit a parametric linear model to the response and to use a centered GP to fit the fluctuations around the trend. This model allows one to include expert knowledge in the definition of the trend and is one of the simplest way to extend GP regression to non-stationary mean functions. Universal kriging has found use in areas as diverse as the study of air pollutants (Romary et al., 2011), the modelling of forest inventory (Mandallaz, 2000) and the prediction of road traffic (Selby and Kockelman, 2013). Considering the broad potential applications, it is natural to try to extend universal kriging to inverse problems. The first attempt in this direction was the seminal paper (Kitanidis, 1995). We here expand on this contribution to bring it up to date with the state of the art Bayesian inversion framework and apply it to real-world large-scale inverse problems (whereas (Kitanidis, 1995) only considered one-dimensional toy examples).

As in Chapter 4, this chapter will use the gravimetric inversion problem from Section 2.4.1 as an applicative red thread. The inclusion of user-defined trends in such gravimetric inversion problems is of great interest since it allows the incorporation of a-priori knowledge about the underground geology in the domain of interest, be it layer structures, depth dependencies or tunnel-like networks. We note that the inclusion of such structural knowledge in the inversion process has already gathered some attention, be it via layer deformations (Berrino and Camacho, 2008), or via object-based methods that leverage knowledge of the possible geological processes that have led to the current underground situation (Guillen et al., 2008). Nevertheless, those methods are either purely deterministic, or Monte Carlo based, whereas our universal inversion algorithm provides fast, analytic expressions for the full posterior.

5.2 Background: Universal Kriging

We start by recalling the main ideas behind universal kriging. Instead of following the usual formulation of the universal kriging predictor as the best linear unbiased predictor (BLUP), we here follow the less common Bayesian approach (Omre and Halvorsen, 1989; Helbert et al., 2009). Readers interested in more conventional treatments of the subject matter are referred to (Chilès and Delfiner, 2012; Cressie, 1993). A good introduction to both approaches can also be found in (Bachoc, 2013).

Universal kriging models the unknown phenomenon of interest as a sum of a trend and fluctuations around the trend described by a centered Gaussian process. The goal is to use the trend to incorporate a priori expert knowledge, while the GP part is meant to fit local "residuals" that are not captured by the trend. The trend is encoded as a linear combination of user specified basis functions $f_i : D \to \mathbb{R}$ with unknown coefficients β_i. The full model writes as:

$$Z_x = \eta_x + \sum_{i=1}^{b} \beta_i f_i(x). \tag{5.1}$$

In the Bayesian approach, one assumes that the trend coefficients are endowed with some prior distribution, here a multivariate Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_{\text{prior}}, \boldsymbol{\Sigma})$. Then, conditionally on observed field values $\boldsymbol{y} \in \mathbb{R}^q$ at locations $\boldsymbol{W} \in D^r$ the posterior of the trend parameters is Gaussian with mean vector and covariance matrix given by:

$$\boldsymbol{\beta}_{\text{post}} = \boldsymbol{\beta}_{\text{prior}} + \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T Q^{-1} \left(\boldsymbol{y} - F_{\boldsymbol{W}} \boldsymbol{\beta}_{\text{prior}}\right) \tag{5.2}$$

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T Q^{-1} F_{\boldsymbol{W}} \boldsymbol{\Sigma}, \tag{5.3}$$

where the data is assumed to be corrupted by additive Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$ and the matrix: $Q := (F_{\boldsymbol{W}} \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T + K_{\boldsymbol{WW}} + \Delta)$ is assumed invertible. We use the shorthand notation $F_{\boldsymbol{W}}$ to denote the $r \times b$ matrix with elements $f_j(\boldsymbol{W}_i)$. The posterior law of the full GP Z is also Gaussian, with mean and covariance function given by:

$$\tilde{m}_{\boldsymbol{X}} = F_{\boldsymbol{X}} \boldsymbol{\beta}_{\text{prior}} + \left(F_{\boldsymbol{X}} \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T + K_{\boldsymbol{XW}}\right) Q^{-1} \left(\boldsymbol{y} - F_{\boldsymbol{W}} \boldsymbol{\beta}_{\text{prior}}\right) \tag{5.4}$$

$$\tilde{K}_{\boldsymbol{XX'}} = K\boldsymbol{XX'} + F_{\boldsymbol{X}} \boldsymbol{\Sigma} F_{\boldsymbol{X'}}^T - \left(F_{\boldsymbol{X}} \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T + K_{\boldsymbol{XW}}\right) Q^{-1} \left(F_{\boldsymbol{W}} \boldsymbol{\Sigma} F_{\boldsymbol{X'}}^T + K_{\boldsymbol{WX'}}\right). \tag{5.5}$$

From the above *Bayesian kriging* equations, universal kriging can be recovered as a limiting case. Universal kriging first estimates the trend coefficients by maximum likelihood and then computes the kriging predictor as the BLUP. Considering Eq. (5.3) and taking the limit of a flat prior, that is the limit where all eigenvalues of the trend prior covariance matrix $\boldsymbol{\Sigma}$ tend to infinity, the optimal trend coefficient vector writes (see (Omre and Halvorsen, 1989) for details):

$$\hat{\boldsymbol{\beta}}_{UK} = \left(F_{\boldsymbol{W}} R^{-1} F_{\boldsymbol{W}}^T\right) F_{\boldsymbol{W}} R^{-1} \boldsymbol{y}, \tag{5.6}$$

where the matrix $R := (K_{\boldsymbol{WW}} + \Delta)$ is assumed invertible. The conditional mean and covariance function of the field itself tend to:

$$\tilde{m}_{\boldsymbol{X}}^{UK} = F_{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{UK} + K_{\boldsymbol{XW}} R^{-1} \left(\boldsymbol{y} - F_{\boldsymbol{W}} \hat{\boldsymbol{\beta}}_{UK}\right) \tag{5.7}$$

$$\tilde{K}_{\boldsymbol{XX'}}^{UK} = K_{\boldsymbol{XX'}} - K_{\boldsymbol{XW}} R^{-1} K_{\boldsymbol{WX'}}, \tag{5.8}$$

and we see that we recover the usual universal kriging equations.

5.3 Universal Inversion

We now extend the universal/Bayesian kriging equations from last section to the case of linear operator observations, to allow for their use in Bayesian inverse problem. In the following, we consider a GP model with a parametric trend as in Eq. (5.1) and (discretized) linear operator observations as in Eq. (4.2). The posterior can then be computed in a similar way as in usual universal kriging.

Theorem 14. *Let Z be a Gaussian process on D with parametric trend as in Eq. (5.1) and let $G \in \mathbb{R}^{q\times r}$ be a linear operator. Assume that one observes data of the form:*

$$\boldsymbol{Y} = GZ_{\boldsymbol{W}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$ is some additive Gaussian noise. Then, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$, the posterior of the trend coefficients is Gaussian, with mean and covariance given by:

$$\boldsymbol{\beta}_{\text{post}} = \boldsymbol{\beta}_{\text{prior}} + \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T G^T Q^{-1} \left(\boldsymbol{y} - GF_{\boldsymbol{W}}\boldsymbol{\beta}_{\text{prior}}\right) \tag{5.9}$$

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma} F_{\boldsymbol{W}}^T G^T Q^{-1} G F_{\boldsymbol{W}} \boldsymbol{\Sigma}, \tag{5.10}$$

where the matrix: $Q := G(F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T + K_{\boldsymbol{WW}})G^T + \Delta$ is assumed invertible. Furthermore, the posterior of the GP is also a GP with mean and covariance given by:

$$\tilde{m}_{\boldsymbol{X}} = F_{\boldsymbol{X}}\boldsymbol{\beta}_{\text{prior}} + \left(F_{\boldsymbol{X}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T + K_{\boldsymbol{XW}}\right) G^T Q^{-1} \left(\boldsymbol{y} - GF_{\boldsymbol{W}}\boldsymbol{\beta}_{\text{prior}}\right) \tag{5.11}$$

$$\begin{aligned}\tilde{K}_{\boldsymbol{XX'}} = K_{\boldsymbol{XX'}} &+ F_{\boldsymbol{X}}\boldsymbol{\Sigma}F_{\boldsymbol{X'}}^T \\ &- \left(F_{\boldsymbol{X}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T + K_{\boldsymbol{XW}}\right) G^T Q^{-1} G \left(F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{X'}}^T + K_{\boldsymbol{WX'}}\right).\end{aligned} \tag{5.12}$$

In the limit of an uninformative prior on the trend coefficients, the posterior mean function and covariance kernel of the GP reduce to:

$$\tilde{m}_{\boldsymbol{X}}^{UK} = F_{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + K_{\boldsymbol{XW}} G^T R^{-1} \left(\boldsymbol{y} - GF_{\boldsymbol{W}}\hat{\boldsymbol{\beta}}\right) \tag{5.13}$$

$$\begin{aligned}\tilde{K}_{\boldsymbol{XX'}}^{UK} = K_{\boldsymbol{XX'}} &- K_{\boldsymbol{XW}} G^T R^{-1} G K_{\boldsymbol{WX'}} \\ &+ \left(F_{\boldsymbol{X}} - K_{\boldsymbol{XW}} G^T R^{-1} F_{\boldsymbol{W}}\right) \left(F_{\boldsymbol{W}}^T R^{-1} F_{\boldsymbol{W}}\right)^{-1} \left(F_{\boldsymbol{X'}} - K_{\boldsymbol{X'W}} G^T R^{-1} F_{\boldsymbol{W}}\right)^T,\end{aligned} \tag{5.14}$$

where the optimal trend coefficients are given by:

$$\hat{\boldsymbol{\beta}} = \left(F_{\boldsymbol{W}} G^T R^{-1} G F_{\boldsymbol{W}}^T\right)^{-1} F_{\boldsymbol{W}} G^T R^{-1} \boldsymbol{y}, \tag{5.15}$$

and the matrix $R := \left(GK_{\boldsymbol{WW}}G^T + \Delta\right)$ is assumed invertible. We note that in this case, the estimator $\hat{\boldsymbol{\beta}}$ corresponds to the maximum likelihood estimator. Interested readers are referred to (Bachoc, 2013) for more details on parameter estimation procedures in universal kriging.

Remark 8 (Covariance Estimation)**.** It is well known in the geostatistics community that universal kriging is fraught with problems when it comes to covariance estimation. Indeed, the universal kriging equations Eqs. (5.4) to (5.6) require the covariance to be known in order to estimate the trend, but the estimation of the

covariance itself depends on the estimated trend, leading to a chicken and egg problem (Armstrong, 1984; Cressie, 1993). While we are well aware of the potential pitfalls originating in the interplay between the trend and the covariance, we leave this question for future work and will content ourselves with MLE for estimating all parameters, as a first demonstration of universal kriging in inverse problems.

By using the MLE of the trend coefficients, the log-likelihood writes as (up to a constant and a prefactor):

$$L \propto \log|R| + \left(\boldsymbol{y} - GF_{\boldsymbol{W}}\hat{\boldsymbol{\beta}}\right)^T R^{-1}\left(\boldsymbol{y} - GF_{\boldsymbol{W}}\hat{\boldsymbol{\beta}}\right) \tag{5.16}$$

5.4 Fast multiple-Fold Cross-Validation for Universal Inversion

One of the main statistical tools for model validation and parameter estimation is cross-validation (CV) (Stone, 1974). Compared to other procedures based on train-test dataset splitting, cross-validation is an approach that allows to make out the most of the available data, which is of particular importance in data-scarce settings, such as if often the case in inverse problems. Apart from these general considerations, in the special case of Universal Inversion, the additional freedom in the choice of the basis functions included in the trend model calls for a solution to detect and prevent overfitting, making cross-validation even more appealing in this context.

There exists a whole array of different cross-validation procedures, among which practitioners should choose according to their statistical goals (Arlot and Celisse, 2010), the most widespread forms of CV being leave-one out (LOOCV), leave-k-out and multiple fold CV. While being the simplest and one of the computationally cheapest form of CV, leave-one-out possesses several flaws and is overoptimistic in the presence of highly correlated datapoints. On the other hand, more sophisticated CV procedures tend to suffer from combinatorial explosion and can be computationally expansive even for small datasets. To overcome these limitations, fast CV formulae have been developed (Ginsbourger and Schärer, 2021). We next leverage those formulae for fast cross-validation in the Universal Inversion setting.

Notation: The basic principle behind CV is to remove parts of the dataset from the fitting step and then compute the error in predicting these held-out datapoints using the model fitted on the restricted dataset. This calls for a notation that is able to express the action of selecting subsets of a given dataset. In the following, we assume that q datapoints are available. Then, a boldface 'i' will be used to denote a strictly ordered (no repetitions) subset of data indices: $\boldsymbol{i} \subset \{1, \ldots, q\}$. Given a q-dimensional vector $\boldsymbol{y}$, we use $\boldsymbol{y_i}$ to denote the subvector built from $\boldsymbol{y}$ by extracting the elements whose indices belong to $\boldsymbol{i}$ and $\boldsymbol{y_{-i}}$ to denote the remaining part of the original vector.

For the rest of this section, assume that one has a GP model as in Theorem 14 with a flat prior on the trend coefficients (the full Bayesian case is not treated here) and (noiseless) observations $\boldsymbol{Y} = GZ_{\boldsymbol{W}}$. As noted in (Ginsbourger and Schärer, 2021), all the information about the distribution of the CV residuals can be encoded

in sub-blocks of a single matrix, which, in the Universal Inversion case, is given by the augmented $(q+b)\times(q+b)$ matrix

$$\tilde{K} = \begin{pmatrix} GK_{\boldsymbol{WW}}G^T & GF_{\boldsymbol{W}} \\ F_{\boldsymbol{W}}^T G^T & \mathbf{0} \end{pmatrix}.$$

Our goal is to derive efficient formulae for the characterization of the distribution of the CV residuals $\boldsymbol{E_i} := \boldsymbol{Y_i} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}}$, where $\hat{\boldsymbol{Y}}_{\boldsymbol{i}} = G\hat{Z}_{\boldsymbol{W}}^{-\boldsymbol{i}}$ denotes the prediction of the left-out data using the remaining data $\boldsymbol{Y}_{-\boldsymbol{i}}$ and $\hat{Z}^{-\boldsymbol{i}}$ denotes the posterior mean Eq. (5.13) conditional on this part of the data. Leveraging (Ginsbourger and Schärer, 2021, Corollary 1) and noting that $GZ_{\boldsymbol{W}}$ is a Gaussian random vector with covariance matrix $GK_{\boldsymbol{WW}}G^T$ and trend $GF_{\boldsymbol{W}}$, the CV residuals can then be computed as:

Theorem 15. *Let Z be a Gaussian process on D with parametric trend as in Eq. (5.1), furthermore let $G \in \mathbb{R}^{q\times r}$ be a linear operator and assume we have data $\boldsymbol{Y} = GZ_{\boldsymbol{W}}$. Then, for any two strictly ordered subset of indices $\boldsymbol{i}, \boldsymbol{j}$, the cross-validation residuals can be written as:*

$$\boldsymbol{E_i} = \left(\tilde{K}_{\boldsymbol{ii}}^{-1}\right)^{-1}\left(\tilde{K}^{-1}[1:q, 1:q]\boldsymbol{Y}\right)_{\boldsymbol{i}}, \tag{5.17}$$

where $\tilde{K}_{\boldsymbol{ii}}^{-1}$ denotes the $\boldsymbol{ii}$ subblock of the $\tilde{K}^{-1}$ matrix and $\tilde{K}[:, 1:q]$ stands for the first q columns of $\tilde{K}$. Furthermore, the residuals are jointly Gaussian distributed, centered and with covariance

$$\mathrm{Cov}\left(\boldsymbol{E_i}, \boldsymbol{E_j}\right) = \left(\tilde{K}_{\boldsymbol{ii}}^{-1}\right)^{-1}\tilde{K}_{\boldsymbol{ij}}^{-1}\left(\tilde{K}_{\boldsymbol{jj}}^{-1}\right)^{-1}$$

The above allows for the computation of any CV residual by extracting sub-blocks from $\tilde{K}^{-1}$, which can lead to substantial computational savings by avoiding recomputation of the posterior at each cross-validation pass.

5.5 Application: Gravimetric Inversion with Trends

Now that we have developed a full-fledged framework for universal inversion, we demonstrate its versatility by applying it to our gravimetric inverse problem example Section 2.4.1. Our main goal is to improve upon the results from Chapter 4 by including expert knowledge in the trend of the GP prior.

In the following, we will consider a few basic trend models to demonstrate the main features of Universal Inversion, leaving the inclusion of more sophisticated, domain-informed trends to future work. In the specific case of volcano gravimetric inversion Section 2.4.1, there are a few natural trend functions that one may want to consider. Using a right-handed coordinate system x, y, z with x pointing south, one can consider the following basis functions:

- **Depth-dependence** Due to the way volcanoes aggregate mass, one expects the density field to have some dependence on the depth. We here model this phenomenon using the naive basis function:

  $$f_{\mathrm{depth}}(x, y, z) = z_{\mathrm{top}} - z$$

  , where z_{top} denotes the altitude of the volcano top.

- **Chimney:** Volcanoes from around a central lava conduit. For most active volcanoes, one can expect higher densities within that region, which can be modelled as a cylindrical dependency around the central axis:

$$f_{\text{chimney}}(x, y, z) = 1 - \kappa\sqrt{(x - x_0)^2 + (y - y_0)^2}$$

, where κ is some fixed scale constant and x_0, y_0 denote the planar coordinates of the chimney axis.

- **Fault-line:** Some active volcanoes such as the Stromboli separate along a fault line, through which dense magma can infiltrate, leading to higher densities around the fault. This can be modelled as a dependence on the distance to a fixed plane Π representing the fault line:

$$f_{\text{fault}}(x, y, z) = \tanh\left(\frac{\pi}{2\kappa}\text{dist}\left((x, y, z), \Pi\right)\right)$$

, where we use a tanh activation function to sharpen the transition and κ is a scale parameter controlling the cut-off of the dependency.

- **Layers:** It is sometimes known from the history of the volcano formation that there exists layers of different materials within the volcano, each layer having a different typical density. This can be modelled by summing

- **Piecewise domains** One might also want to allow for regions with different typical density. This can be achieved by considering trend functions that are given by the characteristic functions of some domains within the volcano.

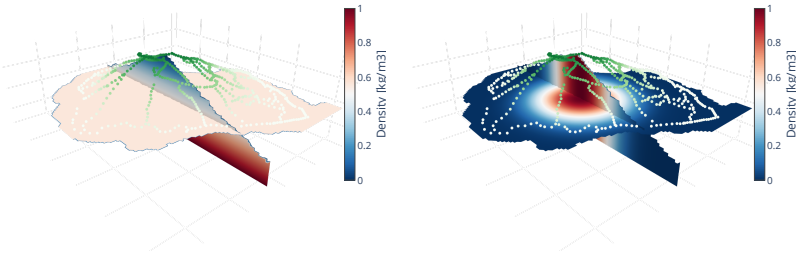
Several of the above mentioned trend functions depend on parameters that have to be fixed beforehand. We stress that universal inversion only estimates the coefficients $\boldsymbol{\beta}$ of the trend functions, but not the free parameters within the functions themselves. Those have to be chosen according to expert knowledge, or via some other estimation procedure, the consideration of which we leave to future work. We note that among all the aforementioned trend functions, the fault line one is the most interesting for our Stromboli example, since it is known that this volcano indeed separates along a fault line whose direction is known (Linde et al., 2014).

5.5.1 Cross-Validation and Model Selection

We now use Universal Inversion with the trend models from the previous section to invert the Stromboli gravimetric data from Section 2.4.1. For each trend model, the hyperparameters of the GP are trained using MLE. The posterior mean for each model is shown in Fig. 5.2.

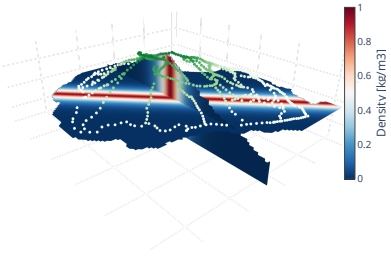
All models are in qualitative agreement with what is known about the interior of the Stromboli volcano (Linde et al., 2014). Owing to the difficulty of collecting direct samples of the density field, the last available option for model selection is cross-validation.

In order to further discriminate among the trend models, we leverage the fast cross-validation formulae from Theorem 15 to bring CV to bear on the problem of model identification. In the following, we consider a similar setting as the one in Theorem 15 and define the k-fold cross-validation criterion following (Arlot and Celisse, 2010).

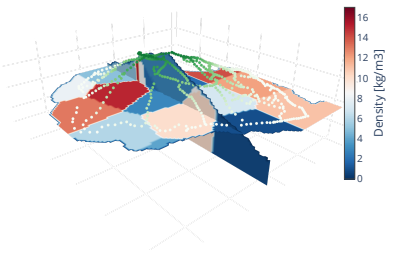


(a) depth-dependence

(b) chimney

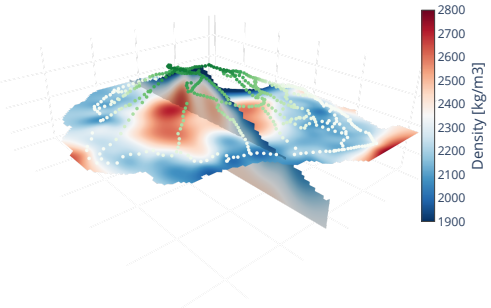


(c) fault-line

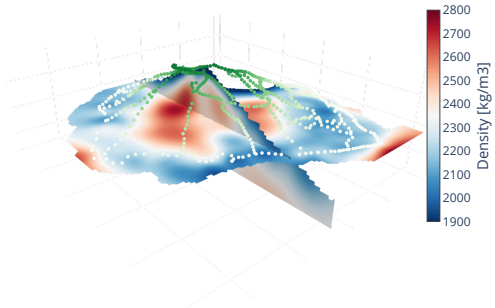


(d) domains

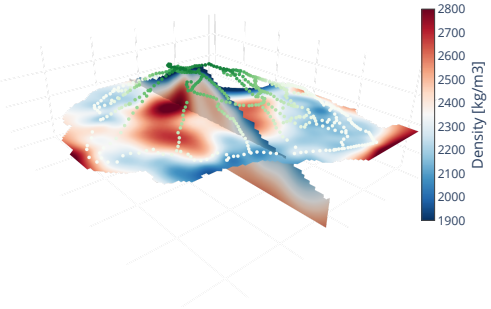
Figure 5.1: Trend basis functions used in the experiments. Solid balls denote the locations of the gravimetry observations (Stromboli dataset).



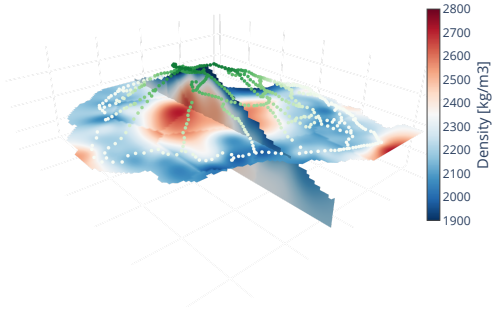
(a) constant



(b) chimney



(c) fault-line



(d) domains

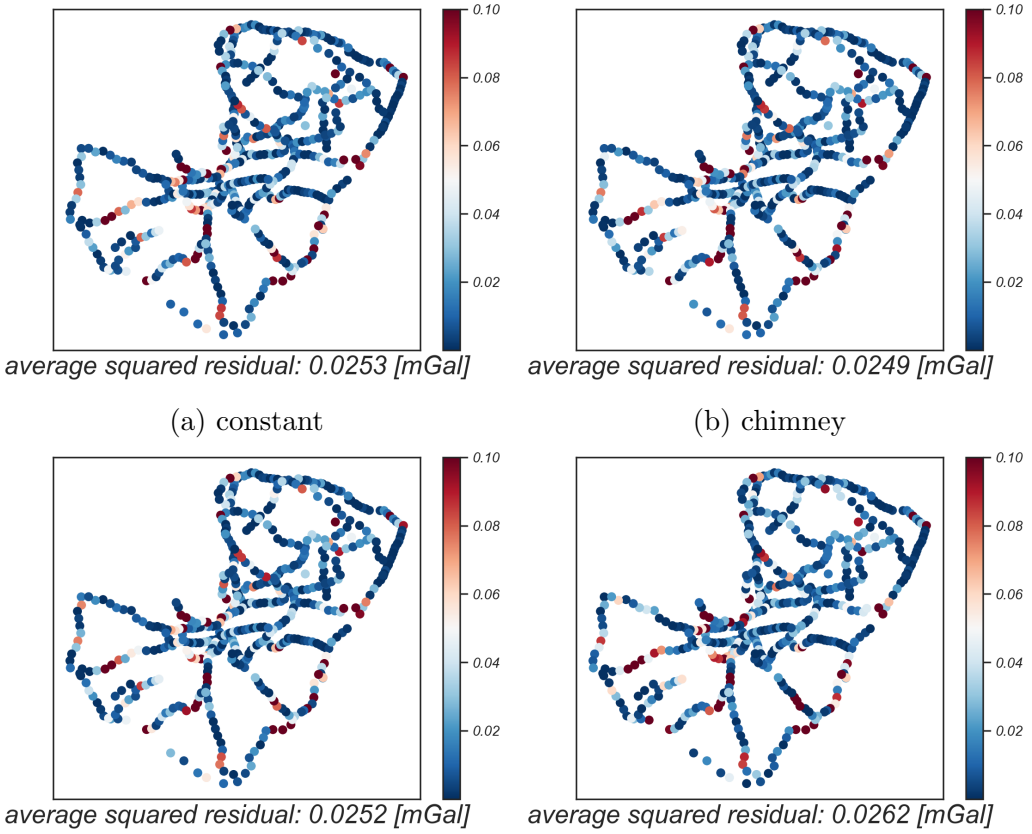
Figure 5.2: Posterior mean for the various trend models (Stromboli data).

k-Fold Cross-Validation Criterion. *Consider a GP model with parametric trend as in Eq.* (5.1) *and a q-dimensional data vector* $\boldsymbol{y}$ *which is a realization of the data model Eq.* (4.2)*. Then, given a k-partition of* $\{1, ..., q\}$ *into folds* $\boldsymbol{I} = (\boldsymbol{i}_1, ..., \boldsymbol{i}_k)$*, the k-folds cross validation criterion is given by:*

$$\mathcal{L}_{\mathrm{CV}}(Z, \boldsymbol{y}, \boldsymbol{I}) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{|\boldsymbol{i}_j|} ||\boldsymbol{e}_{\boldsymbol{i}_j}||^2, \tag{5.18}$$

where $\boldsymbol{e}_{\boldsymbol{i}_j}$ *denotes the cross-validation residual that corresponds to the realization of the (random) residual Eq.* (5.17)*.*

One can then use the above criterion to select between different trend models. We begin by considering leave-one-out residuals. Results for the considered trend models are displayed in Fig. 5.3.



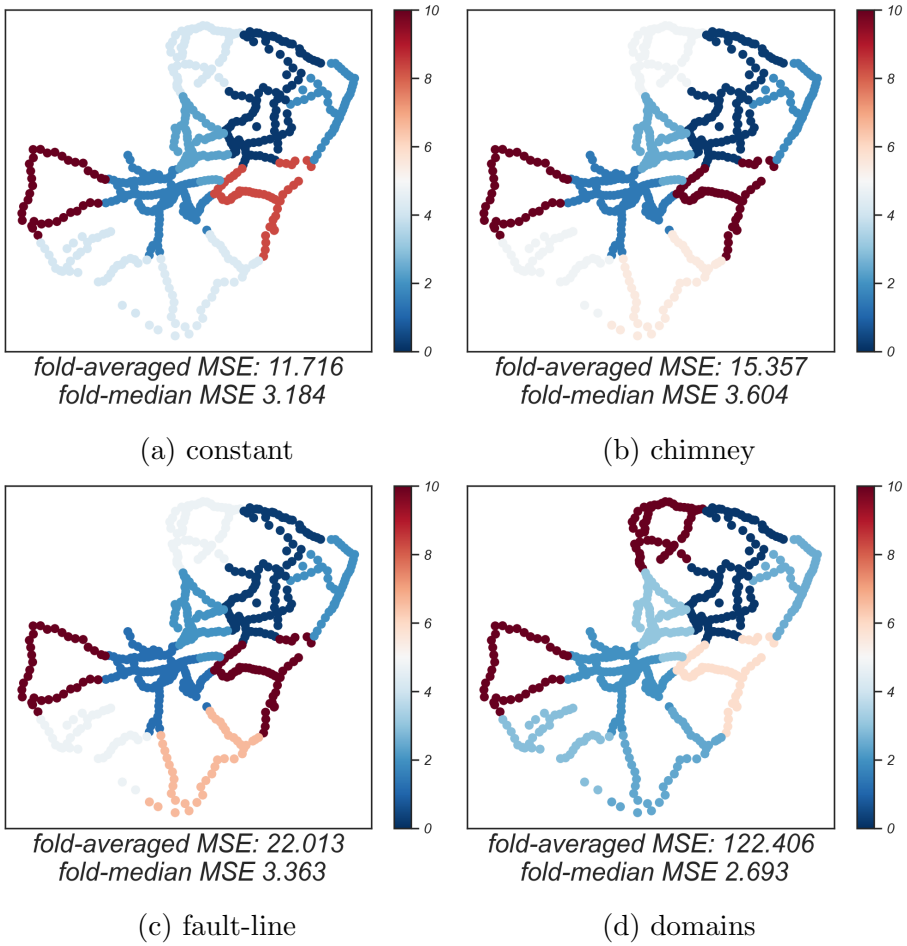
(a) constant (b) chimney

(c) fault-line (d) domains

Figure 5.3: Leave-one-out cross-validation residuals [mGal] for the different trend models. RMS residual over all data points is also given for each trend model.

One sees that the residuals are sensibly the same across all trend models, meaning that leave-one-out is unable to discriminate among different trends for the gravimetric inverse problem under consideration. In retrospect, such an inconclusive result does not seem suprising, considering the high correlation of gravimetric observations at close-by locations that results from the integral nature of the observation operator Eq. (4.12). In order to overcome this difficulty, we next resort to k-fold cross-validation.

The first difficulty that arises when using k-fold CV is that of defining the folds. To the best of our knowledge, no established procedure exists for such a task. As a first candidate procedure for fold definition we use spatial clustering. In our gravimetric inverse problem, this is a sensible procedure since we expect nearby gravimetric observations to be highly correlated, so that folds should be as spatially separated as possible in order to be informative (in a sense that remains to be defined). Figure 5.4 shows the k-fold residuals for 10 folds defined using kMeans clustering.



(a) constant (b) chimney

(c) fault-line (d) domains

Figure 5.4: Root mean square k-fold cross-validation residuals [mGal] over each fold for the different trend models. RMS k-fold residual over all data points is also given for each trend model.

One sees that, unlike the leave-one-out case, cross-validation is now able to distinguish between the different trends, giving a significant preference to the bare constant model. Apart from model ranking, there is more information that is contained in the CV residuals. For example, one sees that all models, have a hard time predicting the cluster in the bottom left-part, indicating that the estimation of the trend is highly sensitive to this region of the volcano for these models.

While this might indeed be a signal that there is some interesting phenomenon to be studied in this region, one should also remember that this particular regions is isolated from the rest of the dataset, owing to its location close to an inaccessible region. We stress that one should not rely too much on any fixed data-partitioning

scheme / choice of folds except if having strong reasons to do so. Indeed, when looking at the median (over folds) of the within-fold mean squared cross-validation residual Fig. 5.4, the model comparison changes drastically from the one done using the mean over folds. The median over fold squared residuals ranks the *domains* model as the best, whereas the mean ranks it as the worst. This suggests that the bottom-left fold exhibits an outlier-like behavior, having a too strong influence over the CV mean. Figure 5.5 shows how the within-fold mean squared residuals vary over the different folds.

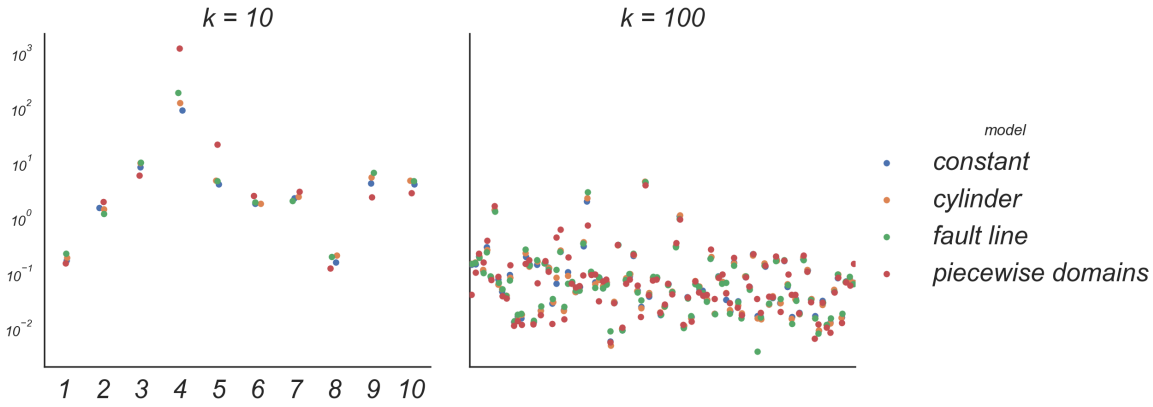


Figure 5.5: Distribution of the MS CV residual over folds ($k = 10$ and $k = 100$) for the different models.

One notices that in 10-fold cross-validation fold nr. 4 is sensibly harder to predict across all model, the *piecewise domains* model being even wildly wrong for that particular fold. The figure also shows how increasing the number of folds (still defined with kMeans clustering) makes each individual fold easier to predict and uniformizes the residuals across folds.

The above figure presents strong evidence that one should be careful in the choice of folds when using cross-validation. It calls for the development of CV diagnostics that enable one to quantify the sensitivity of the chosen CV approach and shows the necessity of elaborating robust cross-validation procedures.

While we leave principled developments to future works, we next present a few heuristic diagnostics that can help to assess the quality of a given cross-validation procedure. Our goal is to present an exhaustive overview of the situation, which we believe can serve as inspiration for further developments in CV research.

Before proceeding further, we also note that the above phenomenon of fold sensitivity can also be of practical interest in the particular inverse problem considered. Indeed, volcanoes tend to have inaccessible regions where no data can be collected. One would thus like to use models that do not depend too strongly on data that could potentially be collected in those regions. Here, cross-validation can be used as an interesting diagnostic for regional sensitivity.

5.6 Cross-Validation for Hyperparameter Training

In the previous sections as well as in most applications of Bayesian inversion, models hyperparameters were trained using maximum likelihood estimates (MLE). While MLE for hyperparameters training is well established and enjoys theoretical guarantees, it suffers from several drawbacks in practice. In particular, it is known that MLE tends to overestimate the range parameter and that it suffers from numerical instabilities (Basak et al., 2022). In this regard, cross-validation can offer an alternative to MLE for hyperparameters estimation.

The usual procedure for training hyperparameters with CV is to pick the set of hyperparameters that minimize the across-fold mean MSE Eq. (5.18). While this constitutes a reasonable training scheme, results from last section have shown how some folds can have outlier-type behaviors for some trend models. By that we mean that, for a given choice of folds, some trend models might be highly sensitive to one given fold, leading to a big MSE for that fold. For example, in our gravimetric inverse problem example, the piecewise domains trend model has a hard time predicting two of the folds (see Fig. 5.4), probably because the estimation of the trend within one domain depends highly on those regions of the dataset. Even though one can wonder whether models with such high sensitivity to parts of the data are appropriate, we leave this question for later and here focus solely on hyperparameters training.

One possible way to alleviate the sensitivity of the training on "outlier folds" is to consider the across-fold *median* MSE instead of the mean. In the context of Eq. (5.18), the *median k-fold CV criterion* writes as:

$$\hat{\boldsymbol{\theta}}_{\mathrm{CV}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CV,median}}(\boldsymbol{\theta}; Z, \boldsymbol{y}, \boldsymbol{I}) \tag{5.19}$$

$$\mathcal{L}_{\mathrm{CV,median}}(\boldsymbol{\theta}, Z, \boldsymbol{y}, \boldsymbol{I}) := \mathrm{median}\left(\left(\frac{1}{|\boldsymbol{i}_j|} ||\boldsymbol{e}_{\boldsymbol{i}_j}||^2 \right)_{=1,\dots,k} \right), \tag{5.20}$$

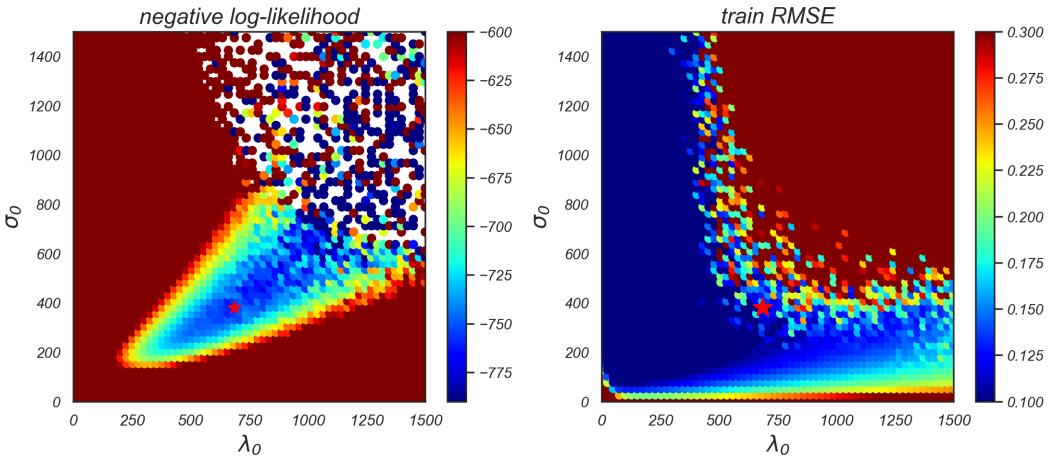
Note that while considering the median is one possible way of mitigating the dependence on outliers, one could also consider other techniques, for example replacing the L_2 norm of the residuals the L_1 norm.

We study the difference between MLE and CV hyperparameters training by applying them to our example gravimetric inverse problem (see Section 4.4.1 for MLE training without trends). Table 5.1 lists the optimal hyperparameters for each training scheme, for different trend models (Matérn 3/2 kernel). The CV training is done with 10 folds, with folds defined by kMeans spatial clustering over the datapoints (same folds as in Fig. 5.4).

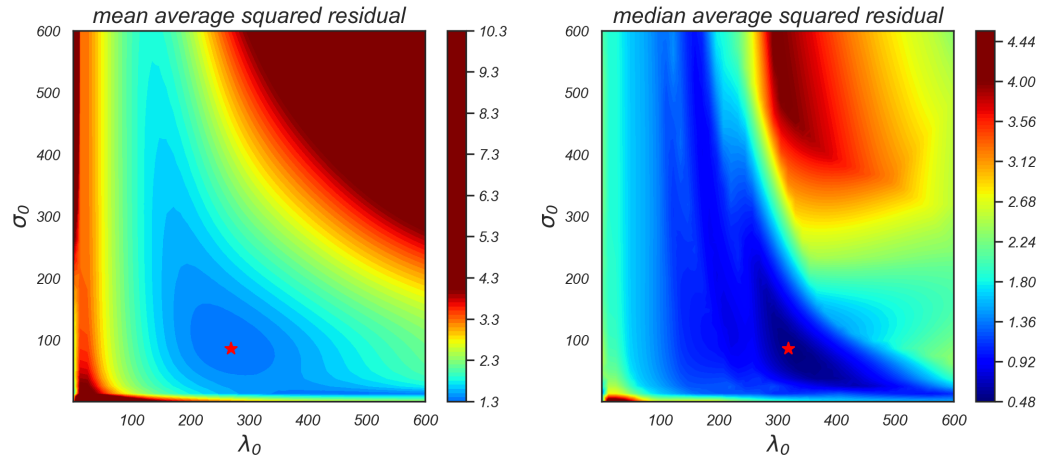
	λ_0		σ_0	
training method	MLE	CV (10 fold)	MLE	CV (10 fold)
constant	651.6	467.3	284.7	49.8
chimney	923.7	574.9	318.8	86.5
fault line	1231.3	790.1	257.7	49.8
domains	539.2	287.9	147.6	245.4

Table 5.1: Optimal hyperparameters (Matérn 3/2) for MLE and 10-fold CV training.

One notices that CV tends to favor noticeably smaller lenghtscale parameters and variances. Also, compared to MLE, CV training does not present any numerical instabilities. Indeed, the log-determinant term in the MLE criterion (Eq. (4.7)) is unstable for large values of λ_0. In practice, one discards unstable regions of the parameter space by excluding hyperparameters for which the prediction error on the training set is significantly larger than the noise standard deviation (see Fig. 5.6).



(a) maximum likelihood



(b) cross-validation

Figure 5.6: Comparison of MLE and 10-fold CV hyperparameters training. Optimal hyperparameters for each training scheme are marked with a red cross. Regions where the log-determinant did not compute are left blank.

5.7 Appendix: Proofs of the Theorems

Proof. (Theorem 14) We begin by noticing that $\boldsymbol{Y}$ and $Z_{\boldsymbol{X}}$ are jointly multivariate Gaussian distributed:

$$\begin{pmatrix} \boldsymbol{Y} \\ Z_{\boldsymbol{X}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

with $\mu_1 := GF_{\boldsymbol{W}}\boldsymbol{\beta}_{\text{prior}}$, $\mu_2 := F_{\boldsymbol{X}}\boldsymbol{\beta}_{\text{prior}}$, $\Sigma_{22} := Q$ and $\Sigma_{12} := GK_{\boldsymbol{WX}} + GF_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{X}}^T$. Then, applying the usual Gaussian conditioning formulae (see e.g. (Rasmussen and Williams, 2006, A.2)) yields the desired result for the posterior mean and covariance of Z. The exact same reasoning gives the posterior distribution of $\boldsymbol{\beta}$. □

For the flat prior limit, we first apply the Woodbury identity to the helper matrix Q to get:

$$Q^{-1} = R^{-1} - R^{-1}GF_{\boldsymbol{W}}\left(\boldsymbol{\Sigma}^{-1} + F_{\boldsymbol{W}}^T G^T R^{-1} G F_{\boldsymbol{W}}\right)^{-1} F_{\boldsymbol{W}}^T G^T R^{-1}.$$

Then, in the limit where all eigenvalues of the trend coefficients prior covariance $\boldsymbol{\Sigma}$ tend to 0, applying the matrix identity (where the matrices A, B are assumed invertible):

$$\left(A^{-1} + B^{-1}\right)^{-1} = A - A\left(A + B\right)^{-1} A,$$

and defining $S := F_{\boldsymbol{W}}^T G^T R^{-1} G F_{\boldsymbol{W}}$, the product involved in the computation of the posterior mean tends to:

$$\begin{aligned}
\left(F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T + K_{\boldsymbol{XW}}\right) G^T Q^{-1} &\overset{\boldsymbol{\Sigma}\to 0}{\longrightarrow} F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T G^T R^{-1} - F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T G^T R^{-1} G F_{\boldsymbol{W}} \left(\boldsymbol{\Sigma}^{-1} + S\right) F_{\boldsymbol{W}}^T G^T R^{-1} \\
&\quad + K_{\boldsymbol{XW}} G^T R^{-1} - K_{\boldsymbol{XW}} G^T R^{-1} G F_{\boldsymbol{W}} S^{-1} \\
&= F_{\boldsymbol{W}}\boldsymbol{\Sigma}F_{\boldsymbol{W}}^T G^T R^{-1} \\
&\quad + K_{\boldsymbol{XW}} G^T R^{-1} - K_{\boldsymbol{XW}} G^T R^{-1} G F_{\boldsymbol{W}} S^{-1} \\
&\quad - F_{\boldsymbol{W}}\boldsymbol{\Sigma}S\left(S^{-1} - S^{-1}\left(S^{-1} + \boldsymbol{\Sigma}\right)^{-1} S^{-1}\right) F_{\boldsymbol{W}}^T G^T R^{-1} \\
&\to F_{\boldsymbol{W}} S^{-1} F_{\boldsymbol{W}}^T G^T R^{-1} + K_{\boldsymbol{XW}} G^T R^{-1} - K_{\boldsymbol{XW}} G^T R^{-1} G F_{\boldsymbol{W}} S^{-1},
\end{aligned}$$

which yields the desired result. A similar calculation give the covariance in the flat prior limit.

Chapter 6

Multivariate Bayesian Inversion and Sequential Uncertainty Reduction

This chapter reproduces the paper Fossum et al. (2021b), co-authored with Trygve Olav Fossum, Jo Eidsvik, David Ginsbourger and Kanna Rajan and published in the Annals of Applied Statistics
(DOI:10.1214/21-AOAS1451). with a few reformulations and corrections.

6.1 Introduction

After having considered extensions of Gaussian processes to linear operator observations and large-scale Bayesian inverse problems in Chapter 4, we now pursue another direction by extending GPs to vector-valued responses. Although GP regression in a multivariate setting is already known in the geostatistics community under the name of cokriging (see for example (Wackernagel, 2003)), the use of Gaussian processes for multivariate inverse problems is still a fresh topic of inquiry. We note that there already exists theoretical works on the topic that apply to vector-valued responses as well (Knapik et al., 2011; Dashti and Stuart, 2016), but these lack practical applications.

In this chapter, we focus on the Bayesian estimation of excursion set of vector-valued fields. To that end, we extend uncertainty reduction criteria to multivariate GPs and provide semi-analytical expressions for their computation. We then turn to the sequential optimization of these criteria and provide algorithms for myopic as well as lookahead optimization. Related works in probabilistic estimation of excursion sets include (Duhamel et al., 2023) and (Fossum et al., 2021a). We also note that, during the elaboration of this thesis, new results concerning the consistency of SUR strategies in multivariate settings have been derived (Stange, 2022).

Our techniques are demonstrated on a river plume mapping problem as presented in Section 2.4.2, performing synthetic benchmarks as well as a real-world field campaign. Our benchmarks demonstrate how adaptive sampling strategies are able to beat traditional static predefined data collection plans, opening new venues of research in oceanography that leverage fruitful interactions between marine robotics and spatial statistics.

6.2 Multivariate Gaussian Processes and UQ on their Excursion Sets

In order to model multivariate fields such as temperature and salinity in a fjord, traditional GP models are insufficient. We thus here introduce a multivariate generalization of GPs. Although kriging for multiple outputs has been known for some time under the name of cokriging (Journel and Huijbregts, 1978; Cressie, 1993; Ver Hoef and Barry, 1998; Wackernagel, 2003) and has been recently revisited in machine learning under the umbrella of multiple output Gaussian processes (Conti and O'Hagan, 2010; Álvarez et al., 2012), most works forego a systematic presentation of the subject. For the sake of completeness, we here start from scratch by extending the definition of GPs to multivariate outputs in a straightforward fashion.

Definition 11 (Multivariate Gaussian Process)**.** A p-dimensional Gaussian process Z on a domain D is an $\mathbb{R}^p$-valued stochastic process, such that for any finite set of locations $x_1, ..., x_n \in D$ and any set of indices $i_1, ..., i_p$ the distribution of the vector

$$(Z_{x_1,i_1}, ..., Z_{x_p,i_p})$$

is n-variate Gaussian.

Here, we use $Z_{x,i}$ to denote the i-th component of Z_x ($1 \leq i \leq p$), specifying the spatial location first, followed by the vector index. For the rest of this work, Greek subscripts will be used to denote vector indices.

Example 1 (Uncorrelated Basis Functions)**.** Let $V_i \sim \mathrm{Gp}(m^{(i)}, k^{(i)}), i = 1, ..., p$ be a set of independent Gaussian processes on D and let $a_i : D \to \mathbb{R}$ be a collection of real-valued functions. Then, the stochastic process

$$Z := \sum_{i=1}^{p} a_i V_i$$

is a p-variate Gaussian process.

Generalized Locations: To simplify notation, we introduce the concept of *generalized locations*, which stands for a couple $\chi = (x, i)$ of spatial location x and vector index i. The notation Z_χ will be used to denote $Z_{x,i}$ and will allow us to think of Z as a scalar-valued random field indexed by $D \times \{1 \dots, p\}$, which will give the co-Kriging equations a particularly simple form that parallels the one of univariate Kriging. From now on, Greek letters will be used to denote generalized locations and the letters x and i will usually denote spatial locations and response indices respectively.
Furthermore, boldface letters (and uppercase in the case of latin letters) will be used to denote concatenated quantities corresponding to batches of observations. Given a dataset consisting of q observations at spatial locations $(x_1, \dots, x_q) \in D^q$ and response indices $(i_1, \dots, i_q) \in \{1, ..., p\}^q$, we use the concatenated notation

$$\boldsymbol{\chi} := (\chi_1, \dots, \chi_q), \text{ with } x_i = (x_i, i_i).$$

We also compactly denote the field values at those different locations by

$$Z_{\boldsymbol{\chi}} := \left(Z_{x_1,i_1}, ..., Z_{x_q,i_q}\right) \in \mathbb{R}^q.$$

For a second order random field Z on D with mean function $\boldsymbol{m}_x$ and matrix covariance function $\boldsymbol{k}(x, x')$, one can straightforwardly extend and vectorize the mean function into a function of the batch-generalized locations $\boldsymbol{m}_{\boldsymbol{\chi}}$. As for $\boldsymbol{k}$, it induces a covariance kernel $\boldsymbol{K}$ on the set of extended locations via $\boldsymbol{K}((x,i),(x',i')) = \boldsymbol{K}(x,x')_{i,i'}$. In vectorized/batch form, $\boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}'}$ then amounts to a matrix with numbers of lines and columns equal to the numbers of generalized locations in $\boldsymbol{\chi}$ and $\boldsymbol{\chi}'$, respectively. For a collection of spatial locations $\boldsymbol{X} = (x_1, \ldots, x_r) \in D^r$, we write $Z_{\boldsymbol{X}} := (Z_{x_1,1}, ..., Z_{x_1,p}, ..., Z_{x_r,1}, ..., Z_{x_r,p}) \in \mathbb{R}^{p\times r}$, concatenating first the vector indices and then the spatial locations. Such vectorized quantities turn out to be useful in order to arrive at simple expressions for the co-Kriging equations presented next.

6.2.1 Cokriging and Update

Given a GRF Z and observations of some of its components at locations in the domain, one can predict the value of the field at some unobserved location $x \in D$ by using the conditional mean of Z_x, conditional on the data. This coincides with co-Kriging equations, which tell us precisely how to compute conditional means and covariances. We will present a general form of co-Kriging, in the sense that it allows inclusion of several (batch) observations at a time; observations at a given location $x \in D$ may only include a subset of the components of $Z_x \in \mathbb{R}^p$ (heterotopic).

Using generalized locations, the simple cokriging equations for generalized observations then amount to kriging with respect to a scalar-valued GP indexed by $D \times \{1 \ldots, p\}$.

Theorem 16. *Let Z be a p-variate GP on D with mean function $\boldsymbol{m}$ and (generalized) covariance function $\boldsymbol{K}$. Assume that one observes the following q-dimensional batch generalized data:*

$$\boldsymbol{Y} := Z_{\boldsymbol{\xi}} + \boldsymbol{\epsilon}, \tag{6.1}$$

where $\boldsymbol{\epsilon}$ is a q-dimensional centered Gaussian vector with covariance matrix $\boldsymbol{\Delta}$ and $\boldsymbol{\xi}$ is a batch of generalized locations. Then, conditionally on $\boldsymbol{Y} = \boldsymbol{y}$, and assuming that $(\boldsymbol{K}_{\boldsymbol{\xi}\boldsymbol{\xi}} + \boldsymbol{\Delta})$ is invertible, the distribution of Z is Gaussian with mean function and (generalized) covariance function:

$$\tilde{\boldsymbol{m}}_{\boldsymbol{\chi}} = \boldsymbol{m}_{\boldsymbol{\chi}} + \boldsymbol{\lambda}(\boldsymbol{\chi})^T(\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{\xi}}) \tag{6.2}$$

$$\tilde{\boldsymbol{K}}_{\boldsymbol{\chi}\boldsymbol{\chi}'} = \boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}'} - \boldsymbol{\lambda}(\boldsymbol{\chi})^T(\boldsymbol{K}_{\boldsymbol{\xi}\boldsymbol{\xi}} + \boldsymbol{\Delta})\boldsymbol{\lambda}(\boldsymbol{\chi}'), \tag{6.3}$$

with (generalized) cokriging weights

$$\boldsymbol{\lambda}(\boldsymbol{\chi}) = (\boldsymbol{K}_{\boldsymbol{\xi}\boldsymbol{\xi}} + \boldsymbol{\Delta})^{-1}\boldsymbol{K}_{\boldsymbol{\xi}\boldsymbol{\chi}}. \tag{6.4}$$

In the following applications, we will always silently assume that invertibility assumptions are met. This is reasonable as long as the observations are not too correlated (many observations of the same component at the same location). Next, we consider a sequential setting similar as that of Section 4.2 where new (batches

of) observations arrive in sequence and one wants to update the posterior in a computationally efficient way, without having to re-assimilate the data from scratch using the bare conditioning equations Eqs. (6.2) and (6.3). We assume that n batches of observations have already been assimilated, so that the current posterior has mean function $\boldsymbol{m}^{(n)}$ and generalized covariance $\boldsymbol{K}^{(n)}$. Then, a new batch of generalized observations made at $\boldsymbol{\xi}_{n+1}$ with observed values $\boldsymbol{y}_{n+1}$ is made available. Then, extending Corollary 6 to the multivariate case, one obtains the analogon of Theorem 13 that allows to update the posterior by only computing a subset of the cokriging weights.

Theorem 17. *Assume that $Z^{(n)}$ is a multivariate GP with mean function $\boldsymbol{m}^{(n)}$ and (generalized) covariance function $\boldsymbol{K}^{(n)}$. Then, given a new batch of q_n observations made at generalized locations $\boldsymbol{\xi}_{n+1}$ with values $\boldsymbol{y}_{n+1}$ and following observations model Eq.* (6.1)*, the posterior of $Z^{(n)}$ is multivariate Gaussian with mean function and generalized covariance given by*

$$\boldsymbol{m}_{\boldsymbol{\chi}}^{(n+1)} = \boldsymbol{m}_{\boldsymbol{\chi}} + \boldsymbol{\lambda}_{n+1}(\boldsymbol{\chi})^T(\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{\xi}}^{(n)}) \tag{6.5}$$

$$\boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}'}^{(n+1)} = \boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}'}^{(n)} - \boldsymbol{\lambda}_{n+1}(\boldsymbol{\chi})^T(\boldsymbol{K}_{\boldsymbol{\xi}_{n+1}\boldsymbol{\xi}_{n+1}}^{(n)} + \boldsymbol{\Delta})\boldsymbol{\lambda}_{n+1}(\boldsymbol{\chi}'). \tag{6.6}$$

Here $\boldsymbol{\lambda}_{n+1}(\boldsymbol{\chi})$ denotes cokriging weights Eq. (6.4) *when conditioning the field $Z^{(n)}$ on the new data $\boldsymbol{y}_{n+1}$.*

These update formulae are essentially a multivariate extension of the batch-sequential Kriging update formulae from (Chevalier et al., 2014b) As noted in (Chevalier et al., 2015) in the case of scalar-valued fields, these update formulae naturally extend to universal Kriging in second-order settings and apply without Gaussian assumptions. Apart from offering computational savings, these formulae We will enable us to derive semi-analytical expressions for step-wise uncertainty reduction criteria for vector-valued random fields.

6.2.2 Multivariate Excursion Sets and UQ

We now turn to the estimation of excursion sets of multivariate random fields. For the rest of this section, we assume that Z is a p-variate GP on a domain D. We are interested in predicting regions where the values of the field lie in a certain range, i.e. sets of the form

$$\Gamma := Z^{-1}(T) = \{x \in D : Z_x \in T\}, \tag{6.7}$$

where $T \subset \mathbb{R}^p$ is some set of specified values. If we assume that Z has continuous trajectories and T is closed, then Γ becomes a Random Closed Set (Molchanov, 2005) and concepts from the theory of random sets will prove useful to study Γ. Note that while some aspects of the developed approaches do not call for a specific form of T, we will often, for purposes of simplicity, stay with the case of orthants: $T = ((-\infty, t_1] \times \cdots \times (-\infty, t_p])$ where $t_1, \ldots, t_p \in \mathbb{R}$, as this allows for efficient calculation of several key quantities. Note that changing some $\leq$ inequalities to $\geq$ ones would lead to immediate adaptations. Figure 6.1 shows a realization of a bivariate GP together with the corresponding excursion set above two arbitrary thresholds.

Apart from the precise location of the excursion set Γ, one might consider the (less ambitious) objective of estimating its volume. In the following, let ν be a

(locally finite, Borel) measure on D. Our goal is to investigate the distribution of the (random) excursion volume $\nu(\Gamma)$. Centered moments of this distribution may be computed using the following theorem (we refer the reader to the section on *generalized locations* for a reminder of the notation).

Theorem 18. *Let Z be a measurable stochastic process defined on a domain D and ν be a locally finite measure on D. Then, $\nu(\Gamma)$ is a random variable and for any $r \geq 1$,*

$$\mathbb{E}[\nu(\Gamma)^r] = \int_{D^r} \mathbb{P}\left(Z_{\boldsymbol{X}} \in T^r\right) \mathrm{d}\nu^{\otimes}\left(\boldsymbol{X}\right),$$

where the product measure is denoted as $\nu^{\otimes} := \bigotimes_{i=1}^{r} \nu$, T^r denotes the Cartesian product and $\boldsymbol{X} \in D^r$.

Furthermore, in the case where Z is a p-variate GP with mean function $\boldsymbol{m}$ and (generalized) covariance function $\boldsymbol{K}$, for any $\boldsymbol{X} \in D^r$ the probability of r-joint excursion can be computed with the help of the $p \times r$-dimensional Gaussian probability density function $\varphi_{p\times r}(\cdot;\ \boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{XX}})$ as

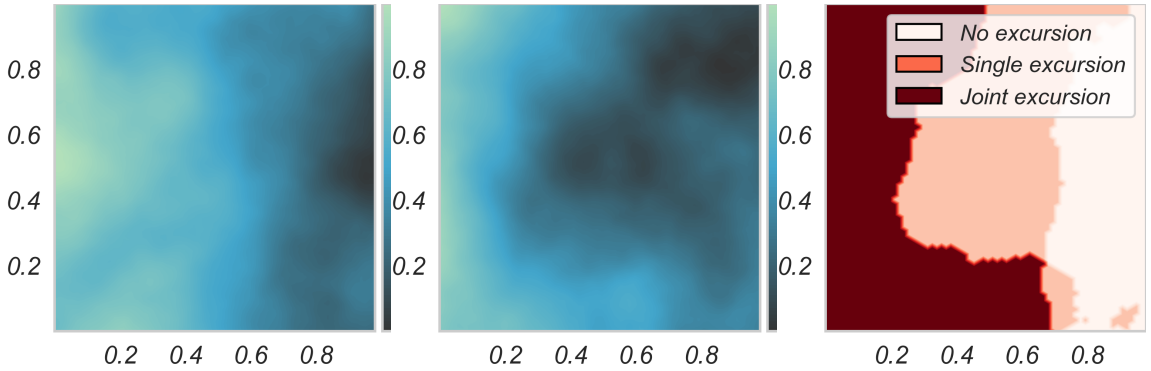
$$\mathbb{P}\left(Z_{\boldsymbol{X}} \in T^r\right) = \int_{T^r} \varphi_{p\times r}\left(\boldsymbol{t};\ \boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{XX}}\right) \mathrm{d}\boldsymbol{t},$$

where $\boldsymbol{m}_{\boldsymbol{X}} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{K}_{\boldsymbol{XX}} \in \mathbb{R}^{(p\times r)\times(p\times r)}$ is assumed to be non-singular.

Corollary 9. *In the case where the excursion range is an orthant, $\boldsymbol{t} = (t_1, ..., t_p) \in \mathbb{R}^p$ and $T = ((-\infty, t_1] \times \cdots \times (-\infty, t_p])$, the joint excursion probability directly writes in terms of the multivariate Gaussian cumulative distribution:*

$$\mathbb{P}\left(Z_{\boldsymbol{X}} \in T^r\right) = \Phi_{p\times r}\left(1_r \otimes \boldsymbol{t};\ \boldsymbol{m}_{\boldsymbol{X}}, \boldsymbol{K}_{\boldsymbol{XX}}\right),$$

where $\Phi_{p\times r}$ denotes the $p \times r$-variate Gaussian cumulative distribution function (CDF) and we use the notation $1_r = (1, ..., 1) \in \mathbb{R}^r$ and $1_r \otimes \boldsymbol{t} = (t_1, \ldots, t_p, \ldots, t_1, \ldots, t_p) \in \mathbb{R}^{p\times r}$.



(a) $Z_{x,1}$. (b) $Z_{x,2}$. (c) Regions of interest.

Figure 6.1: Realization of a bivariate Gaussian process Z (first component (a) and second computational (b)) and excursion set above some threshold (c). Joint excursion in red and excursion of a single variable in light-red.

A natural way of quantifying the uncertainty on the excursion volume is through its (residual) variance (Chevalier et al., 2014a). This is the goal of the *excursion measure variance* (EMV), which is the second centered moment of the excursion measure distribution.

Definition 12 (Excursion Measure Variance)**.** Given a p-variate GP Z defined on a domain D, and a target range $T \subset \mathbb{R}^p$, the **excursion measure variance** of the excursion set of Z in T is defined as:

$$\begin{aligned}\mathrm{EMV}(Z) := \mathrm{Var}[\nu(\Gamma)] = &\int_{D^2} \mathbb{P}\left(Z_x \in T, Z_{x'} \in T\right) d\nu^{\otimes}(x, x') \\ &- \left(\int_D \mathbb{P}\left(Z_x \in T\right) d\nu(x)\right)^2,\end{aligned}$$

where the random set Γ is defined as in Eq. (6.7) and the dependence on T is left silent for the sake of readability.

In the orthant case of Corollary 9 and denoting as usual by $\boldsymbol{m}$ and $\boldsymbol{K}$ the mean and covariance function of Z, the above can be computed semi-analytically as

$$\begin{aligned}\mathrm{EMV}(Z) = &\int_{D\times D} \Phi_{2p}\left(1_2 \otimes \boldsymbol{t}; \boldsymbol{m}_{[x,x']}, \boldsymbol{K}_{[x,x'][x,x']}\right) \, \mathrm{d}\nu^{\otimes}(x, x') \\ &- \left(\int_D \Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x, \boldsymbol{K}_{xx}\right) d\nu(x)\right)^2,\end{aligned}$$

where the brackets denote concatenation. We note that in practice, the above integral can be efficiently evaluated numerically using (Genz and Bretz, 2009), but still require integration over the product domain $D \times D$. In contrast, the integrated Bernoulli variance (IBV) of Bect et al. (2019) involves solely an integral over the domain.

Definition 13 (Integrated Bernoulli Variance)**.** Given the same setting as in Definition 12, the **integrated Bernoulli variance** of the excursion of Z in T is defined as:

$$\mathrm{IBV}(Z) := \int_D \mathbb{P}\left(Z_x \in T\right)\left(1 - \mathbb{P}\left(Z_x \in T\right)\right) d\nu(x).$$

This functional also provides a natural way of measuring the uncertainty in the excursion volume, since it is equal to the integral of the variance of the pointwise excursion indicator function $\mathrm{IBV} = \int_D \mathrm{Var}[\mathbb{1}(Z_x \in T)] d\nu(x)$. Again, in the orthant case of Corollary 9, the above can be expanded as:

$$\mathrm{IBV}(Z) = \int_D \Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x, \boldsymbol{K}_{xx}\right) - \left(\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x, \boldsymbol{K}_{xx}\right)\right)^2 \mathrm{d}\nu(x).$$

6.2.3 SUR Strategies for Multivariate Excursion Sets

We now focus on the engineering of stepwise uncertainty reduction strategies (SUR) in the vein of (Bect et al., 2012) for the estimation of excursion sets of multivariate GPs. To that end, we consider the same sequential setting as in Theorem 17 and denote by $Z^{(n)}$ a GP that is distributed according to the conditional law of the

original GP Z conditional on n stages of already performed assimilations. The introduction of the stochastic process Z is useful in simplifying notations, but we stress that in the end all quantities only depend on the conditional law and not on the process itself. Our goal is to compute the expected effect of the inclusion of new observations on the uncertainty functionals EMV and IBV, thereby extending results from (Chevalier et al., 2014a; Bect et al., 2019) to the multivariate setting.

Remark 9. The uncertainty measures $\mathrm{EMV}(Z)$ and $\mathrm{IBV}(Z)$ (Definitions 12 and 13) only depend on the law of the GP Z in the sense that if Z' is another GP that is equal in law to Z, then $\mathrm{EMV}(Z) = \mathrm{EMV}(Z')$ and $\mathrm{IBV}(Z) = \mathrm{IBV}(Z')$. Owing to this fact, the notation $\mathrm{IBV}(Z^{(n)})$, where $Z^{(n)}$ is any GP having as law the current conditional law of Z, is well-defined, and we will liberally use it next.

In order to study the effect of the inclusion of a new data point, we denote by $\mathrm{IBV}(Z^{(n)}|Z^{(n)}_{\boldsymbol{\chi}})$ the IBV under the current law (the law of $Z^{(n)}$), conditioned on observing $Z^{(n)}_{\boldsymbol{\chi}}$ (generalized, possibly batch observation) at generalized location $\boldsymbol{\chi}$. Our goal is then to compute the expected effect of a new observation on the IBV:

$$\mathrm{EIBV}(\boldsymbol{\chi}; Z^{(n)}) := \mathbb{E}_{Z^{(n)}}\left[\mathrm{IBV}(Z^{(n)}|Z^{(n)}_{\boldsymbol{\chi}})\right], \tag{6.8}$$

where $\boldsymbol{\chi}$ is any (batch) generalized location.

We next present a result that allows efficient computation of EIBV as an integral of CDFs of the multivariate Gaussian distribution. This will prove useful when designing sequential expected uncertainty reduction strategies. For the next two propositions, assume that the target range is an orthant as in Corollary 9 with upper threshold $\boldsymbol{t} \in \mathbb{R}^p$.

Proposition 1. *Let the current law of the field be that of the p variate GP $Z^{(n)}$ with mean function $\boldsymbol{m}^{(n)}$ and covariance $\boldsymbol{K}^{(n)}$. Also let $\boldsymbol{\chi}$ be any batch-generalized location and denote by $\boldsymbol{K}^{(n+1)}$ the conditional covariance functions of $Z^{(n)}$ conditionally on observations at $\boldsymbol{\chi}$ (note that according to Eq. (6.3) this does not depend on the values of the observations). Then the expected IBV at $\boldsymbol{\chi}$ can be computed as:*

$$\begin{aligned}\mathrm{EIBV}(\boldsymbol{\chi}; Z^{(n)}) = &\int_D \Phi_p\left(\boldsymbol{t}; \boldsymbol{m}^{(n)}_x, \boldsymbol{K}^{(n)}_{xx}\right) \mathrm{d}\nu(x) \\ &- \int_D \Phi_{2p}\left(\begin{pmatrix}\boldsymbol{t} - \boldsymbol{m}^{(n)}_x \\ \boldsymbol{t} - \boldsymbol{m}^{(n)}_x\end{pmatrix}; \boldsymbol{\Sigma}^{(n)}(x)\right) \mathrm{d}\nu(x),\end{aligned} \tag{6.9}$$

where the $2p \times 2p$ matrix $\boldsymbol{\Sigma}^{(n)}(x)$ is given by:

$$\boldsymbol{\Sigma}^{(n)}(x) := \begin{pmatrix}\boldsymbol{K}^{(n)}_{xx} & \boldsymbol{K}^{(n)}_{xx} - \boldsymbol{K}^{(n+1)}_{xx} \\ \boldsymbol{K}^{(n)}_{xx} - \boldsymbol{K}^{(n+1)}_{xx} & \boldsymbol{K}^{(n)}_{xx}\end{pmatrix},$$

A similar semi-analytical expression can also be derived for the expected EMV, which is defined in the same way as the expected IBV.

Proposition 2. *Consider the same setting as in Proposition 1, then the expected EMV at the batch-generalized location $\boldsymbol{\chi}$ can be computed as:*

$$\begin{aligned}\mathrm{EEMV}(\boldsymbol{\chi}; Z^{(n)}) = &\int_D \Phi_{2p}\left(1_2 \otimes \boldsymbol{t}; \boldsymbol{m}_{[x,x']}, \boldsymbol{K}_{[x,x'][x,x']}\right) \mathrm{d}\nu^{\otimes}(x, x') \\ &- \int_D \Phi_{2p}\left(\begin{pmatrix}\boldsymbol{t} - \boldsymbol{m}^{(n)}_x \\ \boldsymbol{t} - \boldsymbol{m}^{(n)}_{x'}\end{pmatrix}; \tilde{\boldsymbol{\Sigma}}^{(n)}(x, x')\right) \mathrm{d}\nu^{\otimes}(x, x'),\end{aligned}$$

where the matrix $\tilde{\boldsymbol{\Sigma}}^{(n)}(x, x')$ *is defined blockwise as*

$$\tilde{\boldsymbol{\Sigma}}^{(n)}(x, x') = \begin{pmatrix} \tilde{\Sigma}_{1,1}(x, x) & \tilde{\Sigma}_{1,2}(x, x') \\ \tilde{\Sigma}_{2,1}(x', x) & \tilde{\Sigma}_{2,2}(x', x') \end{pmatrix}$$

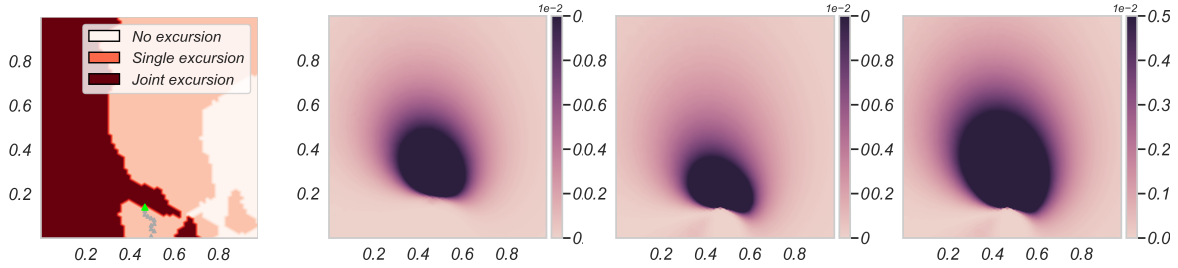
with blocks given, for $i, j \in \{1, 2\}$ *and* $x, x' \in D$*, by*

$$\tilde{\Sigma}_{i,j}(x, x') = \boldsymbol{\lambda}_{n+1}(x)^T \boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}}^{(n)} \boldsymbol{\lambda}_{n+1}(x') + \delta_{i,j} \boldsymbol{K}_{xx'}^{(n+1)}.$$

Here we use the obvious convention that the cokriging weights for the spatial location $x \in D$ *are given by Eq.* (6.4) *for the generalized location* $((x, 1), ..., (x, p))$*, i.e. the weights for all components of the field at location* x*.*

We remark that Propositions 1 and 2 are twofold generalizations of results from Chevalier et al. (2014a): they extend previous results to the multivariate setting and also allow for the inclusion of batch or heterotopic observations through the concept of generalized locations. A key element for understanding these propositions is that the conditional co-Kriging mean entering in the EPs depend linearly on (batch) observations. The conditional equality expressions thus become linear combinations of Gaussian variables whose mean and covariance are easily calculated. Related closed-form solutions have been noted in similar contexts (Bhattacharjya et al., 2013; Stroh, 2018), but not generalized to our situation with random sets for multivariate GPs.

Figure 6.2 displays of the EBV is reduced depending on which component of a bivariate GP is observed. In the situation of the figure, a first set of observations are done at the locations depicted in gray (see Fig. 6.2a), and the data is used to update the GP model. We then consider the green triangle as a potential next observation location and plot the EBV reduction (at each location) that would result from observing only one component of the field ($Z_{x,1}$ or $Z_{x,2}$), or both at location x.



(a) Regions. (b) $Z_{x,1}$ observed. (c) $Z_{x,2}$ observed. (d) Both observed.

Figure 6.2: Pointwise Bernoulli variance reduction for observation of a single or both components of the random field at one location. Data collection locations in green. True excursion set in red. Places where only one response is above threshold are depicted in pink. EBV reduction associated to observing one or both responses at the green location are shown in 6.2b, 6.2c and 6.2d.

6.3 Application: Sequential Design for Multivariate Excursion Set Estimation

Now that we have uncertainty functionals for quantifying the reduction of uncertainty on excursion sets of multivariate GPs that would result from adding obser-

vations at a given (batch of) location, we want to use these to develop sequential design strategies for reducing the uncertainty on said excursion sets. This section will be guided by the river plume estimation example from Section 2.4.2.
For the rest of this section, we will assume that Z is a p-variate GP on a domain D and that n data collection steps have been performed so that the current conditional distribution of the GP is the same as that of a p-variate GP $Z^{(n)}$ having mean function $\boldsymbol{m}^{(n)}$ and generalized covariance function $\boldsymbol{K}^{(n)}$. Our goal is to estimate the excursion set Eq. (6.7) for which the value of the GP lie in the pre-specified range $T \subset \mathbb{R}^p$.
The data collection process is performed in a sequential way, so that at stage n one selects the next observation location among a set of candidates using some criterion and then proceeds to gather data at the chosen location. Once the data has been gathered, one updates the model with the observed data values using Eqs. (6.2) and (6.3) to get a new GP $Z^{(n+1)}$. This process is then repeated iteratively.
Note that the type of data collected at each stage can be of various type (all components of the field at a single location, only some components at a subset of selected locations, etc.) because of the concept of *generalized location* in the co-Kriging expressions. In general, a design strategy must choose the spatial location as well as the components to observe (heterotopic), or where several observations are allowed at each stage (batch). In this work, we will limit ourselves to the case where one observation location has to be chosen at each stage and all components of the field there are then observed (isotopic), since this corresponds to the limitations encountered in our river plume mapping example Section 2.4.2. We nevertheless stress that, in principle, the ideas presented next can be applied to any type of sampling situations.

6.3.1 Sampling Strategies

Although, for each criterion, one could write down the Bell equations for the solution of the full dynamic program describing the optimal design problem under the given criterion, in practice this involves a series of intermixed minimizations over designs and integrals over data so that the optimal solution is intractable because of the enormous growth over stages (see e.g. Powell (2016)). We thus resort to heuristic strategies, of which we present a few below.
We assume that at each stage n the next location is chosen from a set of candidate locations $\mathcal{J}$. This set depends on the current location (where the last observation has been performed), but we drop the dependence in the notation for the sake of readability. The next location is then chosen by minimizing a criterion C over the candidates $\mathcal{J}$.

Strategy. [**Naive Greedy Sampling**]

One of the simplest heuristic for adaptive sampling is to select the candidate location that has (current) excursion probability closest to 1/2:

$$C_{\text{naive}}(u) = |\mathbb{P}(Z_u^{(n)} \in T) - 1/2| \tag{6.10}$$

While easy to implement, this strategy does not account for the expected reduction in uncertainty and ignores the effects the locations other that the one observed, leading it to exhibit a too greedy behavior, spending many stages in excursion boundary regions.

In order to go beyond naive sampling, one needs to use more sophisticated uncertainty functionals such as EMV or IBV and take into account the effects of the added observation on the subsequent design stages. For the rest of this section, we will only consider the IBV, tough the dynamic programming approaches presented here can be applied to the EMV as well. The easiest way to account for the future effects of the assimilation of a new observation is to compute its expected effect on the uncertainty functional, under the current distribution of the GP model. This results in the *myopic* sampling strategy.

Strategy. [Myopic]

At each stage n, the location minimizing the expected future uncertainty is chosen:

$$C_{\text{myopic}}(x) = \text{EIBV}\left(u, Z^{(n)}\right). \tag{6.11}$$

This criterion can be efficiently computed using Proposition 1. Even though this myopic strategy is non-anticipatory, it still provides a reasonable approach for creating designs in many applications and is reasonably easy to compute. We also note that this strategy is optimal if there is only one last observation allowed.
The myopic strategy can be extended by considering two stages of measurements, resulting in a *two-step look-ahead* strategy. The principle of is to select as next observation location the one that yields the biggest reduction in EIBV if we were to (optimally) add one more observation after that again.

Strategy. [2-step Look-ahead]

The next observation location is chosen among the minimizers in $\mathcal{J}$ of the criterion

$$C_{\text{2-steps}}(u) = \mathbb{E}_Y\left[\min_{u' \in \mathcal{J}(u)} \text{EIBV}_n\left(u', Z^{(n)} | Z_u^{(n)} = Y\right)\right] \tag{6.12}$$

where Y is the random data realization of $Z_u^{(n)}$ given the observation model Eq. (6.1).

In practice, the outer expectation is computed by Monte Carlo sampling of data Y from the current conditional distribution of the GP. For each sample, the second expectation is then solved using the closed-form expressions for EIBV provided by Proposition 1.

6.3.2 Benchmarks on a Synthetic Test Case

We now study the performances of the above strategies on a synthetic test case that is meant to reproduce the characteristics of a real river plume mapping problem like the one presented in Section 2.4.2. As a reminder, the goal is to estimate the excursion set (ocean)

$$\Gamma^* := \{x \in D : \rho_x^{(1)} \geq T_1, \rho_x^{(2)} \geq T_2\},$$

where $\rho D \to \mathbb{R}^2$ denotes the temperature (first component) and salinity (second component) field in a region D around a river mouth, and T_1, T_2 are prespecified thresholds. The experiments are performed by first generating a bi-variate function on the unit square $\rho : D = [0,1]^2 \to \mathbb{R}^2$ by sampling from some prespecified bivariate GP model. This realization is then used as ground truth to mimic the data

collection process. The data collection is performed by an agent that is allowed to move on a discrete 31×31 rectangular grid over the domain.

GP Model: In the experiments, we use a bi-variate GP model Z that is meant to reproduce the characteristics of a real river plume situation. To that end, the mean function is endowed with a liner trend:

$$\boldsymbol{m}_x = = \boldsymbol{a} + Bx,$$

where $\boldsymbol{a}$ is a two dimensional vector and B a 2×2 matrix. For the covariance part we use a separable covariance model, that is we assume that the covariance can be written as a product of a spatial part and a vector part:

$$\boldsymbol{K}((x,i),(x',j)) = h(x,x')\gamma_{ij},\ \gamma_{ij} = \begin{cases}\sigma_i^2,\ i=j\\ \gamma\sigma_i\sigma_j,\ i\neq j,\end{cases}$$

where h is some covariance kernel on D and $\gamma, \sigma_1, \sigma_2 \in \mathbb{R}$. We note that in theory one could consider non-separable covariance models for multivariate GPs such as (Gneiting et al., 2010; Genton and Kleiber, 2015), but in practice those would require extensive data to fit the model. For the rest of this section, we fix the spatial covariance kernel h to be a Matérn 3/2 kernel with lenghtscale 0.495. The other parameters are set to:

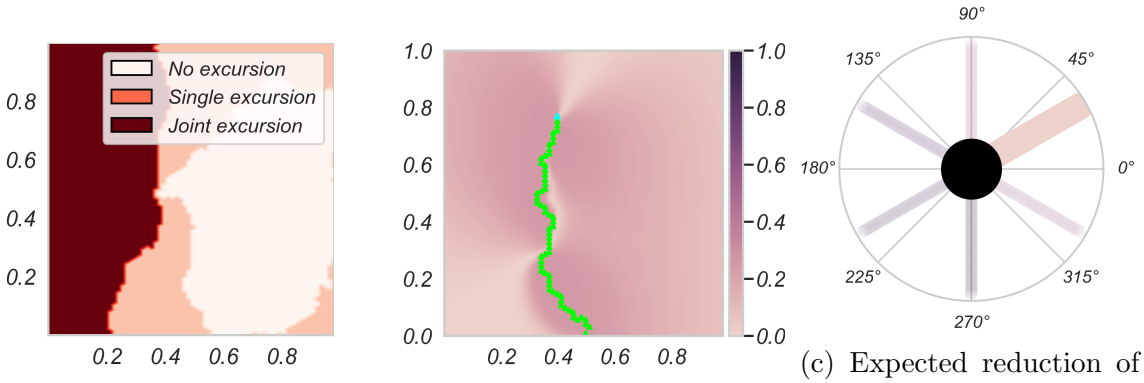
$$\boldsymbol{a} = \begin{pmatrix}5.8\\24.0\end{pmatrix},\ B = \begin{pmatrix}0.0 & -4.0\\0.0 & -3.8\end{pmatrix},\ \sigma_1 = 2.5,\ \sigma_2 = 2.25,\ \gamma = 0.2.$$

Those values are meant to approximate real river plume situations. Figure 6.3 show a realization of the aforementioned GP model, together with a run of the myopic strategy. Plots and experiments are generated with a Python toolbox called MESLAS [1], developed for the needs of this work.

Static Strategies: In order to compare our sampling strategies with the ones usually used in design problems involving autonomous vehicles, we consider static designs, where the data collection plan is pre-scripted. In the experiments, we will consider a *static_north* and *static_east* strategy, which basically travel the domain along the vertical, respectively horizontal middle line. We also consider a *static_zigzag* strategy which travels along the vertical middle line in a zigzag pattern.

Results: The strategies are compared by running them on 100 different ground truth sampled from the bi-variate GP model presented before. For the adaptive strategies, the same GP model is used as the one for sampling, so that we are in a well-specified setting. All strategies are allowed to collect 10 data points. For each strategy, we monitor the decrease in uncertainty on the excursion set (decrease in IBV) as well as the predictive performance on the whole field, measured by root mean square error (RMSE) between the conditional mean and the ground truth, as well as the reduction in predictive variance. It is important to note that the objective function used by the AUV is focused on reducing the EIBV, but we nevertheless expect that we will achieve good predictive performance for criteria such as RMSE as well. Another non-statistical criterion that is relevant for practical purposes is

[1] https://github.com/CedricTravelletti/MESLAS



(a) Excursion realization. (b) BV reduction. (c) Expected reduction of BV.

Figure 6.3: Example run of the myopic strategy on a realization of the given GP model. Reduction in Bernoulli variance compared to the prior is shown in 6.3b, with past observation locations in green and current agent position in cyan. The expected IBV reduction associated to data collection at neighbouring nodes of the current location is shown in 6.3c. The thick and light color indicates the node at 30° to be the best possible choice.

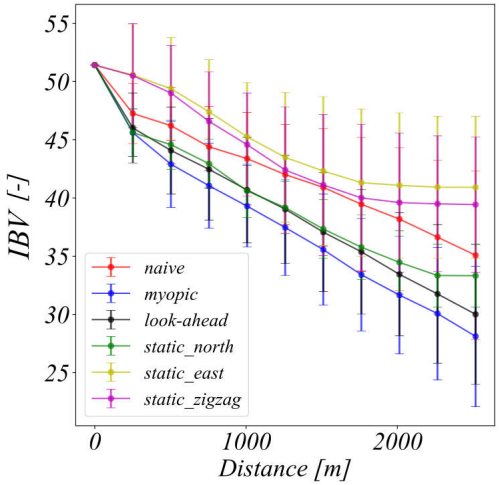
the computational time needed for the strategy. The results of the replicate runs are shown in Fig. 6.4, where the different criteria are plotted as a function of survey distance.

We see that both *myopic* and *look-ahead* strategies perform well here, but some of the *static_east* and *static_zigzag* also achieve good results because they cover large parts of the domain without re-visitation. Sequential strategies targeting IBV will sometimes not reach similar coverage, as interesting data may draw the AUV into twists and turns. There is a relatively large variety in the replicate results as indicated by the vertical lines. Nevertheless, the ordering of strategies is similar. On the computational side, the *naive* strategy is on par with the static designs, while the *myopic* strategy is slower because it evaluates expected values for all candidate directions at the waypoints. But it is still able to do so in reasonable time, which allows for real-world applicability. The *look-ahead* strategy is much slower, reaching levels that are nearly impractical for execution on an AUV. We also studied the sensitivity of the results by modifying the input parameters to have different correlations between temperature and salinity, standard deviations, and spatial correlation range. In all runs, the *myopic* and *look-ahead* strategies perform the best in terms of realized IBV, and much better than *naive*. The *look-ahead* strategy seems to be substantially better than the *myopic* design only for very small initial standard deviations or very large spatial correlation range.

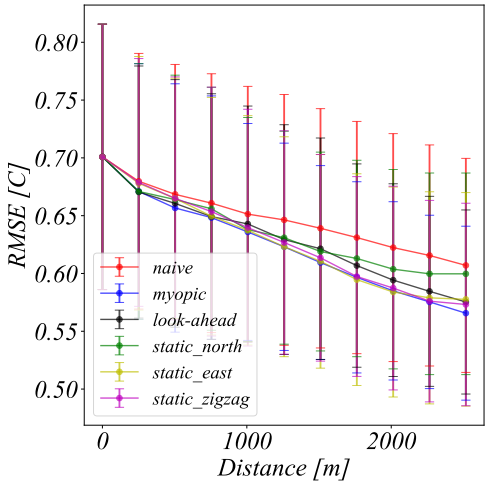
6.3.3 Real-world Application: River Plume Mapping in Trondheim Fjord

This subsection summarizes field experiments performed by Trygve Olav Fossum and Jo Eidsvik (NTNU Trondheim). These results are included for the sake of completeness but are not original work of the author of this thesis.

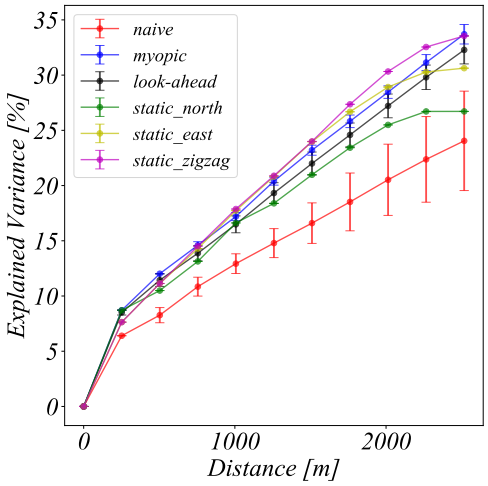
To demonstrate the applicability of using multivariate EPs and the IBV to inform



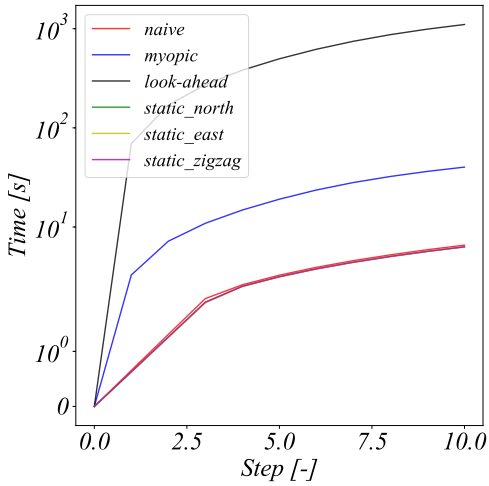
(a) IBV.



(b) RMSE



(c) Explained variance.



(d) Computational time for inference (the lines for *naive*, *static_north*, *static_east*, and *static_zigzag* overlap).

Figure 6.4: Simulation results from 100 replicate simulations for 10 sampling choices/stages on the grid. Vertical lines show variation in replicate results.

oceanographic sampling, we present a case study mapping a river plume with an AUV. The experiment was performed in Trondheim, Norway, surveying the Nidelva river (Fig. 2.4b). The experiments were conducted in late Spring 2019, when there is still snow melting in the surrounding mountains so that the river water is colder than the water in the fjord. The experiment was focused along the frontal zone that runs more or less parallel to the eastern shore. The GP model parameters were specified based on a short preliminary survey where the AUV made an initial transect to determine the trends in environmental conditions and correlation structures. Based on the initial runs we get a reasonable idea of the temperature and salinity of river and ocean waters, and also specify the trend by linear regression, where both temperature and salinity were assumed to increase linearly with the west coordinate. Next, the residuals from the regression analysis were analyzed to specify

the covariance parameters of the GRF model, leading to the parameters listed in Table 6.1. The regression parameters shown here are scaled to represent the east and west boundaries of the domain as seen in the preliminary transect data, and the thresholds are intermediate values. These parameter values were then used in field trials, where we explored the algorithm's ability to characterize the river plume front separating the river and fjord water masses.

Parameter	Value	Source
Cross correlation temperature and salinity	0.5	AUV observations
Temperature variance	0.20	AUV observations (variogram)
Salinity variance	5.76	AUV observations (variogram)
Correlation range	0.15 km	AUV observations (variogram)
River temperature	10.0 °C	AUV observations
Ocean temperature T_{ocean}	11.0 °C	AUV observations
River salinity S_{river}	14.0 g/kg	AUV observations
Ocean salinity S_{ocean}	22.0 g/kg	AUV observations
Threshold in temperature	10.5 °C	User specified
Threshold in salinity	18.0 g/kg	User specified

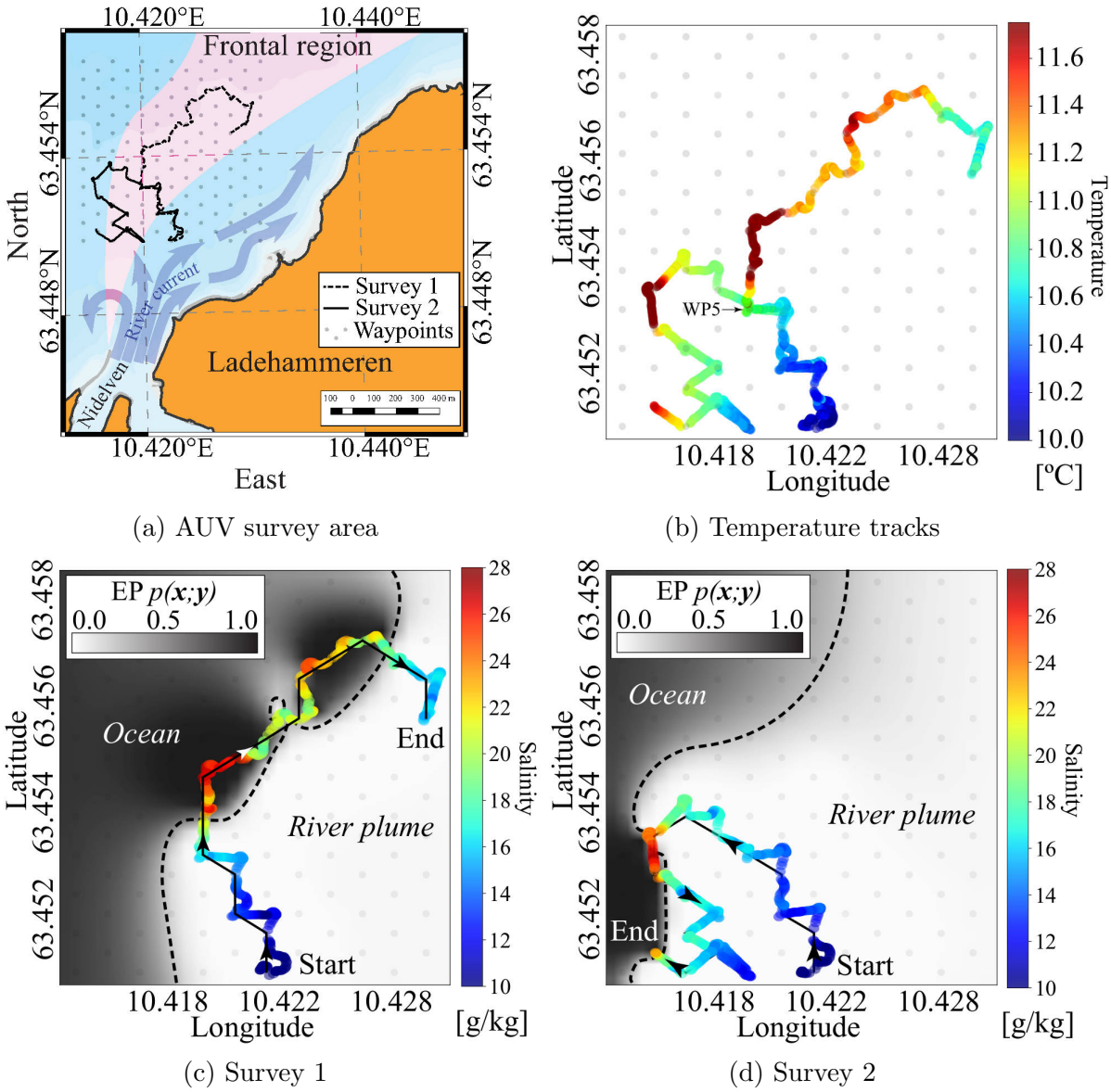
Table 6.1: Model and threshold parameters from an initial survey.

The platform used for data collection is a Light AUV (Sousa et al., 2012) (Fig. 6.5) equipped with a 16 Hz Seabird Fastcat-49 conductivity, temperature, and depth (CTD) sensor was used to provide salinity and temperature measurements. The AUV is a powered untethered platform that operates at 1-3 m/s in the upper water column. We assume that the measurements are conditionally independent because the salinity is extracted from the conductivity sensor which is different from the temperature sensor. We specify variance 0.25^2 for both errors, which is based on a middle ground between the nugget effect in the empirical variogram and the sensor specifications.

Figure 6.5: The commercially available Light Autonomous Underwater Vehicle (LAUV) platform for upper water-column exploration used in our experiments.

The AUV was guided using the *myopic strategy* from Section 6.3.1, computed on a equilateral grid discretization of the survey domain. At each stage, it takes the AUV about 30 seconds to assimilate data and evaluate the EIBV for all the possible candidates (grid nearest neighbors). The survey was set to take approximately 40 minutes, visiting 15 grid points in total, with the vehicle running near the surface to capture the plume. On its path from one grid point to the next, the AUV collects data with an update frequency of 30 seconds, giving three measurements per batch

in the updates at each stage. The resulting designs are shown in Fig. 6.6 for two surveys performed 2 hours apart.



(a) AUV survey area

(b) Temperature tracks

(c) Survey 1

(d) Survey 2

Figure 6.6: Results from mapping the Nidelva river, Trondheim, Norway over two survey missions. 6.6a shows an overview of the survey area overlaid with the AUV path in black and dashed line. Note the shaded region indicating a typical frontal region. 6.6b shows the collected temperature data as colored trails. Note waypoint 5 (WP5) which indicates where the two surveys diverge. 6.6c and 6.6d shows the collected salinity data overlaid on the final EP, which indicate the AUVs statistical impression of the front. For both missions, the temperature and salinity data correspond with an indication of the EP front. About 2 hours time separated the two runs.

6.3.4 Results

The recorded temperatures are shown as colored trails in Fig. 6.6b, clearly indicating the temperature difference between fjord and riverine waters. The salinity

data are then shown separately, overlaid with the estimated EP for each survey in Fig. 6.6c and Fig. 6.6d. We see that both surveys successfully estimated and navigated the separation zone, crossing the frontal boundary multiple times. As conditions changed slightly between the two surveys, the resulting trajectory (after waypoint 5) is shown to deviate. Survey 1 continued northwards, tracking the north-eastern portion of the front, while Survey 2 turned west, mapping the south-western region.

The final predictions of the front location, represented by conditional EPs in Fig. 6.6c and Fig. 6.6d as dashed lines, correspond with one another. In both surveys they yield a picture of the front being to the west in the southern portions of the region and gradually bending off toward the north east. The amount of exploration done by Survey 1 which turned north is greater than Survey 2 which was coming close to the survey area borders in the south-western corner.

6.4 Conclusion

By extending excursion set uncertainty functionals to multivariate Gaussian process, this work is able to provide strategies for estimation of multivariate excursion sets. In particular, the characterization of uncertainties in random sets is extended in the vector-valued case with new results for the expected integrated Bernoulli variance reduction achieved by spatial sampling designs. This is provided in semi-analytical form for static designs, and then extended to the adaptive situations. The sequential derivations provide new insights into efficient applications of adaptive data collection, as demonstrated in our application. characterizing water mass properties.

6.5 Appendix: Proofs of the Theorems

Proof. (Theorem 18) That $\nu(\Gamma)$ defines indeed a random variable follows from Fubini's theorem relying on the joint measurability of $(x,\omega) \to \mathbb{1}_{\Gamma(\omega)}(x)$, itself inherited from the assumed measurability for $(x,\omega) \to Z_x(\omega)$ and T, respectively. From there, following the steps of Robbins' theorem Robbins (1944), we find that

$$\begin{aligned}\mathbb{E}[\nu(\Gamma)^r] &= \mathbb{E}\left[\left(\int_D \mathbb{1}_{Z_x\in T}\, d\nu(x)\right)^r\right] = \mathbb{E}\left[\prod_{i=1}^r \left(\int_D \mathbb{1}_{Z_{x_i}\in T}\, d\nu(x_i)\right)\right] \\ &= \mathbb{E}\left[\int_{D^r} \mathbb{1}_{Z_{x_1}\in T,\ldots,Z_{x_r}\in T}\, \mathrm{d}\nu^{\otimes}(\boldsymbol{X})\right] = \int_{D^r} \mathbb{P}\left(Z_{\boldsymbol{X}} \in T^r\right) \mathrm{d}\nu^{\otimes}(\boldsymbol{X}),\end{aligned}$$

where $\boldsymbol{X} = (x_1, ..., x_r) \in D^r$. The rest consists in expliciting the probability of $T \times \cdots \times T$ under the multivariate Gaussian distribution of $(Z_{x_1}, \ldots, Z_{x_r})$. □

The propositions below provide formulae for computations of expectations of moments of multivariate Gaussian CDFs.

Proposition 3. *Let $p, q, h \geq 1$, $a \in \mathbb{R}^p$, $B \in \mathbb{R}^{p\times q}$, and C, C_V be two covariance matrices in $\mathbb{R}^{p\times p}$ and $\mathbb{R}^{q\times q}$, respectively. Then, for $V \sim \mathcal{N}_q(0_q, C_V)$,*

$$\mathbb{E}\left[\Phi_p\left(a + BV; C\right)^h\right] = \Phi_{ph}\left(\boldsymbol{a};\ \boldsymbol{\Sigma}\right),$$

where the vector $\boldsymbol{a} \in \mathbb{R}^{ph}$ is defined as $\boldsymbol{a} := 1_h \otimes a = (a, \ldots, a)'$ and the $ph \times ph$ covariance matrix is given by $\boldsymbol{\Sigma} := 1_h 1_h' \otimes BC_V B' + I_h \otimes C$.

Remark 10. In blockwise representation, $\mathbf{\Sigma}$ can be expressed as follows:

$$\begin{pmatrix} C & & \\ & \ddots & \\ & & C \end{pmatrix} + \begin{pmatrix} BC_VB' & \dots & BC_VB' \\ \vdots & & \vdots \\ BC_VB' & \dots & BC_VB' \end{pmatrix}$$

Proof. By definition of Φ_p, for $N \sim \mathcal{N}_p(0_p, C)$,

$$\mathbb{P}(N \leq a + BV|V) = \Phi_p\left(a + BV; C\right).$$

Now for $\Phi_p\left(a + BV; C\right)^h$, provided that the probability space is sufficiently large to accommodate h independent Gaussian random vectors $N_i \sim \mathcal{N}_p(0, C)$ (which is silently assumed here), using the former equality delivers

$$\Phi_p\left(a + BV; C\right)^h = \prod_{i=1}^{h} \mathbb{P}(N_i \leq a + BV|V).$$

Now by independence of the N_i's we obtain the joint conditional probability

$$\prod_{i=1}^{h} \mathbb{P}(N_i \leq a + BV|V) = \mathbb{P}(N_1 \leq a + BV, \dots, N_h \leq a + BV|V),$$

whereof, by virtue of the law of total expectation,

$$\begin{aligned} \mathbb{E}\left[\Phi_p\left(a + BV; K^{(n)}\right)^h\right] &= \mathbb{E}\left[\mathbb{P}(N_1 \leq a + BV, \dots, N_h \leq a + BV|V)\right] \\ &= \mathbb{P}(N_1 \leq a + BV, \dots, N_h \leq a + BV) \\ &= \mathbb{P}(W_1 \leq a, \dots, W_h \leq a) \\ &= \Phi_{ph}\left(1_h \otimes a; (1_h 1_h') \otimes (B\Sigma_V B') + I_h \otimes C\right), \end{aligned}$$

where $\mathbf{W} = (W_1, \dots, W_h)$ with $W_i = N_i - BV$ $(1 \leq i \leq h)$ and the last line follows $\mathbf{W}$ forming a Gaussian vector (by global independence of the N_i's and V) and from the definition of Φ_{ph}. The covariance matrix $\mathbf{\Sigma}$ of $\mathbf{W}$ is obtained by noting that $\mathrm{cov}(W_i, W_j) = BC_VB' + \delta_{ij}C$ $(i, j \in \{1, \dots, h\})$. □

We now generalize Proposition 3 to the case of multivariate monomials in orthant probabilities with thresholds affine in a common Gaussian vector.

Proposition 4. *Let $g, p, q \geq 1$, $h_1, \dots, h_g \geq 1$ with $H = \sum_{i=1}^{g} h_i$, $a_i \in \mathbb{R}^p$, $B_i \in \mathbb{R}^{p \times q}$, and covariance matrices $C_i \in \mathbb{R}^{p \times p}$ $(1 \leq i \leq g)$. Then, for any covariance matrix $C_V \in \mathbb{R}^{q \times q}$ and $V \sim \mathcal{N}_q(0_q, C_V)$,*

$$\mathbb{E}\left[\prod_{i=1}^{g} \Phi_p\left(a_i + B_iV; C_i\right)^{h_i}\right] = \Phi_{pH}\left(\boldsymbol{a}; \mathbf{\Sigma}\right), \tag{6.13}$$

with $\boldsymbol{a} = (1_{h_1} \otimes a_1, \dots, 1_{h_g} \otimes a_g) \in \mathbb{R}^{pH}$ and $\mathbf{\Sigma} \in \mathbb{R}^{pH \times pH}$ is defined blockwise by $(\Sigma_{i,j})_{i,j \in \{1,\dots,g\}}$ where, for any $i, j \in \{1, \dots, g\}$,

$$\Sigma_{i,j} = (1_{h_i} 1_{h_j}') \otimes (B_i \Sigma_V B_j') + \delta_{i,j}(I_{h_i} \otimes C_i) \in \mathbb{R}^{ph_i \times ph_j}. \tag{6.14}$$

Remark 11. Using blockwise representation for the blocks themselves delivers

$$\Sigma_{ij} = \begin{pmatrix} B_i\Sigma_V B_j' & \dots & B_i\Sigma_V B_j' \\ \vdots & & \vdots \\ B_i\Sigma_V B_j' & \dots & B_i\Sigma_V B_j' \end{pmatrix} + \delta_{ij} \begin{pmatrix} C_i & & \\ & \ddots & \\ & & C_i \end{pmatrix}$$

Here each Σ_{ij} is made of h_i times h_j (vertically/horizontally) $p \times p$ sub-blocks, hence possesses ph_i lines and ph_j columns.

Proof. The proof relies (again) heavily on the fact that, by definition of Φ_p, for any covariance matrix $C \in \mathbb{R}^{p\times p}$, $a \in \mathbb{R}^p$, $B \in \mathbb{R}^{p\times q}$, and $N \sim \mathcal{N}_p(0_p, C)$,

$$\mathbb{P}(N \leq a + BV|V) = \Phi_p\left(a + BV; C\right).$$

In particular, for globally independent $N_{i,j} \sim \mathcal{N}_p(0_p, C_i)$ $(1 \leq j \leq h_i, 1 \leq i \leq g)$,

$$\begin{aligned}\prod_{i=1}^{g} \Phi_p\left(a_i + B_iV; C_i\right)^{h_i} &= \prod_{i=1}^{g}\prod_{j=1}^{h_i} \mathbb{P}(N_{i,j} \leq a_i + B_iV|V) \\ &= \mathbb{P}(N_{1,1} \leq a_1 + B_1V, \dots, N_{g,h_g} \leq a_g + B_gV|V),\end{aligned}$$

so that, by the law of total expectation,

$$\mathbb{E}\left[\prod_{i=1}^{g} \Phi_p\left(a_i + B_iV; C_i\right)^{h_i}\right] = \mathbb{P}(W_1 \leq 1_{h_1} \otimes a_1, \dots, W_g \leq 1_{h_g} \otimes a_g)$$

where $W_1 = (N_{1,1} - B_1V, \dots, N_{1,h_1} - B_1V), W_2 = (N_{2,1} - B_2V, \dots, N_{2,h_2} - B_2V), \dots, W_g = (N_{g,1} - B_gV, \dots, N_{g,h_g} - B_gV)$. Noting that $\mathbf{W} = (W_1, \dots, W_g)$ is a centred pH-dimensional Gaussian random vector, we finally obtain that

$$\mathbb{E}\left[\prod_{i=1}^{g} \Phi_p\left(a_i + B_iV; C_i\right)^{h_i}\right] = \Phi_{pH}\left(\boldsymbol{a}; \boldsymbol{\Sigma}\right),$$

with $\boldsymbol{a} = (1_{h_1} \otimes a_1, \dots, 1_{h_g} \otimes a_g)$ and $\boldsymbol{\Sigma} = (\mathrm{cov}(W_i, W_j))_{i,j\in\{1,\dots,g\}}$. □

Those two general results allow us to derive simple expressions for the expected effect of the inclusion of new datapoints on the IBV (Proposition 1) and on the EMV (Proposition 2)) for which we provide proofs below.

Proof. (Proposition 1) Applying Tonelli-Fubini followed by the law of total expectation first delivers

$$\begin{aligned}\mathrm{EIBV}(\boldsymbol{\chi}; Z^{(n)}) &= \int_D \mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{P}\left(Z_x^{(n)} \in T|Z_{\boldsymbol{\chi}}^{(n)} + \boldsymbol{\epsilon} = \boldsymbol{Y}\right)\left(1 - \mathbb{P}\left(Z_x^{(n)} \in T|Z_{\boldsymbol{\chi}}^{(n)} + \boldsymbol{\epsilon} = \boldsymbol{Y}\right)\right)\right] d\nu(x) \\ &= \int_D \Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x^{(n)}, \boldsymbol{K}_{xx}^{(n)}\right) \mathrm{d}\nu(x) \\ &\quad - \int_D \mathbb{E}_{\boldsymbol{Y}}\left[\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x^{(n+1)}(\boldsymbol{Y}), \boldsymbol{K}_{xx}^{(n+1)}\right)^2\right] \mathrm{d}\nu(x)\end{aligned}$$

where $\boldsymbol{m}_x^{(n+1)}(\boldsymbol{Y})$ and $\boldsymbol{K}_{xx}^{(n+1)}$ denote the conditional mean and covariance function conditionally on the data $\boldsymbol{Y}$. Now, by using the cokriging update formulae Eqs. (6.5) and (6.6) we get:

$$\begin{aligned}
&\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x^{(n+1)}(\boldsymbol{Y}), \boldsymbol{K}_{xx}^{(n+1)}\right)\\
=&\Phi_p\left(\boldsymbol{t} - \boldsymbol{m}_x^{(n+1)}(\boldsymbol{Y}); \boldsymbol{K}_{xx}^{(n+1)}\right)\\
=&\Phi_p\left(\boldsymbol{t} - \boldsymbol{m}_x - \boldsymbol{\lambda}_{n+1}(x)^T(\boldsymbol{Y} - \boldsymbol{m}_x^{(n)}); \boldsymbol{K}_{xx}^{(n+1)}\right)\\
=&\Phi_p\left(a + BV, \boldsymbol{K}_{xx}^{(n+1)}\right),
\end{aligned}$$

with $a = \boldsymbol{t} - \boldsymbol{m}_x$, $B = -\boldsymbol{\lambda}_{n+1}(x)^T$ and $V = \boldsymbol{Y} - \boldsymbol{m}_x^{(n)}$. Applying Proposition 3 then delivers that

$$\mathbb{E}_{\boldsymbol{Y}}\left[\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x^{(n+1)}(\boldsymbol{Y}), \boldsymbol{K}_{xx}^{(n+1)}\right)^2\right] = \Phi_{2p}\left(\begin{pmatrix}\boldsymbol{t} - \boldsymbol{m}_x^{(n)}\\ \boldsymbol{t} - \boldsymbol{m}_x^{(n)}\end{pmatrix}; \boldsymbol{\Sigma}^{(n)}(x)\right),$$

with $\boldsymbol{\Sigma}^{(n)}(x)$ as in the formulation of the proposition. This completes the proof. □

Proof. (Proposition 2)

$$\begin{aligned}
\mathrm{EEMV}(\boldsymbol{\chi}; Z^{(n)}) = &\int_{D\times D} \Phi_{2p}\left(1_2 \otimes \boldsymbol{t}; \boldsymbol{m}_{[x,x']}, \boldsymbol{K}_{[x,x'][x,x']}\right)\ \mathrm{d}\nu^{\otimes}(x,x')\\
&- \int_{D\times D} \mathbb{E}_{\boldsymbol{Y}}\left[\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_x, \boldsymbol{K}_{xx}\right)\Phi_p\left(\boldsymbol{t}; \boldsymbol{m}_{x'}, \boldsymbol{K}_{x'x'}\right)\right]\mathrm{d}\nu^{\otimes}(x,x').
\end{aligned}$$

and the proof follows by applying Proposition 4 with

$$V = \boldsymbol{Y} - \boldsymbol{m}_{\boldsymbol{\chi}}^{(n)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{\chi}\boldsymbol{\chi}}^{(n)})$$

and $a_1 = \boldsymbol{t} - \boldsymbol{m}_x$, $B_1 = -\boldsymbol{\lambda}_{n+1}(x)^T$, $a_2 = \boldsymbol{t} - \boldsymbol{m}_{x'}$, and $C_1 = \boldsymbol{K}_{xx}^{(n)}$, $C_2 = \boldsymbol{K}_{xx}^{(n)}$. □

Chapter 7

Conclusion and Perspectives

In this thesis, we have proposed new approaches for implicit set estimation in Bayesian inverse problems and demonstrated them on real-world inverse problems originating in the natural sciences. More specifically, we have focused on linear inverse problems with Gaussian process priors and have devoted special attention to large-scale settings.

In Chapter 2 we have introduced the necessary background in inverse problem theory and Gaussian processes. We have also presented the basics of Gaussian measure theory that is used to build connections between the Gaussian process and Gaussian measure point of view of Bayesian inversion in Chapter 3. We have also presented the most recent developments in Bayesian set estimation (Azzimonti et al., 2016; Chevalier et al., 2013) based on the theory of random sets (Molchanov, 2005). Finally, we have introduced two concrete inverse problems that are used throughout the thesis to demonstrate our contributions in realistic settings. The first of these is a gravimetric inverse problem where the aim is to reconstruct the interior density field of the Stromboli volcano from observations of the gravity field on the surface. This problem exemplifies the challenges arising in large-scale Bayesian inversion owing to the non-sparse nature of the involved observation operators and to the three-dimensional nature of the problem. The second problem introduced is a river plume mapping task that is of special interest for demonstrating Bayesian estimation of excursion sets of multivariate functions.

In Chapter 3 we have studied the connection between the Gaussian process and Gaussian measure point of view of Bayesian inversion. By bridging classical results in Gaussian measure theory (Rajput and Cambanis, 1972) with recent developments in the study of GP sample paths properties (Steinwart, 2019), we are able to provide conditions under which the two point of views are equivalent. Then, we leverage the framework of disintegrations of measures to derive a purely functional formulation of Gaussian process updating under linear operator observations, providing an abstract and generic version of the Gaussian update formulae.

In Chapter 4, we have introduced a new representation of the posterior covariance of GPs. We have shown how this representation allows for efficient updating of GPs under linear operator data and how it enables substantial computational savings in the computation of the posterior. This new representation is given sound foundations by basing on the developments in Gaussian measure theory from Chapter 3. We have demonstrated how our techniques allow for sequential estimation of excursion sets in a gravimetric inverse problem. To the best of our knowledge, this is

the first time that sequential design criteria for excursion set estimation are applied in Bayesian inverse problems. We have shown how the criteria are able to significantly reduce the uncertainty on the excursion volume. In passing, we have also shown how to efficiently sample from the posterior in large-scale inverse problems.

In Chapter 5 we have introduced *universal inversion* as an extension of universal kriging that allows for the inclusion of parametric trends in Bayesian inversion. We have derived explicit formulae for the posterior when trend coefficients are treated in a Bayesian way and have computed the uninformative prior limit. We have demonstrated how our framework allows for the incorporation of expert knowledge in Bayesian inversion and have leveraged fast k-fold cross-validation results Ginsbourger and Schärer (2021) to provide heuristic diagnostics for model selection. We have also shown how cross-validation can be used for hyperparameter training. To the best of our knowledge, this is the first time that k-fold cross-validation is used for hyperparameter training in Bayesian inversion.

In Chapter 6 we have considered multivariate extensions of GPs. We have introduced the novel concept of *generalized locations* and shown how it allows for the co-kriging equation to be written in a from-invariant way in the most general setting (heterotopic observations). Then, by extending excursion set uncertainty functionals to the multivariate setting, we have developed sequential uncertainty reduction strategies for the estimation of excursion sets of multivariate functions, providing semi-analytic formulae for the computation of the sampling criteria. We have demonstrated how our techniques on a river plume mapping problem.

Overall, this thesis shows how traditional GP techniques can be made to scale to large-scale Bayesian inverse problems and how they can be extended to multivariate settings, while still enjoying strong theoretical foundations.

Perspectives: While we have shown how Bayesian inversion techniques can be successfully applied to real-world inverse problems, these still rely on simplifying assumptions that need to be replaced if one wants to build a fully realistic framework for implicit set estimation in Bayesian inverse problems. We next list what we think are the most promising directions for future research.

- **Beyond Gaussianity:** During this whole thesis, we only considered Gaussian process priors. While this class of priors is of great interest owing to its tractability, it suffers from clear limitations. In most applications there is no reason to expect the underlying phenomenon to be a realization of a GP. For example, in gravimetric inversion, density fields can only take on positive values, which clearly violates the Gaussian assumptions. Also, most natural phenomenon exhibit (spatial) non-stationarity, which cannot be described by the usual GP covariance kernels (Matérn family, ...). While still remaining in the Gaussian realm, one can build more realistic models by incorporating non-stationarity. One possible such approach is to formulate the GP prior as a solution to a stochastic partial differential equation with spatially varying coefficients (Lindgren et al., 2011). When considering non-Gaussian models, analytical formulae for the posterior usually aren't available, forcing one to resort to MCMC to approximate the posterior. This tends to make non-Gaussian models computationally expansive and inapplicable to real-world problems. One class of non-Gaussian models that still enjoys closed form formulae for the posterior is that of skew-Gaussian processes (Benavoli et al.,

2021), and we believe these could be successfully applied to Bayesian inversion. Another class of non-Gaussian models that could be of potential interest for Bayesian inversion are Besov priors. While their theoretical properties in this context have been thoroughly studied (Dashti et al., 2012), convincing practical applications are stilly lacking.

- **Cost-aware Path Planning:** While the sequential uncertainty reduction criterion studied in Section 4.4.3 provide good performance for excursion set estimation, they are out of touch with reality, in that the feasibility of the proposed data collection plan is never questioned and the terrain-specific constraints are not included the design process. Indeed, in most inverse problems, data collection is limited or influenced by domain specificities. For example, in our Stromboli gravimetric inverse problem, some locations may be harder to reach than others and some are even inaccessible. Furthermore, there may exist topographic feature that influence the optimal path between two data collection locations. Apart from these local factors, there may also exist global features that influence the data collection process, such as the presence of shortcuts between two distant locations (in the case of the Stromboli island a ferry line) or global constraints (having to be back to base before nightfall).

  While all these factors could, in theory, be taken into account by adding a cost term to the sampling criterion Eq. (4.10), myopic optimization as performed in Section 4.4.3 performs poorly in the presence of global features. Such settings require more sophisticated dynamic programming approaches for path optimization that include a given amount of lookahead. While there has been some preliminary work in adding path constraints to sequential design (Ge et al., 2022), we believe that a fully realistic framework is still lacking.

- **Cross-Validation Diagnostics for Model Selection:** In Chapter 5 the potential of cross-validation for model selection has been briefly touched upon in a heuristic fashion, without providing any theoretical development. We believe that the availability of fast formulae for the computation of the k-fold cross-validation residuals, as well as their theoretical covariance Theorem 15 allows for the creation of sophisticated models selection procedures and diagnostics, as sketched in (Ginsbourger and Schärer, 2021).

  For example, the residual covariance matrix can be used to decorrelate the residuals, allowing one to apply the usual tests available for Gaussian data. Figure 7.1 shows a QQ-plot of the decorrelated residuals for the first fold in the 10-fold cross-validation example of Section 5.5.1.

  While these plots suggest that the behavior of this fold is captured surprisingly well by all models, we stress that the situation is drastically different for other folds. Overall, one should instead look at QQ-plots of the concatenated vector of all folds residuals. Nevertheless, the computation of the associated residual covariance is fraught with numerical instabilities, owing to the multiple inversions of ill-conditioned matrices involved in the process. We believe that the development of numerically stable computation techniques for the covariance of the residuals is a necessary next step towards

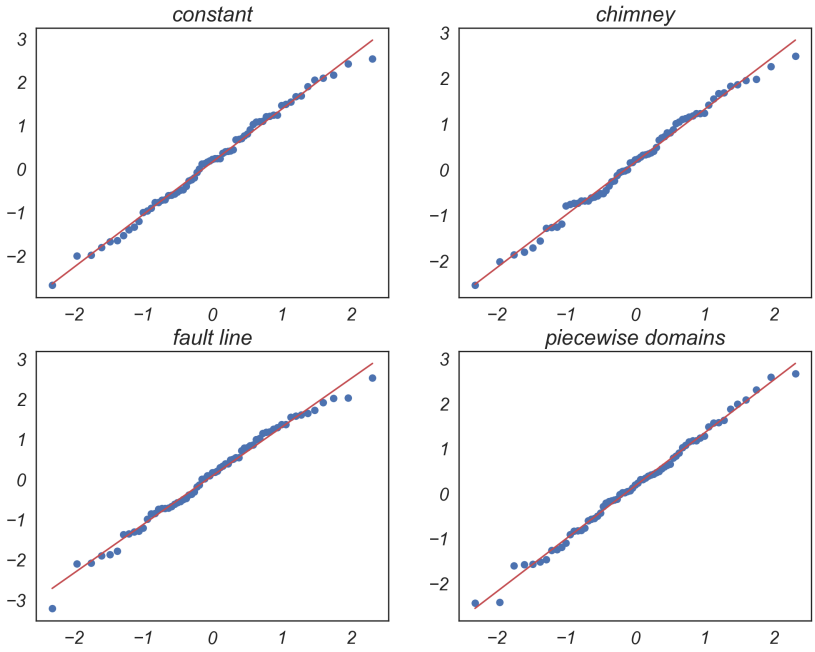


Figure 7.1: Quantile-quantile plot of decorrelated cross-validation residuals for fold nr.1 in 10-fold CV.

more sophisticated CV diagnostics. Apart from that, the phenomenon of "fold outliers" identified in Section 5.5.1, where some given folds are significantly harder to predict has to be examined further. Towards that end, one should investigate the behaviour of CV under different clustering schemes. Finally, in order to come up with a principled model selection criterion, one should study how model complexity is handled by cross-validation, since it seems that CV already penalizes complex models under the hood.

Apart from these main directions, we also think that one venture worth pursuing is the application of more sophisticated set uncertainty quantification criteria in Bayesian inversion. One natural class of criteria that can be of interest in inverse problems is that of conservative set estimation (Azzimonti et al., 2021). On the more theoretical side, the question of the consistency of the sequential designs in the multivariate setting (Chapter 6) should be elucidated and we believe that extending the results from (Bect et al., 2019) to the multivariate case should prove a fruitful effort.

Overall, optimal design for implicit set estimation in Bayesian inverse problems is a research topic with promising applications in the natural sciences. We strongly think that it has the potential to dramatically change the way data is collected and the way field campaigns are planned in the natural sciences and hope that this thesis can serve as a first step in that direction.

Bibliography

Agrell, C. (2019). Gaussian processes with linear operator inequality constraints. *Journal of Machine Learning Research*, 20(135):1–36.

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Armstrong, M. (1984). Problems with universal kriging. *Journal of the International Association for Mathematical Geology*, 16(1):101–108.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Attia, A., Alexanderian, A., and Saibaba, A. (2018). Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems*, 34.

Azzimonti, D., Bect, J., Chevalier, C., and Ginsbourger, D. (2016). Quantifying uncertainties on excursion sets under a Gaussian random field prior. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):850–874.

Azzimonti, D., Ginsbourger, D., Chevalier, C., Bect, J., and Richet, Y. (2021). Adaptive design of experiments for conservative estimation of excursion sets. *Technometrics*, 63(1):13–26.

Bachoc, F. (2013). *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments*. PhD thesis, Université Paris-Diderot-Paris VII.

Banerjee, B. and Das Gupta, S. (1977). Gravitational attraction of a rectangular parallelepiped. *Geophysics*, 42(5):1053–1055.

Barnes, R. J. and Watson, A. (1992). Efficient updating of kriging estimates and variances. *Mathematical Geology*, 24(1):129–133.

Basak, S., Petit, S., Bect, J., and Vazquez, E. (2022). Numerical issues in maximum likelihood parameter estimation for gaussian process interpolation. In *Machine Learning, Optimization, and Data Science: 7th International Conference, LOD 2021, Grasmere, UK, October 4–8, 2021, Revised Selected Papers, Part II*, pages 116–131. Springer.

Bect, J., Bachoc, F., and Ginsbourger, D. (2019). A supermartingale approach to gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919.

Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793.

Ben-Israel, A. and Greville, T. N. (2003). *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media.

Benavoli, A., Azzimonti, D., and Piga, D. (2021). A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with skew gaussian processes. *Machine Learning*, 110(11-12):3095–3133.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Berrino, G. and Camacho, A. G. (2008). 3d gravity inversion by growing bodies and shaping layers at mt. vesuvius (southern italy). *Pure and Applied Geophysics*, 165(6):1095–1115.

Bhattacharjya, D., Eidsvik, J., and Mukerji, T. (2013). The value of information in portfolio problems with dependent projects. *Decision Analysis*, 10(4):341–351.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley, New York.

Blakely, R. J. (1995). *Potential Theory in Gravity and Magnetic Applications*. Cambridge University Press.

Bogachev, V. I. (1998). *Gaussian measures*. Number 62. American Mathematical Soc.

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent gaussian models. *Journal of the Royal Statistical Society, Series B Methodology*, 77(1):85–106.

Bovier, A. (2015). Stochastic Processes: Lecture, Summer term 2013, Bonn. URL:`https://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Patrik_Ferrari/Lectures/SS16StochProc/wt2-new.pdf`. Last visited on 2019/10/29.

Bühler, T. and Salamon, D. A. (2018). *Functional analysis*. American Mathematical Society, Providence, Rhode Island.

Calvetti, D. and Somersalo, E. (2018). Inverse problems: From regularization to bayesian inference. *WIREs Computational Statistics*, 10(3):e1427.

Cambanis, S. (1973). On some continuity and differentiability properties of paths of gaussian processes. *Journal of Multivariate Analysis*, 3(4):420–434.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014a). Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465.

Chevalier, C., David, G., and Emery, X. (2015). Fast update of conditional simulation ensembles. *Mathematical Geosciences*, 47:771–789.

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models. In *mODa 10–Advances in Model-Oriented Design and Analysis*, pages 35–43. Springer.

Chevalier, C., Ginsbourger, D., and Emery, X. (2014b). Corrected kriging update formulae for batch-sequential data assimilation. In Pardo-Igúzquiza, E., Guardiola-Albert, C., Heredia, J., Moreno-Merino, L., Durán, J., and Vargas-Guzmán, J., editors, *Mathematics of Planet Earth. Lecture Notes in Earth System Sciences*. Springer, Berlin, Heidelberg.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley, 2 edition.

Chilès, J.-P. and Desassis, N. (2018). *Fifty Years of Kriging*, pages 589–612. Springer International Publishing, Cham.

Conti, S. and O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140(3):640–651.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). Mcmc methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (revised edition)*. Wiley.

Dashti, M., Harris, S., and Stuart, A. (2012). Besov priors for bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200.

Dashti, M. and Stuart, A. M. (2016). The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 1–118.

de Fouquet, C. (1994). Reminders on the conditioning kriging. In Armstrong, M. and Dowd, P. A., editors, *Geostatistical Simulations*, pages 131–145, Dordrecht. Springer Netherlands.

Driscoll, M. F. (1973). The reproducing kernel hilbert space structure of the sample paths of a gaussian process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26:309–316.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

Duhamel, C., Helbert, C., Munoz Zuniga, M., Prieur, C., and Sinoquet, D. (2023). A sur version of the bichon criterion for excursion set estimation. *Statistics and Computing*, 33.

Emery, X. (2009). The kriging update equations and their application to the selection of neighboring data. *Computational Geosciences*, 13(3):269–280.

Folland, G. B. (2013). *Real analysis: modern techniques and their applications*. John Wiley & Sons.

Fossum, T. O., Norgren, P., Fer, I., Nilsen, F., Koenig, Z. C., and Ludvigsen, M. (2021a). Adaptive sampling of surface fronts in the arctic using an autonomous underwater vehicle. *IEEE Journal of Oceanic Engineering*, 46(4):1155–1164.

Fossum, T. O., Travelletti, C., Eidsvik, J., Ginsbourger, D., and Rajan, K. (2021b). Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean sampling. *The Annals of Applied Statistics*, 15(2):597 – 618.

French, J. and Sain, S. (2013). Spatio-temporal exceedance locations and confidence regions. *Annals of Applied Statistics*, 7 (3):1421–1449.

Gao, H., Wang, J., and Zhao, P. (1996). The updated kriging variance and optimal sample design. *Mathematical Geology*, 28(3):295–313.

Ge, Y., Olaisen, A. J. H., Eidsvik, J., Jain, R. P., and Johansen, T. A. (2022). Long-horizon informative path planning with obstacles and time constraints. *IFAC-PapersOnLine*, 55(31):124–129. 14th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2022.

Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2):147–163.

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media.

Ginsbourger, D. (2018). Sequential design of computer experiments. *Wiley StatsRef: Statistics Reference Online*, 99:1–11.

Ginsbourger, D. and Schärer, C. (2021). Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances. *arXiv preprint arXiv:2101.03108*.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Gowrisankaran, K. (1972). Measurability of functions in product spaces. *Proceedings of the American Mathematical Society*, 31(2):485–488.

Guillen, A., Calcagno, P., Courrioux, G., Joly, A., and Ledru, P. (2008). Geological modelling from field data and geological knowledge: Part ii. modelling validation using gravity and magnetic data inversion. *Physics of the Earth and Planetary Interiors*, 171(1-4):158–169.

Hadamard, J. (1902). Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52.

Hairer, M. (2009). An introduction to stochastic pdes. *arXiv preprint arXiv:0907.4178*.

Hairer, M., Stuart, A. M., Voss, J., and Wiberg, P. (2005). Analysis of SPDEs arising in path sampling. Part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587 – 603.

Hamelijnck, O., Wilkinson, W. J., Loppi, N. A., Solin, A., and Damoulas, T. (2021). Spatio-temporal variational gaussian processes. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc.

Hansen, P. C. (2010). *Discrete Inverse Problems*. Society for Industrial and Applied Mathematics.

Helbert, C., Dupuy, D., and Carraro, L. (2009). Assessment of uncertainty in computer experiments from universal to bayesian kriging. *Applied Stochastic Models in Business and Industry*, 25(2):99–113.

Hendriks, J. N., Jidling, C., Wills, A., and Schön, T. B. (2018). Evaluating the squared-exponential covariance function in gaussian processes with integral observations.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Huber, M. F. (2014). Recursive gaussian process: On-line regression and learning. *Pattern Recognition Letters*, 45:85–91.

Jidling, C., Hendriks, J., Schön, T. B., and Wills, A. (2019). Deep kernel learning for integral measurements.

Jidling, C., Hendriks, J., Wahlström, N., Gregg, A., Schön, T. B., Wensrich, C., and Wills, A. (2018). Probabilistic modelling and reconstruction of strain. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 436:141–155.

Jidling, C., Wahlström, N., Wills, A., and Schön, T. B. (2017). Linearly constrained Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1215–1224.

Journel, A. and Huijbregts, C. (1978). *Mining Geostatistics*. Academic Press.

Kallenberg, O. (2021). *Foundations of Modern Probability*. Springer International Publishing.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *ArXiv*, abs/1807.02582.

Karvonen, T. (2021). Small sample spaces for gaussian processes. *arXiv preprint arXiv:2103.03169*.

Kitanidis, P. K. (1995). Quasi-linear geostatistical theory for inversing. *Water Resources Research*, 31(10):2411–2419.

Kitanidis, P. K. (2015). Compressed state Kalman filter for large systems. *Advances in Water Resources*, 76:120 – 126.

Klebanov, I., Sprungk, B., and Sullivan, T. (2021). The linear conditional expectation in Hilbert space. *Bernoulli*, 27(4):2267 – 2299.

Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626 – 2657.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa*, 52(6):119–139.

Kuo, H.-H. (1975). *Gaussian measures in Banach spaces*. Lecture Notes in Mathematics. Springer, Berlin, Germany, 1975 edition.

Le Gratiet, L., Cannamela, C., and Iooss, B. (2015). Cokriging-based sequential design strategies using fast cross-validation for multi-fidelity computer codes. *Technometrics*, 57:418–427.

Linde, N., Baron, L., Ricci, T., Finizola, A., Revil, A., Muccini, F., Cocchi, L., and Carmisciano, C. (2014). 3-D density structure and geological evolution of Stromboli volcano (Aeolian Islands, Italy) inferred from land-based and sea-surface gravity data. *Journal of Volcanology and Geothermal Research*, 273:58–69.

Linde, N., Ricci, t., Baron, L., A., S., and G., B. (2017). The 3-D structure of the Somma-Vesuvius volcanic complex (Italy) inferred from new and historic gravimetric data. *Scientific Reports*, 7(1):8434.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Longi, K., Rajani, C., Sillanpää, T., Mäkinen, J., Rauhala, T., Salmi, A., Haeggström, E., and Klami, A. (2020). Sensor placement for spatial gaussian processes with integral observations. In Peters, J. and Sontag, D., editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1009–1018. PMLR.

Lukić, M. N. and Beder, J. H. (2001). Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.

Mandallaz, D. (2000). Estimation of the spatial covariance in universal kriging: application to forest inventory. *Environmental and Ecological Statistics*, 7(3):263–284.

Mandel, J. (2006). Efficient implementation of the ensemble Kalman filter. Technical Report 231, University of Colorado at Denver and Health Sciences Center.

Mandelbaum, A. (1984). Linear estimators and measurable linear transformations on a hilbert space. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(3):385–397.

Mantoglou, A. and Wilson, J. L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research*, 18(5):1379–1394.

Matheron, G. (1962). *Traité de géostatistique appliquée, Tome I.* Mémoires du Bureau de Recherches Géologiques et Minières, no. 14. Editions Technip, Paris.

Matheron, G. (1969). Le krigeage universel (universal kriging). vol. 1. *Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris, Fontainebleau, 83pp.*

Molchanov, I. (2005). *Theory of Random Sets.* Springer, London.

Montesinos, F. G., Arnoso, J., Benavent, M., and Vieira, R. (2006). The crustal structure of El Hierro (Canary Islands) from 3-D gravity inversion. *Journal of Volcanology and Geothermal Research*, 150(1-3):283–299.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1):1–42.

Omre, H. and Halvorsen, K. B. (1989). The bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21:767–786.

Owhadi, H. and Scovel, C. (2015). Conditioning gaussian measure on hilbert space.

Park, J.-S. and Baek, J. (2001). Efficient computation of maximum likelihood estimators in a spatial linear model with power exponential covariogram. *Computers & Geosciences*, 27(1):1–7.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pavliotis, G. A. (2014). *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer.

Penrose, R. (1956). On best approximate solutions of linear matrix equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(1):17–19.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R., and Kim, N. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132:071008.

Poloczek, M., Wang, J., and Frazier, P. (2017). Multi-information source optimization. In *Advances in Neural Information Processing Systems 30.*

Powell, W. B. (2016). Perspectives of approximate dynamic programming. *Annals of Operations Research*, 241(1-2):319–356.

Purisha, Z., Jidling, C., Wahlström, N., Schön, T. B., and Särkkä, S. (2019). Probabilistic approach to limited-data computed tomography reconstruction. *Inverse Problems*, 35(10):105004.

Rajput, B. S. and Cambanis, S. (1972). Gaussian processes and Gaussian measures. *Ann. Math. Statist.*, 43(6):1944–1952.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Represas, P., Catalão, J. a., Montesinos, F. G., Madeira, J., Mata, J. a., Antunes, C., and Moreira, M. (2012). Constraints on the structure of Maio Island (Cape Verde) by a three-dimensional gravity model: imaging partially exhumed magma chambers. *Geophysical Journal International*, 190(2):931–940.

Ribaud, M. (2018). *Krigeage pour la conception de turbomachines : grande dimension et optimisation multi-objectif robuste*. PhD thesis, École centrale de Lyon. Thèse de doctorat dirigée par Helbert, Céline, Blanchet-Scalliet, Christophette et Gillot, Frédéric Mathématiques Lyon 2018.

Robbins, H. E. (1944). On the Measure of a Random Set. *The Annals of Mathematical Statistics*, 15(1):70 – 74.

Romary, T., de Fouquet, C., and Malherbe, L. (2011). Sampling design for air quality measurement surveys: An optimization approach. *Atmospheric Environment*, 45(21):3613–3620.

Rudin, W. (1974). *Real and complex analysis*. McGraw-Hill Book Co., New York, second edition. McGraw-Hill Series in Higher Mathematics.

Ryan, J. P., Kudela, R. M., Birch, J. M., Blum, M., Bowers, H. A., Chavez, F. P., Doucette, G. J., Hayashi, K., Marin III, R., Mikulski, C. M., Pennington, J. T., Scholin, C. A., Smith, G. J., Woods, A., and Zhang, Y. (2017). Causality of an extreme harmful algal bloom in monterey bay, california, during the 2014–2016 northeast pacific warm anomaly. *Geophysical Research Letters*, 44(11):5571–5579.

Särkkä, S. (2011). Linear operators and stochastic partial differential equations in gaussian process regression. In Honkela, T., Duch, W., Girolami, M., and Kaski, S., editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 151–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Särkkä, S., Solin, A., and Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Process. Mag.*, 30(4):51–61.

Scheuerer, M. (2010). Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120(10):1879–1897.

Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaus associés (noyaux reproduisants). *J. Analyse Math.*, 13:115–256.

Selby, B. and Kockelman, K. M. (2013). Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29:24–32.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. (2003). Derivative observations in gaussian process models of dynamic systems. In *Advances in neural information processing systems*, pages 1057–1064.

Solin, A., Kok, M., Wahlström, N., Schön, T., and Särkkä, S. (2015). Modeling and interpolation of the ambient magnetic field by gaussian processes. *IEEE Transactions on Robotics*, PP.

Sousa, A., Madureira, L., Coelho, J., Pinto, J., Pereira, J., Sousa, J., and Dias, P. (2012). LAUV: The man-portable autonomous underwater vehicle. In *Navigation, Guidance and Control of Underwater Vehicles*, volume 3, pages 268–274.

Stange, P. A. (2022). Consistency of some gaussian process based sequential experimental design strategies in the vector-valued case. Master's thesis, University of Bern.

Steinwart, I. (2019). Convergence types and rates in generic karhunen-loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Stroh, R. (2018). *Planification d'expériences numériques en multi-fidélité: Application à un simulateur d'incendies*. PhD thesis, Université Paris-Saclay.

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics.

Tarantola, A. and Valette, B. (1982). Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics*, 20(2):219–232.

Tarantola, A., Valette, B., et al. (1982). Inverse problems= quest for information. *Journal of geophysics*, 50(1):159–170.

Tarieladze, V. and Vakhania, N. (2007). Disintegration of Gaussian measures and average-case optimal algorithms. *Journal of Complexity*, 23(4):851 – 866. Festschrift for the 60th Birthday of Henryk Woźniakowski.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Travelletti, C. and Ginsbourger, D. (2022). Disintegration of gaussian measures for sequential assimilation of linear operator data.

Travelletti, C., Ginsbourger, D., and Linde, N. (2023). Uncertainty quantification and experimental design for large-scale linear inverse problems under gaussian process priors. *SIAM/ASA Journal on Uncertainty Quantification*, 11(1):168–198.

Vakhania, N. N., Tarieladze, V. I., and Chobanyan, S. A. (1987). *Probability Distributions on Banach Spaces*. Springer Netherlands.

Vargas-Guzmán, J. A. and Jim Yeh, T.-C. (1999). Sequential kriging and cokriging: Two powerful geostatistical approaches. *Stochastic Environmental Research and Risk Assessment volume*, 13:416–435.

Ver Hoef, J. M. and Barry, R. P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69(2):275–294.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer.

Wagner, P.-R., Marelli, S., and Sudret, B. (2021). Bayesian model inversion using stochastic spectral embedding. *Journal of Computational Physics*, 436:110141.

Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., and Wilson, A. G. (2019). Exact Gaussian processes on a million data points. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 14622–14632. Curran Associates, Inc.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

Wilson, J. T., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. (2020). Efficiently sampling functions from Gaussian process posteriors. *ArXiv*, abs/2002.09309.