Towards Analytic Informative Path Planning

Conrad Stevens

A thesis submitted in fulfillment
of the requirements of the degree of
Master of Philosophy

THE UNIVERSITY OF
SYDNEY

Australian Centre for Field Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney

Submitted March 2024; revised 7 November 2024

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the University or other institute of higher learning, except where due acknowledgement has been made in the text.

Conrad Stevens

31 March 2024

Abstract

Informative path planning algorithms are used to guide agents in carrying out observation actions to best learn about their environment. Prior information, predictive modelling and operational constraints are some of the considerations that go into this engineering problem. This thesis take a head on, analytic Bayesian approach towards the problem with aims to reduce the dependence on prior assumptions and to improve the adaptability of autonomous information gathering algorithms.

First, conjugate priors are used to construct Bayesian sensor models. These keep the Bayesian update step in closed-form, giving more readily applicable solutions and algebraic context for numeric approximations that go beyond such forms.

These sensor models are then integrated into stochastic processes that model the quantity of interest across the environment. Further, a novel alternative to Gaussian processes is derived using the semi conjugate prior to the multivariate Gaussian distribution. This allows for a Conjugate prior over the covariance matrix of the Gaussian process while also permitting observations that have independent Gaussian noise. In a novel finding, the covariance of this proposed model is more stable than that of a canonical noisy t processes and its covariance converges to the that of the ground truth when given the true mean and kernel function.

These models are then integrated into informative path planning algorithms using Monte Carlo Tree search driven by the mutual information of each stochastic process. The algorithms were then tested on a series of potassium density maps over New South Wales Australia. The theoretical properties of each model were demonstrated on the data set, with t processes performing equally or better than Gaussian processes depending on the priors used.

Acknowledgements

This thesis would not have been possible without the support of my supervisors Salah Sukkarieh and Timothy Bailey.

To my Bunica and my Grandmother.

Contents

List of Figures

Nomenclature

List of Symbols

X	The quantity of interest across the environment
T	The spatial indices of X
Z	The space of possible observations about X
x	Quantity of interest in one dimensional case
x_τ	True value of x in one dimensional case
x_i	One instance of the quantity of interest
t_i	The spatial index of x_i
z_i	The space of possible observations of x_i
$\check{z}_i$	One particular instance of an observation of x_i
X_n	$\{x_1, ..., x_n\}$
$\mathcal{N}(x; \mu, \sigma^2)$	Normal distribution about x with mean μ and variance σ^2
$\mathcal{T}(x; \mu, \tau, v)$	t-distribution about x with mean μ, shape τ and degrees of freedom v
$\mathcal{G}(\lambda; \alpha, \beta)$	Gamma distribution about λ parametrized in terms of shape and rate
$\mathcal{IG}(\sigma^2; \alpha, \beta)$	Inverse gamma distribution about σ^2 parametrized in terms of shape and rate
$\mathcal{IX}(\sigma^2; \tau, v)$	Inverse Chi squared distribution about σ^2 with shape τ and degrees of freedom v
$\mathcal{IW}(\Sigma; \Psi, v)$	Inverse Wishart distribution about Σ with shape Ψ and degrees of freedom v
Σ_x	Covariance of X
Σ_z	Covariance of Z_n given X_n
λ	Precision (σ^{-2})
λ_z	Precision of z given x in one dimensional case
σ_z	standard deviation of z given x in one dimensional case
$\lambda_{z\tau}$	True precision of z given x in one dimensional case
q	$\lvert X \rvert$
$k(.,.)$	Kernel function
K_{qq}	Covariance matrix of X
K_{nn}	Covariance matrix of X_n

K_{**}	Covariance matrix of $X_{n+1:q}$
K_{n*}	Rectangular covariance matrix between X_n and $X_{n+1:q}$
S	Mahalanobis distance
$\phi(t)$	Expectation of x at spatial location t
$\psi(t)$	Variance of z given x at spatial location t
$H(.)$	Quantification of uncertainty
$I(.,.)$	Mutual information
$\Gamma(.)$	Gamma function
$\Gamma(.)_q$	Multivariate Gamma function
$\backslash$	left matrix divide: $A\backslash b$ is the vector x which solves $Ax = b$
tr(.)	the trace function

List of Acronyms

NSW	New South Wales
IPPA	informative path planning algorithms
QOI	quantity of interest
PDF	probability distribution function
PMF	probability mass function
KLD	Kullback–Leibler divergence
MCTS	Monte-Carlo Tree Search
MSE	mean squared error
MSPE	mean squared prediction error
SLAM	simultaneous localisation and mapping
POMDPs	partially observable markov decision processes
GP	Gaussian process
TP	t process
RBF	radial basis function

Chapter 1

Introduction

1.1 Classical Robotics and Machine Learning

A robot's ability to observe and act upon its surroundings is what distinguishes it from a mere machine. Autonomy, adaptability and robustness are all qualities engineers strive for motivating much of robotic research.

Due to the unpredictability of the world and the imperfect nature of sensors and controls, robots must make autonomous decisions using incomplete information. Statistics is the language used to quantify a robot's incomplete understanding of the world, often manifesting itself as a distribution over states. This may be a distribution over the state of a robot's joints, the environment, input controls and more.

In particular, Bayesian statistics has cemented itself as a pillar of classical robotics (Thrun et al. (2005)), offering fundamental approaches to rich fields such as simultaneous localisation and mapping (SLAM), partially observable markov decision processes (POMDPs), control theory and more. In recent years, exciting developments in the machine learning space, coupled with increasingly accessible compute have given rise to new approaches towards these robotics problems. Deep reinforcement learning based control, neural network driven state estimation and more. Machine learning has had a particularly strong impact on perception models and sensor fusion. Envi-

ronmental state estimation required for self driving vehicles has been a catalyst for much of this research (Soori et al. (2023)).

Besides logistic challenges such as data collection and compute, a frequent limitation of machine learning driven modes is poor explainability and statistical interpretability. In contrast, Bayesian models are analytically tractable and statistically explainable. Among many benefits, this is a desirable property when informing potentially dangerous or costly decisions.

Intractable machine learning models are also often designed to be integrated within classical methods. For example the integration of auto-encoders and convolutional neural networks within visual SLAM systems (Gao and Zhang (2017); Taheri and Xia (2021)).

Thus, despite the evolving fields in robotics featuring modern machine learning techniques, classical Bayesian techniques are still relied upon to provide grounded results and model architectures.

1.2 Informative Path Planning

This thesis will take an analytic Bayesian approach towards information gathering and informative path planning. Information gathering is the process of making observations to gain information about some quantity of interest (QOI). It takes many forms depending on the nature of the environment, modalities of observation, and all other facets of the problem. In industry for example, focus groups are used to gain information about a consumer market and geologic surveys are used to gain information about the presence of minerals. Adjacently in research, information gathering is the basis of scientific experiments (Lindley (1956)).

The "*no free lunch theorem*" from Wolpert (1996)[1] states there is no universally best statistic model, and hence there is no universally best information gathering algorithm. This makes designing autonomous information gathering algorithms challeng-

[1] Section 4.

ing especially given limited prior information. They must be problem specific while maintaining enough flexibility to adapt to unknowns in the environment.

Situations were the information gathering process is manual, laborious, dangerous or inaccessible to humans motivates the intervention of robotic systems. To Automate these processes, robots must be equipped with the sensors required to observe the QOI, the controls capable of situating these sensors and the the autonomy to make informative actions in the field. Some examples include: SwagBot, an autonomous swerve steer robot used to make observations across livestock pastures (Wallace (2020); Sirius, an autonomous submarine used to make benthic observations (Jackson (2023)); Zöe, an autonomous rover used to make spectral observations (Garza (2021)) and autonomous aerial vehicles used for terrain mapping (Popović et al. (2020)).

These platforms have automated laborious surveying tasks, optimized the information gathering process and used robots to access fields of view and environments not easily accessible to humans. Further, each case proved to be a challenging robotics problem.

In recent years, this problem has gained more attention with particular emphasis often placed on optimizing the agents action sequence under time and energy constraints. This will be discussed in more detail in Chapter 2.

This thesis aims to contribute to the field of informative path planning by focusing on belief models, their information acquisition functions, and the integration of sensor models into both. This thesis will approach the problem head on from an analytic Bayesian perspective.

1.3 Thesis Approach to Information Gathering

Bayesian Solutions

When applied to information gathering, Bayesian models are capable of consolidating multiple, potentially conflicting results, to arrive at one posterior belief about

the QOI. This allows for smooth, statistically justified transitions across beliefs as data is collected. It also makes information-based metrics like entropy and mutual information more readily available. Parametric approaches are discussed in how they may support Bayesian models, however, they are not the focus of this thesis.

Analytic Solutions

Information gathering problems will be defined in a way that yields analytic solutions, giving interpretable and computationally efficient solutions. Further, closed-form solutions are also often easier to integrate into greater encompassing systems.

However convenient, they may have difficulty modelling an environment who's ground truth cannot be analytically expressed, or their algebraic constraints may inhibit the model in other capacities. In any case, analytic models still give the foundation required to derive and approximate intractable models when they are necessary. Consider for example how the Kalman filter (closed-form model) serves as a launch point for the extended Kalman filter (approximation of intractable Bayes filtering model).

From Chapter 3 onwards this thesis specializes in cases where the QOI and observed values belong to the real numbers. This excludes informative path planning problems whose objective is to categorize elements in the environment. However, the stochastic processes discussed in this thesis have shown to also be capable of represent categoric beliefs Rasmussen et al. (2006)[2].

This thesis will further specialize in cases where the QOI is continuously present in the environment and thus may be mathematically described as a scalar field. Some examples of physically observable scalar fields include, pasture height measured by a depth camera, ocean surface temperatures measured by a thermometer or the vegetation indices of a forest measured by a multispectral camera. The models presented in this thesis may be applied to each of these information gathering problems.

[2]Chapter 3.

1.3.1 Thesis Contribution in the Context of Related Work

Analytic Bayesian sensor models are a well studied topic primarily motivated by control theory (Thrun et al. (2005)). Linear Gaussian sensor models are frequently used for their analytic properties, however, it is rarely discussed why they have such properties and what the extent of these properties are. Chapter 3 studies these models from the perspective of conjugate priors building off the statistic works of Fink (1997) and Murphy (2007). The full breadth of analytic sensor models over the real numbers is outlined from this perspective and one step further is explored leveraging semi conjugate priors. Some works have discussed this in a robotics context, particularly the Bayes filtering works of Roth et al. (2017) and Huang et al. (2017). However, few works have discussed this in the context of informative path planning. Chapter 3 further explores the informatic properties of these sensor models including the pertinent, mutual information between observations and the QOI. These become relevant in later chapters when integrating these sensor models into stochastic processes and informative path planning algorithms (IPPA)s.

Once observations have been made, machine learning / stochastic models may be used to generated a belief model about the QOI that goes beyond observed locations. There are many approaches to this depending on the available prior information, properties of the QOI, available compute and more. When the QOI is categoric, there are a variety of approaches proposed by researches (Section 2.4). When the QOI can be represented as a scalar field however, Gaussian processes (GP)s are almost unanimously used (Section 2.4.1).

There are many enhancements to classical Gaussian process regression people use to improve their performance in an informative path planning context. These generally involve incorporating prior beliefs about the QOI, learning correlations between other environmental quantities, developing heuristic objective functions, developing multi agent systems and more. Related work in this specific domain will be discussed in Section 2.4.1. Further, GPs and their analytic and informative properties are explored in great detail in Chapter 4.

In accordance with the analytic Bayesian themes of this paper, Chapter 5 provides a novel enhancement to the application of GPs to informative path planning, by placing a conjugate prior over the covariance matrix, giving rise to t processes TPs. This gives the following advantages that cannot be addressed by GPs:

1. TPs scale the covariance of the data in a Bayesian way.
2. TPs are more robust to outliers.
3. TPs give an uncertainty model over the covariance of the QOI.

This chapter starts by exploring the history and limitations of canonical TPs building a foundation of the works of Shah et al. (2014) and Rasmussen et al. (2006). However, the inability to integrate Bayesian sensor models into canonical TPs motivates a new unique stochastic process derived from the normal distribution's semi conjugate prior. Addressing this problem has been explored in few works (Yu et al. (2007) and Roetzer (2020)) each with limited applications. This thesis presents a novel Monte-Carlo mixture model based approach to approximate the posterior beliefs of this stochastic process in a computationally efficient way. It was found, that the covariance of this proposed model is more stable then that of a canonical TP and has the potential to converge to the true variance of the data given the correct parameters.

This thesis further presents a novel approach to training canonical TPs building off the techniques used to train GPs in Rasmussen et al. (2006).

In Chapter 6, a series simple yet principled IPPAs are outlined leveraging each stochastic process. The solutions to each algorithm are approximated using Monte Carlo tree search, a common approach among IPPAs. These IPPAs are then evaluated on a data set of potassium density over New South Wales Australia. The advantages and short comings of each stochastic process are then demonstrated across a variety of statistic metrics.

Figure 1.1 shows how the fields discussed in this thesis are related and where the discussions and contributions of this thesis lie.

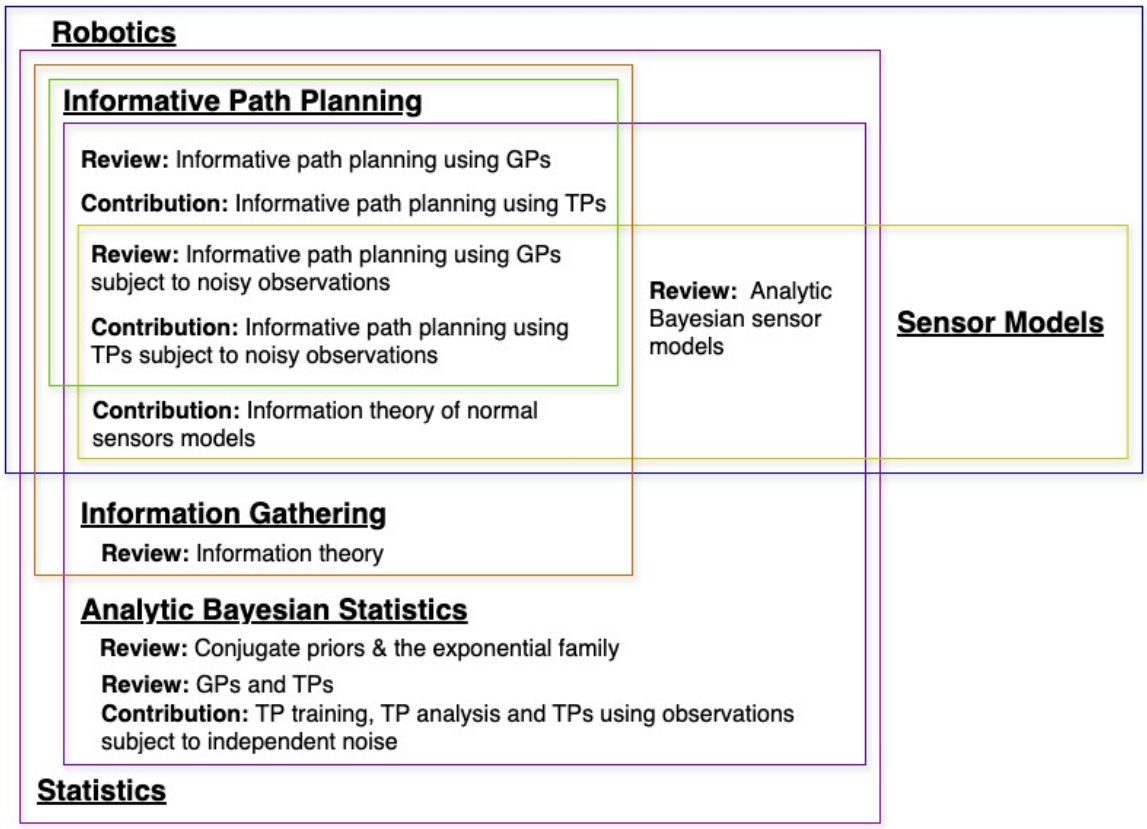


Figure 1.1 – Venn diagram showing the relationship between the fields discussed in this thesis. The discussions and contributions of this thesis are also placed within their respective fields.

1.4 Thesis Outline

- **Chapter 1** - Introduces analytic Bayesian statistics in the context of robotics and informative path planning. The thesis approach, contributions and structure is also outlined.
- **Chapter 2** - Introduces concepts of Bayesian statistics and information theory in order to define the Bayesian informative path planning constraint problem. Related informative path planning works are then discussed.
- **Chapter 3** - Introduces Bayesian sensor models, conjugate priors, the exponential family and their application to robotics and informative path planning.
- **Chapter 4** - Introduces GPs in great technical detail and outlines how the normal distribution's conjugate prior is used to derive noisy GPs.
- **Chapter 5** - Introduces TPs in great technical detail and outlines how the covariance's conjugate prior is used to derive the model. The limitations of noisy TPs are then discussed and addressed.
- **Chapter 6** - Integrates the models discussed in this thesis into their respective IPPAs. The models are then evaluated on a geographic data set.

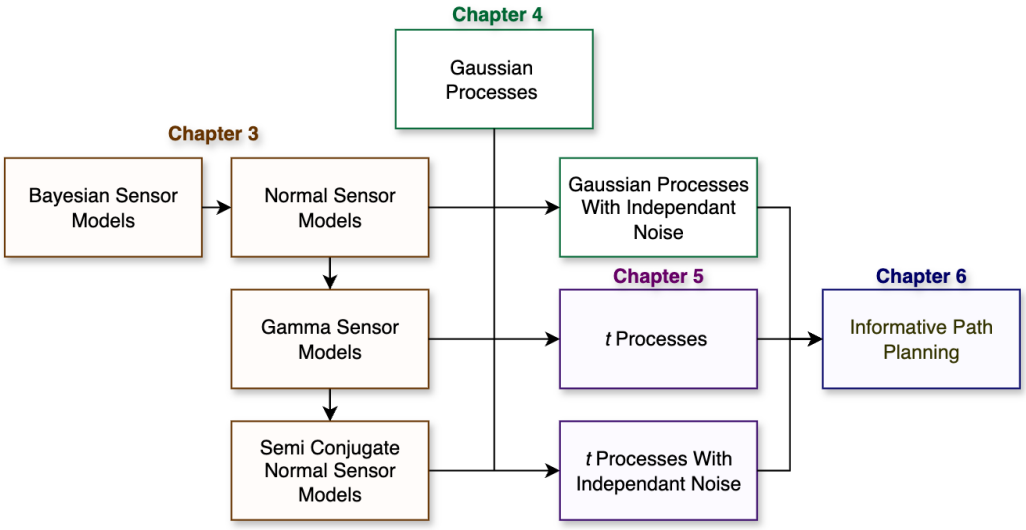


Figure 1.2 – Flow chart showing how the technical elements of this thesis are related.

Chapter 2

Informative Path Planning Background and Related Work

This Chapter will discuss the Bayesian structure of informative path planning algorithms (IPPA)s. Of particular interest in regards to this thesis is how uncertainty about the quantity of interest is quantified and how this motivates the utility function of information gathering algorithms.

2.1 Uncertainty About the Quantity of Interest

The following will be used consistently throughout this thesis:

- Let X denote the quantity of interest (QOI) considered within the full scope of the information gathering problem.
- Let $x_i \in X$ denote the QOI at index i.

Index i often corresponds to a location in the environment or a point in time. In the applications discussed in Chapter 6, i corresponds to a point in a two dimensional region. In the derivations discussed in Chapters 4 and 5 it more generally corresponds to the independent variable on the horizontal axis.

The QOI is the variable researchers wish to gain information about. It may be continuously present across the environment (e.g. temperature, air pressure or presence of gold) or may be only present in discrete locations (e.g. tree girth, battery life or plant type). It may refer to a continuous quantity or may be categoric.

As the objective of IPPAs is to reduce uncertainty about the QOI, it is crucial to quantify this metric. Building of principles of encoding theory (MacKay (2003))[1], entropy is used to quantify uncertainty about the true state of X. For example the uncertainty of tree girth considered across all trees in the forest of interest.

Interestingly, independent of encoding theory, entropy can also be derived axiomatically. Csiszár (2008) shows this by stating the desired postulates of a quantification of uncertainty, then derives it's algebraic form.

In an informative path planning context, most applications tend to quantify uncertainty about the QOI via the entropy of a probability distribution over the QOI (see related work Section 2.4). The following are different forms of entropy that are used depending on the nature of the QOI and it's relationship to the environment. $H(X)$ will denote the entropy of X in all cases. The cardinality of X and x_i's state space will prompt the type of entropy required.

Shannon Entropy

Shannon Entropy quantifies uncertainty about categoric variables that may be one of m categories. It is a function of a probability mass function (PMF) over the QOI. When the natural log is used (Equation 2.1), the unit of entropy is nats, when log base 2 is used entropy is quantified in bits.

$$H(x) = -\sum_{j=0}^{m} p(x_j) \ln(p(x_j)) \tag{2.1}$$

Where x_j denotes the j^{th} category of x.

[1] Chapter 4.

Differential Entropy

Differential Entropy quantifies uncertainty about continuous variables. It is a function of a probability distribution function (PDF) over the QOI. Unlike Shannon Entropy, it does not have meaning directly rooted in encoding theory, but is rather a continuous analogues of Shannon Entropy, as can be seen by its formulation. Differential Entropy may be negative and changes depending on the parametrization of the QOI.

$$H(x) = -\int p(x)\ln(p(x))\,dx \tag{2.2}$$

Entropy Rate

Entropy Rate quantifies uncertainty about a variable over some continuous latent space (often physical space or time). It is frequently used to quantify the entropy of a stochastic process that returns a PDF about x at each point in some continuous latent space (Golshani (2015)).

- Let T be a continuous bounded latent space.
- Let x_i be the QOI at $t_i \in T$.

Then the Entropy rate over T is defined as:

$$\begin{aligned} H_n(X) &= H(x_1, x_2, ..., x_n) \\ H(X) &= \lim_{n\to\infty} \frac{1}{n} H_n(x) \end{aligned} \tag{2.3}$$

2.2 Informativeness of Research Actions and Observations

The following will be used consistently throughout this thesis:

- Let $a_i \in A$ be an observation action made about x_i, where A is the space of possible observation actions.

- Let z_i be the space of possible observation outcomes when making an observation about x_i.
- Let $\check{z}_i$ be an observation yielded from exercising observation action a_i. Note $\check{z}_i$ is a particular instance of z_i.

If the same observation action a_i is exercised more than once, multiple observed values ($\check{z}_i$) may be returned due to sensor noise. Thus additional notation is required to differentiate multiple observations about the same x_i. In cases where the sensor is assumed to have no error or observations cannot be repeated, this notation is not necessary.

- Let $z_i^{[j]} = \check{z}_i^{[j]}$ indicates the j^{th} observation about x_i has value $\check{z}_i^{[j]}$.
- Let $z_i^{[1:n]} = \check{z}_i^{[1:n]} \implies \{z_i^{[1]} = \check{z}_i^{[1]}, ..., z_i^{[n]} = \check{z}_i^{[n]}\}$.

Again, the purpose of making observations in the context of information gathering is to reduce the uncertainty about X. Though as alluded to, sensor noise often results in incomplete information. This motivates the quantification of the informativeness of an observation.

The following will be defined over continuous variables. However, analogous forms exist for the different types of entropy defined in Section 2.1. The below will also exclude serial notation $[j]$, as the below are defined for only one observation of x_i.

Entropy of a Posterior Distribution

When an observation action (a_i) is exercised it yields an observed value ($\check{z}_i$). The conditional entropy of X given observed value $\check{z}_i$ is written as $H(X|z_i = \check{z}_i)$:

$$H(X|z_i = \check{z}_i) = -\int p(X|z_i = \check{z}_i) \ln(p(X|z_i = \check{z}_i))\, dx \tag{2.4}$$

Note, this is simply the entropy of the posterior distribution $p(X|z = \check{z}_i)$.

Conditional Entropy

Before an observation action (a_i) is exercised, the observation outcome ($\check{z}_i$) is unknown, thus the entropy of the posterior (Equation 2.4) is unknown. However, given the required probabilities the expected entropy of the posterior distribution can be solved. This defines the conditional entropy between X and z_i denoted $H(X|z_i)$:

$$\begin{aligned} H(X|z_i) &= -\int p(z_i) H(X|z_i = \check{z}_i)\, dz_i \\ &= -\iint p(X, z_i) \ln(p(X|z_i = \check{z}_i))\, dX dz_i \end{aligned} \tag{2.5}$$

Informativeness

The informativeness of an observed value $\check{z}_i$ denoted $I(X; z_i = \check{z}_i)$ quantifies the amount of information gained about X (reduction in uncertainty) given observation outcome $\check{z}_i$:

$$I(X; z_i = \check{z}_i) = H(X) - H(X|z_i = \check{z}_i) \tag{2.6}$$

Mutual Information

As in defining conditional entropy (Equation 2.5), before an observation action (a_i) is exercised, the observation outcome ($\check{z}_i$) is not known. Again however, given the required probabilities, the expected informativeness can be solved. This defines the mutual information between X and z_i denoted $I(X; z_i)$:

$$I(X; z_i) = H(X) - H(X|z_i) \tag{2.7}$$

Mutual Information is always positive, that is to say observations are always expected to be informative (MacKay (1992))[2]. However, the informativeness of an observation (Equation 2.6) may still be negative. This may be the case given a surprising observation, suggesting less is known about X than previously thought (Lindley (1956))[3].

[2] Equation 8.8.
[3] Discussion of Theorem 1, Page 991.

2.3 Informative Path Planning

Suppose there is some QOI X distributed across an environment consisting of many observable components x_i. To reduce uncertainty about the QOI, observation actions (a_i) may be exercised giving observed values $(\check{z}_i)$.

Informative path planning applies to an agent in the environment capable of exercising observation actions. However, there are almost always some constraints, usually in the form of time and energy, limiting the agent's ability to exercise these actions. A research mission[4] will refer to the process of an agent exercising observation actions until its budget has been exhausted. To motivate the canonical informative path planning constraint problem, the following are defined, and will be used consistently throughout this thesis:

- Let $A_n = \{a_1, ..., a_n\}$ be a sequence of observation actions.
- Let B be the budget of the research mission.
- Let $C : A_n \to \mathbb{R}^+$ be the cost of exercising a sequence of observation actions.

With all the above defined in this chapter, the most informative course of actions can be expressed as:

$$\max_{C(A_n)<B} I(X; Z_n) \tag{2.8}$$

Given $m < n$ observation actions have been exercised, the subsequent most informative actions can be recursively defined. At a high level Murphy (2012)[5] defines this class of problem as a *sequential decision theory problem*. This may also be defined as an orienteering Problem, and in many cases a sub-modular or correlated orienteering problem(**?**.

Once the informative path planning constraint problem has been properly defined, (IPPA)s[6] define a solution to the constraint problem. As the action space grows

[4] Some works refer to this as a campaign
[5] Section 5.7.3.3.
[6] Sometimes referred to as policies.

exponentially, heuristics or analytic policies are almost always used in practice. Online or adaptive IPPAs update the action sequence considering information collected throughout the research mission. Generally IPPAs follow the below structure (Figure 2.1):

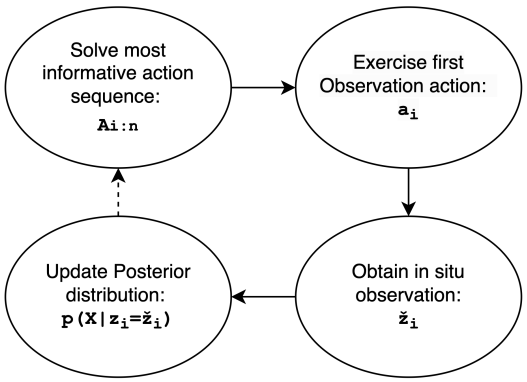


Figure 2.1 – IPPA flow chart. The dotted line indicates the next iteration of the cycle.

The right two bubbles of Figure 2.1 are primarily operational steps, and though difficult engineering problems, they will not be the focus of this thesis. The left two bubbles are the statistic and informatic steps of the algorithm, these will be the primary topics of this thesis.

When using a Bayesian framework, these steps generally depend on the following algebraic components:

- The prior $p(X)$ defines the beliefs about X before the informative path planner has begun. This may be an uninformative prior (Murphy (2012))[7] under circumstances with no prior information.
- Sensor model $p(z_i|x_i)$. Note, $p(x_i|z_i)$ is almost never given, as z_i can rarely be fixed as x_is are sampled. Further, it goes against canonical model structures.
- Probability update step $p(X|x_i)$ or $p(X|z_i)$. This is typically given by a stochastic or machine learning model. Prior information about the correlation and behaviour of X is often embedded in these models.

[7] Section 5.4.

- Mutual information $I(X; z_i)$ determines which observation actions are likely to be most informative. This is essentially the utility function[8] of the IPPA.
- Tree search over the action space (A) returning the most informative action sequence while adhering to the agent's operational constraints.

Autonomous Informative Path Planning

Autonomous IPPAs, require minimal intervention before and during a research mission. This invokes the need for robust IPPAs that can adapt to unknown environments and unexpected observations. These are often related to online or active learning algorithms taken from the field of machine learning.

2.4 Related Work - Applications of Informative Path Planning Algorithms

To begin with, consider a naive IPPA. This usually takes the form of a coverage or loan mower policy. When there is little concern for operational constraints or coverage is a priority, this is an appropriate algorithm. Although naive, this is not always a trivial problem, particularly when observing spatially complex environments or when multiple agents are deployed (Viseras (2018)). This can also be used as a baseline for comparison against other IPPAs (Arora (2018)).

If operational constraints limit the agent's ability to observe the entire environment, prior information and model assumptions may be used to develop a more strategic IPPA. Garza (2021) discusses several settings where the QOI can be determined as a function of multispectral data. Low resolution multispectral satellite data was used to generate a prior distribution over the environment, and in-situ high resolution multispectral data was collected as observation actions were exercised. Under this

[8] Sometimes referred to as an acquisition function in the context of IPPAs

framework, locations where the satellite data offered less information about the QOI became a more informative place for the agent to observe.

In the case where the QOI is a ratio of categories (eg. geologic categories: 20% igneous rock, 80% sedimentary), Furlong (2018) assigns Dirichlet priors (conjugate prior to the multinomial distribution) over the ratio of interest to regions across the environment. The priors determine the mutual information of an observation in each region, guiding the agent throughout the research mission.

Furlong (2018) also considers scenarios where there are multiple sensing modalities on a robot. In the scenario he explores, there is a trade-off between sensor fidelity and energy cost. To address this problem, changes in the environment were detected using a low fidelity energy efficient sensor, prompting observations from the high fidelity energy intensive sensor.

In other multi modal scenarios, there is a trade-off of high fidelity low field of view sensors, versus low fidelity high field of view sensors. In a simulated environment Arora (2018) addresses this problem by approximating the most informative action sequence using different tree search algorithms. Observations made from aerial vehicles face a similar trade off, where low fidelity high field of view observations can be made from higher altitudes and vice-versa. Popović et al. (2017) propose a combination of global viewpoint selection and evolutionary optimization techniques in this context.

Following observations, models may be used to propagate beliefs about the QOI across the environment. How the model does this critically effects the most informative action sequence. In settings with multiple QOIs or latent variables, learning correlations between variables in the field often becomes a driving factor of the IPPA. Harrison et al. (2024) uses GP regression on meteorological data to simultaneously model the QOIs and its correlations to other known latent variables in the environment. Similarly, Jackson (2023) learns the correlations of Benthic categories and other environmental variables using an autonomous underwater vehicle. As these examples demonstrate, learning correlations may be an adjacent objective of the research mission and simultaneously a tool to model the QOI.

Quantifying the informativeness of an observation sequence is an analytically challenging task that must often be approximated. It depends on all the elements discussed thus far, though particularly the model used to propagate beliefs about the QOI. In the Benthic example, Jackson (2023) uses the Mahalanobis distance to choose observations that maximize the variety of categories, to best learn the covariance matrix. In the multispectral case Garza (2021) uses the maxima of informativeness to approximate the utility function. Popović et al. (2017) weigh the priority of minimizing the entropy of unclassified areas and the absolute number of unclassified elements in the environment.

Many of the methods discussed above rely on numeric approximations to model the QOI or to determine the next best observation. Alternatively, this thesis focuses on the exponential family and their conjugate priors to construct closed form IPPAs. These outline what is tractably possible and serve as a strong foundation for models that go beyond such forms.

2.4.1 Applications of Gaussian Processes to Informative Path Planning

The IPPAs considered in this thesis apply to when the QOI can be mathematically described as a scalar field. Thus the QOI is numeric and continuously present across the environment. In this case of informative path planning, GPs are almost unanimously used to model beliefs about of the QOI. In short, GPs are an analytic Bayesian regression model driven by the covariance function relating instances of the QOI. In geostatistic applications they are sometimes referred to as kriging models. GPs and their informative properties will be discussed in great detail in Chapter 4.

Among the examples already mentioned; Binney et al. (2010) use GPs to model the salinity of ocean water using measurements collected by an autonomous underwater vessel; Wei and Zheng (2020) use GPs to model the strength of a wifi radio signal using observations collected by an autonomous agent; Hitz et al. (2017) use GPs to model the density of harmful microorganisms using samples collected by an autonomous

surface vessel; Marchant and Ramos (2012) use GPs to model the ambient light in an indoor environment using observations taken from an autonomous wheeled robot; Rayas Fernández et al. (2022) use GPs to model plant greenness as a proxy for plant health, using observations taken from autonomous aerial vehicle.

Some of the properties of GPs that make them so well suited for informative path planning are:

- Their ability to spatially correlate the QOI in the environment.
- Their ability to train model parameters on observed data.
- Their ability to incorporate observations subject to Gaussian noise.
- Their ability to quantify uncertainty of about the QOI.
- Their ability to incorporate prior beliefs about the QOI.
- Their non parametric properties giving model flexibility.

A well cited disadvantage of GPs are their $\mathcal{O}(n^3)$ time complexity. This limits the amount of observations the model may be conditioned on without reaching computational constraints. To address this, Rasmussen et al. (2006)[9] outlines several methods to approximate the posterior distribution of a GP and its likelihood functions while reducing computational demands. Mishra et al. (2018) use sparse GPs to address these computational constraints in a multi robot informative path planning application. However, the majority of works simply limit the amount of observations the agent may make.

Another challenge of applying GPs to informative path planning, is their inability to calculate the mutual information of noisy observations. Thus, the objective function of these IPPA must be approximated. In many applications this is done by discretizing the environment into a finite set (Binney et al. (2010); Wei and Zheng (2020)) thus giving:

$$I(X; Z_n = \check{Z}_n) \approx I(X_q; Z_n = \check{Z}_n) \tag{2.9}$$

[9] Chapter 8

Where X_q is a sufficiently large and representative set of points. This discretization is also often used to define a finite action space ($|A| \in \mathbb{N}$).

Much of the theoretical research in this space aims at finding the best action sequence given the intangible constraints of an infinite action space and an un-analytic objective function.

In cases where GPs are used to models the QOI over a discretized environment (Equation 2.9); Mishra et al. (2018) use an A* based objective function to prioritize near by high variance points; Wei and Zheng (2020) use a Q learning model to approximate the most informative action sequence; Muñoz et al. (2023) use a fast marching square based function to determine the most unexplored areas; Harrison et al. (2024) use the expected improvement for global fit (Lam (2008)) to capture how more observations are required to capture volatile regions and more.

Instead of discretizing the environment, Hitz et al. (2017) discretize a set of paths in continuous space then uses an evolutionary algorithm to shape the paths and maximize information gain. Jakkala and Akella (2024) are able to optimize over a continuous action space by using a converging sampling method driven by the sparse GP optimization function. In a secondary step, a travelling salesman algorithm is used to connect the set of observation actions.

Other unique applications of GPs to informative path planning include the use of non-stationary GPs, allowing for a blend of kernel functions ot be used (Chen et al. (2024)); the use of acquisition functions designed to find the maximum value of the QOI in the environment rather than to reduce uncertainty (Marchant and Ramos (2012)); and the use of GP mixture models to unify the beliefs of multiple agents (Luo and Sycara (2018)).

As discussed in Section 1.3.1, this thesis will focus on the analytic properties of informative Bayesian regression models. From this perspective, why GPs have such favourable properties in the context of informative path planning is explored. Using the same conjugate principals, a novel application of TPs to informative path planning is then proposed as well as a unique Bayesian regression model based on the normal

distribution's semi conjugate prior.

As the emphasis of this research is placed on the application of stochastic processes to informative path planning, when implementing these models into IPPAs, a standard informative objective function, discretized action space and Monte Carlo Tree Search algorithm is used. These will be discussed in greater detail in Chapter 6.

Chapter 3

Application of Conjugate Priors to Normal Sensor Models

Bayesian sensors models play a crucial role within many fields of classical robotics.

This chapter begins by introducing the general Bayesian sensor model and how it relates to Bayes filtering. The Bayesian update step following noisy measurements of a scalar is then demonstrated in more detail.

It is shown how the normal distribution's conjugate priors can be used to make closed-form inferences about the quantity of interest and the sensor model's likelihood function. The informative properties of these distributions are also explored relating the material to information gathering. Lastly, The analytic limits of the conjugate prior are discussed motivating approximation methods used in later chapters.

3.1 Bayesian Sensor Model

Building off the definitions of Chapter 2:

- Let $x \in \mathbb{R}$ represent the QOI.
- Let $z \in \mathbb{R}$ be the observation space about x.
- Let $\check{z} \in \mathbb{R}$ be an observed value.

Using Murphy (2012)'s[1] nomenclature, the probabilities relevant to Bayesian sensor models are:

- $p(x)$ is the prior.
- $p(z = \check{z}|x)$ is the likelihood function[2] of x given that $z = \check{z}$.
- $p(z)$ is the posterior predictive.
- $p(x|z)$ is the posterior distribution.

In a practical setting, typically only the prior $p(x)$ and likelihood model $p(z|x)$ are given. These are parametrized by prior information and the sensor model respectively. The posterior predictive can be solved by the sum rule giving:

$$p(z) = \int p(z|x)p(x)dx \tag{3.1}$$

In many situations Equation 3.1 is intractable. Though, fortunately in many applications it can be ignored as a normalizing constant.

The posterior distribution may be updated using Bayes rule giving:

$$p(x|z = \check{z}) = \frac{p(x)p(z = \check{z}|x)}{p(z = \check{z})} \tag{3.2}$$

[1] Chapter 3.
[2] Also referred to as an the inverse probability MacKay (2003) Section 2.3.

3.2 Bayes Filtering and Kalman Models

Bayes filters are closely related to the general Bayesian sensor model outlined in the previous section (3.1). They are essentially an extension of the Bayesian sensor model to include control actions that change the state of x prompting recursive state estimation. See Thrun et al. (2005)[3] for a detailed outline of this framework.

When there is no control action and x is constant over time, Bayes filtering gives the same framework as a Bayesian sensor model. From this perspective, much of the methods used in Bayes filtering can be applied to the Bayesian sensor model.

In many cases, x and z are variables in separate domains representing different quantities in the world. For example z may be a radar frequency and x may be the location of a target. In such cases a function is required to relate x and z.

$$z = f(x) \tag{3.3}$$

When z is subject to measurement error, the error is propagated through the function manifesting itself as error about x. Generally, Gaussian white noise about z is no longer Gaussian when propagated through f.

$$z = f(x) + \epsilon \;\; st. \;\; \epsilon \sim \mathcal{N}(\epsilon; 0, \sigma_z^2) \tag{3.4}$$

Linear Case - The Kalman Filter

When f is linear and the measurement noise and prior are Gaussian, the Bayesian sensor model can be expressed a Kalman filter estimating a stationary state.

$$z = ax + b + \epsilon \;\; st. \;\; \epsilon \sim \mathcal{N}(\epsilon; 0, \sigma_z^2) \tag{3.5}$$

In this special case, due to the linear properties of the normal distribution, $p(z|x)$, $p(x|z)$ and $p(z)$ are all tractable and normally distributed (Thrun et al. (2005))[4].

[3] Section 2.4.3.
[4] Section 3.2.

Non Linear Case - The Extended Kalman Filter

When f is not a linear function, normal distributions are not preserved throughout the update step as they are in the Kalman filter. Further, in most cases the posterior distribution is intractable.

The Extended Kalman Filter handles this case by approximating f as a linear function using the Taylor series. Typically only the first two terms of the Taylor series are used to give a linear approximation. However, higher order extended Kalman filters also exist (Einicke (2012))[5].

3.3 Conjugate Priors and the Exponential Family

If the Bayesian sensor mode is not intentionally designed to avoid tractability issues, it will likely encounter such challenges. Numeric and computational methods can be used to approximate integrals when necessary. However, these require time and compute resources that often do not scale well to high dimensions (Fink (1997)).

A good place to start when deriving a tractable posterior distribution, is to choose a prior that is naturally conjugate to the likelihood function $p(z|x)$. When this is the case, the prior is termed a conjugate prior. This guarantees the posterior distribution will be a member of the same family as the prior, and will hence also be naturally conjugate to the likelihood function (Raiffa and Schlaifer (1961))[6].

In short, conjugate families are constructed by extracting the likelihood function's parameters and sufficient statistics[7] from the data. If there are a finite number of sufficient statistics to describe the parameters of interest, this permits the dimensionality of the data handled by Bayes's theorem to be reduced. Note this does not necessarily imply tractability, but it guarantees an analytic simplification in the Bayes update

[5] Chapter 10
[6] Theorem 3.2.2.
[7] The statistics of the data that contain all the information required to update the parameters DeGroot (2005) - Section 9.1.

step and for the prior distribution to be closed under sampling. See Fink (1997) for a detailed explanation.

One crucial limitation of this framework however, is that sufficient statistics of finite dimensions only typically exist for distributions in the exponential family (DeGroot (2005))[8]. Fink (1997) extrapolates on this stating that "it is only for these processes that there must exist a set of simple operations for updating the prior into the posterior". This limits viable conjugate priors over the real numbers to just the normal distribution.

Implications for Analytic Informative Path Planning

When designing Bayesian IPPAs, the Bayesian update step $p(X|Z_n = \check{Z}_n)$ is essential in updating beliefs about the QOI and in determining the subsequent best observation. If this step is to be analytic, $P(X)$ must be naturally conjugate to $P(Z_n|X)$.

If $x_i, z_i \in \mathbb{R}$, this limits $p(x)$ and $p(z)$ to being normal distribution. If the domain space of x_i is changed to be different from $\mathbb{R}$, the probability distribution belonging to the exponential family over this domain is the only distribution that guarantees a closed form Bayesian update step. For example if $x_i, z_i \in \mathbb{R}^+$ the corresponding distribution in the exponential family would be $x_i, z_i \sim \mathcal{G}$ where $\mathcal{G}$ is the gamma distribution.

Fink (1997) outlines a "complete compendium of conjugate priors" covering all domains.

[8] Section 9.2.

3.4 Noisy Measurements of a Scalar

Related work

The remaining sections of this chapter largely build on the works Kevin Murphy: *Bayesian Analysis of the normal distribution* Murphy (2007) and *Inferring an unknown scalar from noisy measurements* Murphy (2012)[9]. Gelman et al. (2013)[10] discuss similar Bayesian models in the normal and categoric case with applications to experiment design.

Normal Distribution Parametrized Using Precision

This chapter will express the variance of a normal distribution in terms of precision:

$$\lambda := \frac{1}{\sigma^2} \tag{3.6}$$

Thus:

$$\mathcal{N}(x; \mu, \sigma^2) := \mathcal{N}(x; \mu, \lambda^{-1}) \tag{3.7}$$

This makes for less cumbersome equations.

Gaussian White Noise Sensor

This section will consider the simple case where z is a direct measurement of x. Thus the function relating z to x (Equation 3.3) is simply the identity function. The prior over x will be a normal distribution and the measurement noise will be Gaussian white noise with known precision $\lambda_{z\tau}$. x_τ will be used to denote the true value of x.

$$p(x) = \mathcal{N}(x; \mu_0, \lambda_0^{-1}) \tag{3.8}$$

$$x = z + \epsilon \;\; st. \;\; \epsilon \sim \mathcal{N}(\epsilon; 0, \lambda_{z\tau}^{-1}) \tag{3.9}$$

[9] Section 4.4.2.1.
[10] Chapter 3.

Where μ_0 is the expected value given by the prior, λ_0 is the precision of the prior and $\lambda_{z\tau}$ is the true precision of the sensor's Gaussian white noise. This gives the conditional distribution:

$$p(z|\lambda_{z\tau}, x) = \mathcal{N}(z; x, \lambda_{z\tau}^{-1}) \tag{3.10}$$

Given Equation 3.8 and 3.10 the posterior predictive of z is calculated as (Murphy (2007))[11]:

$$\begin{aligned} p(z|\lambda_{z\tau}) &= \int p(z|\lambda_{z\tau}, x)p(x)dx \\ &= \mathcal{N}(z; \mu_0, \lambda_{z\tau}^{-1} + \lambda_0^{-1}) \end{aligned} \tag{3.11}$$

Combining the above Equations 3.8, 3.10 and 3.11 the posterior distribution is calculated as (Murphy (2007))[12]:

$$\begin{aligned} p(x|\lambda_{z\tau}, z = \check{z}) &= \frac{p(x)p(z = \check{z}|\lambda_{z\tau}, x)}{p(z = \check{z}|\lambda_{z\tau})} \\ &= \mathcal{N}(x; \mu_1, \lambda_1^{-1}) \end{aligned} \tag{3.12a}$$

where:

$$\mu_1 = \frac{\mu_0\lambda_0 + \check{z}\lambda_{z\tau}}{\lambda_0 + \lambda_{z\tau}} \tag{3.12b}$$

$$\lambda_1 = \lambda_0 + \lambda_{z\tau} \tag{3.12c}$$

From the Perspective of Conjugate Priors

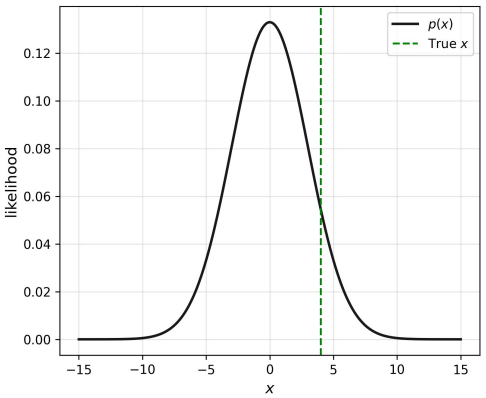
Remark the convenient property shown in Figure 3.1:

$$\left.\begin{aligned} x &\sim \mathcal{N} \\ z|x &\sim \mathcal{N} \end{aligned}\right\} \implies \left\{\begin{aligned} z &\sim \mathcal{N} \\ x|\check{z}^{[1:n]} &\sim \mathcal{N} \\ z|\check{z}^{[1:n]} &\sim \mathcal{N} \end{aligned}\right.$$

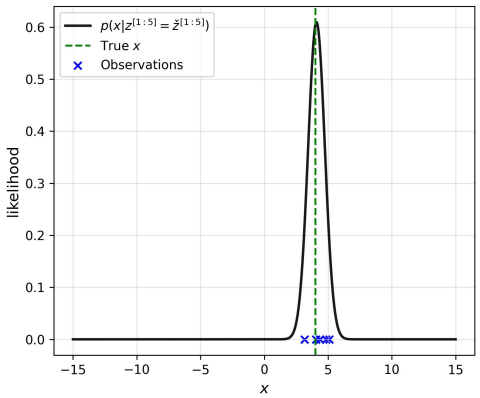
Figure 3.1 – If $p(x)$ and $p(z|x)$ are normally distributed then $p(z)$, $p(z|\check{z}^{[1:n]})$ and $p(z|\check{z}^{[1:n]})$ are all normally distributed.

[11] Equation 36.
[12] Equations 28-30.

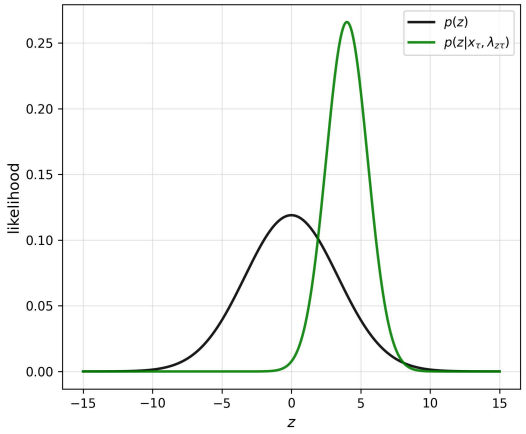
This is because of the unique property that the normal distribution is the conjugate prior to the mean of a normal distribution. From this perspective, the problem may be interpreted as solving for the true mean of the posterior predictive. Figure 3.2 shows how the posterior distributions are updated following five observations.



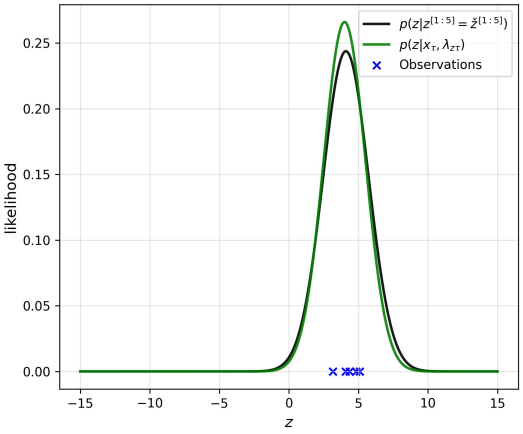
(a) Prior $p(x) = \mathcal{N}(x; 0, 3^2)$ and $x_\tau = 4$.



(b) Posterior $p(x|z^{[1:5]} = \check{z}^{[1:5]})$.



(c) Posterior predictive $p(z|\lambda_{z\tau})$ and true distribution of z, $p(z|\lambda_{z\tau}, x_\tau) = \mathcal{N}(z; 4, 1.5^2)$.



(d) Posterior predictive $p(z|z^{[1:5]} = \check{z}^{[1:5]})$ approaching $p(z|\lambda_{z\tau}, x_\tau)$.

Figure 3.2 – Prior and posterior distributions of x and z using a Normal Sensor model with known variance. It can be seen the posterior predictive $p(z|z^{[1:5]} = \check{z}^{[1:5]})$ approaches the true likelihood model and the posterior distribution ($p(x|z^{[1:5]} = \check{z}^{[1:5]})$) more closely models the ground truth.

3.4.1 Informativeness of an Observation

Before an observation is made, the differential entropy about x provided by the prior is (Michalowicz et al. (2014))[13]:

$$\begin{aligned} H(x) &= \int \mathcal{N}(x; \mu_0, \lambda_0^{-1}) \ln(\mathcal{N}(x; \mu_0, \lambda_0^{-1}))\, dx \\ &= \frac{1}{2} \ln(2\pi\lambda_0^{-1}) + \frac{1}{2} \end{aligned} \tag{3.13}$$

Following an observation, the differential entropy of x provided by the posterior distribution (Equation 3.12a) is:

$$\begin{aligned} H(x|z = \check{z}) &= \int \mathcal{N}(x; \mu_1, \lambda_1^{-1}) \ln(\mathcal{N}(x; \mu_1, \lambda_1^{-1}))\, dx \\ &= \frac{1}{2} \ln(2\pi(\lambda_0 + \lambda_{z\tau})^{-1}) + \frac{1}{2} \end{aligned} \tag{3.14}$$

Taking the difference between Equation 3.13 and Equation 3.14 gives the informativeness $\check{z}$ provides about x:

$$\begin{aligned} I(x; z = \check{z}) &= H(x) - H(x|z = \check{z}) \\ &= \frac{1}{2} \ln(\frac{\lambda_1}{\lambda_0}) \\ &= \frac{1}{2} \ln\left(\frac{\lambda_0 + \lambda_{z\tau}}{\lambda_0}\right) \end{aligned} \tag{3.15}$$

As the informativeness is independent of the observed value $\check{z}$, the mutual information of an observation is $I(x; z) = I(x; z = \check{z})$.

Another unique property of using the conjugate prior to the mean of a normal distribution is that it minimizes the mutual information between z and x. This is true for all parameters and distributions in the exponential family and their conjugate priors (Gutiérrez-Peña and Muliere (2004)).

In an informative path planning context where $x \in X$ is an element of the QOI, Equation 3.15 returns the reduction in uncertainty about the observed instance x.

[13] Section 4.19.

3.5 Solving For the Precision of a Noisy Measurement

Assumptions of this Model

Consider again the basic scenario used in Section 3.4, where z is a direct measurement of x. However, in this case assume the true value of x (x_τ) is given and there is a prior over the precision of the likelihood model λ_z. To serve this Bayesian model, the prior and likelihood function are expressed in relation to λ_z:

$$p(\lambda_z) = \mathcal{G}(\lambda_z; \alpha_0, \beta_0) \tag{3.16}$$

$$p(z|x_\tau, \lambda_z) = \mathcal{N}(z; x_\tau, \lambda_z^{-1}) \tag{3.17}$$

Where $\mathcal{G}$ is the gamma distribution parametrized in terms of shape and rate respectively. This could be used to tune a sensor model's likelihood function in a controlled environment where x_τ is ascertainable.

3.5.1 The Gamma Prior

As the normal distribution is the conjugate prior to the mean of a normal distribution, the gamma distribution is the conjugate prior to the precision of a normal distribution. This makes the gamma prior over λ_z (Equation 3.16) an analytically convenient choice.

Under this parametrization, the posterior predictive $p(z|x_\tau)$ is no longer obtained by integrating over x, as x_τ is given. Rather, λ_z must be integrated over giving (Murphy (2007))[14]

$$\begin{aligned} p(z) &= \int p(z|x_\tau, \lambda_z) p(\lambda_z) d\lambda_z \\ &= \int \mathcal{N}(z; x_\tau, \lambda_z^{-1}) \mathcal{G}(\lambda_z; \alpha_0, \beta_0) d\lambda_z \\ &= \mathcal{T}\Big(z; x_\tau, \frac{\beta_0}{\alpha_0}, 2\alpha_0\Big) \end{aligned} \tag{3.18}$$

[14] Equation 91.

Where $\mathcal{T}$ is the t-distribution parametrized in terms of mean, scale and degrees of freedom respectively. Note, scale is not the same as variance.

Combining the prior $p(\lambda_z)$, the likelihood function $p(z|x_\tau, \lambda_z)$ and posterior predictive $p(z|x_\tau)$, the posterior distribution of λ_z can be calculated as (Murphy (2007))[15]:

$$\begin{aligned} p(\lambda_z|x_\tau, z=\check{z}) &= \frac{p(\lambda_z)p(z=\check{z}|\lambda_z)}{p(z=\check{z})} \\ &= \frac{\mathcal{G}(\lambda_z;\alpha_0,\beta_0)\mathcal{N}(z=\check{z};x_\tau,\lambda_z^{-1})}{\mathcal{T}(z=\check{z};x_\tau,\frac{\beta_0}{\alpha_0},2\alpha_0)} \\ &= \mathcal{G}(\lambda_z;\alpha_1,\beta_1) \end{aligned} \tag{3.19a}$$

Where:

$$\alpha_1 = \alpha_0 + \frac{1}{2} \tag{3.19b}$$

$$\beta_1 = \beta_0 + \frac{1}{2}(\check{z} - x_\tau)^2 \tag{3.19c}$$

Thus updating the posterior predictive to:

$$p(z|z=\check{z}) = \mathcal{T}\left(z; x_\tau, \frac{\alpha_1}{\beta_1}, 2\alpha_1\right) \tag{3.20}$$

From the Perspective of Conjugate Priors

Remark the convenient properties outlined in Figure 3.3:

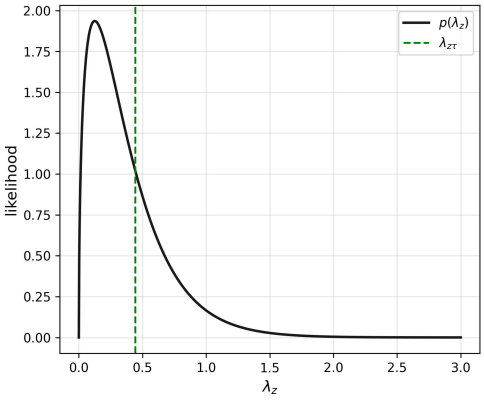
$$\left.\begin{aligned} \lambda_z &\sim \mathcal{G} \\ z|\lambda_z &\sim \mathcal{N} \end{aligned}\right\} \implies \left\{\begin{aligned} z &\sim \mathcal{T} \\ z|\check{z}^{[1:n]} &\sim \mathcal{T} \\ \lambda_z|\check{z}^{[1:n]} &\sim \mathcal{G} \end{aligned}\right.$$

Figure 3.3 – If $p(\lambda_z)$ is gamma distributed and $p(z|x_\tau)$ is normally distributed then $p(\lambda_z|\check{z}^{[1:5]})$ is gamma distributed and $p(z)$ and $p(z|x_\tau, \check{z}^{[1:5]})$ are t distributed.

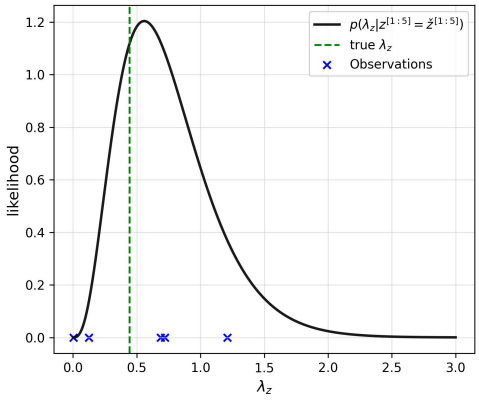
This is because $p(\lambda_z)$ was chosen to be the conjugate prior to the precision of a normal distribution. From this perspective, the problem may be interpreted as solving for

[15] Equation 90.

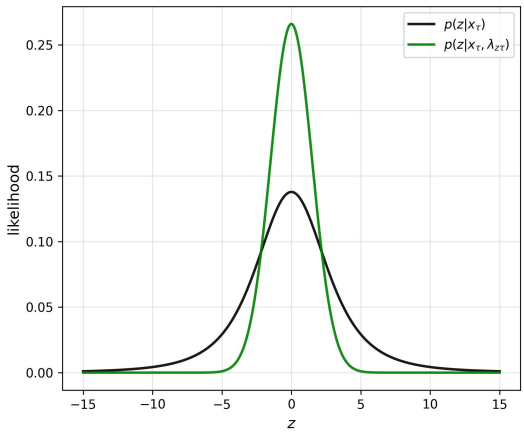
the true precision of the posterior predictive. Figure 3.4 shows how the posterior distributions are updated following five observations.



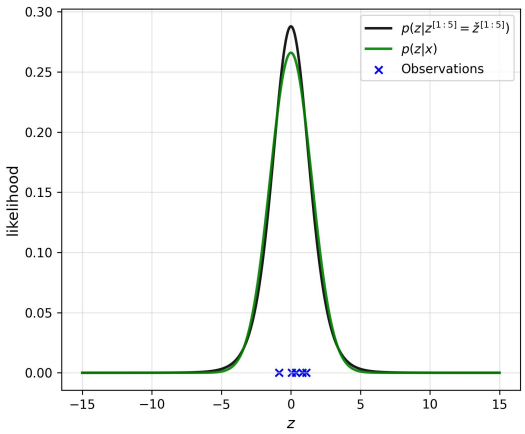
(a) Prior $p(\lambda_z) = \mathcal{G}(\lambda_z; 1.5, 4)$ and $\lambda_z^{-1} = 0.5^2$



(b) Posterior $p(\lambda_z|z^{[1:5]} = \check{z}^{[1:5]})$.



(c) Posterior predictive $p(z) = \mathcal{T}(z; x_\tau, \frac{1.5}{4}, 3)$.



(d) Posterior predictive $p(z|z^{[1:5]} = \check{z}^{[1:5]}))$, approaching $p(z|\lambda_{z\tau}, x_\tau)$.

Figure 3.4 – Prior and posterior distributions of λ_z and z using a Normal Sensor model with unknown precision and given the true value of x (x_τ). It can be seen the posterior predictive $p(z|z^{[1:5]} = \check{z}^{[1:5]})$ approaches the true likelihood model and the posterior distribution ($p(\lambda_z|z^{[1:5]} = \check{z}^{[1:5]})$) more closely models the true sensor noise.

In an informative path planning context, similar methods may be used to train the precision of likelihood models in the field. However, as will be explored in the following section, things become more complicated when the true value of x (x_τ) is not given.

3.6 Noisy Measurements of a Scalar with Unknown Precision

Section 3.4 considers the case where the likelihood function's precision is given and observations are used to reduce uncertainty about the true value of x. Conversely, Section 3.5 considers the case where the true value of x is given and observations are used to reduce the uncertainty about the likelihood function's precision. This section will address both, where neither the true value of x nor the likelihood function's precision is given.

3.6.1 Inference About Joint Probability Distributions

In the case where neither x_τ nor $\lambda_{z\tau}$ is given, the prior is a joint probability over both parameters:

$$p(x, \lambda_z) = p(x|\lambda_z)p(\lambda_z) \tag{3.21a}$$

If x and λ_z are independent random variables, their joint probability can be separated as:

$$x \perp \lambda_z \iff p(x, \lambda_z) = p(x)p(\lambda_z) \tag{3.21b}$$

Continuing with Murphy (2012)[16]'s nomenclature distributions are termed as follows:

- $p(x)$ and $p(\lambda_z)$ are the prior marginal distributions of x and λ_x respectively.
- $p(x|\check{z}^{[1:n]})$ and $p(\lambda_z|\check{z}^{[1:n]})$ are the posterior marginal distributions of x and λ_z respectively.

They can be thought of as marginals of their joint distribution. The marginal distribution of x for example can be calculated as:

$$p(x) = \int p(x|\lambda_z)p(\lambda_z)\, d\lambda_z \tag{3.22a}$$

[16] Chapter 4

or

$$p(x|z=\check{z}) = \int p(x,\lambda_z|z=\check{z})\, d\lambda_z \tag{3.22b}$$

The Bayesian update step in this context is calculated as:

$$p(x,\lambda_z|z=\check{z}) = \frac{p(\lambda_z)p(x|\lambda_z)p(z=\check{z}|\lambda_z,x)}{p(z=\check{z})} \tag{3.22c}$$

3.7 Conjugate Prior and Semi Conjugate Prior of the Normal Distribution

When using Bayesian methods to solve for x_τ and $\lambda_{z\tau}$, using their respective conjugate priors brought convenient analytic solutions (Figures 3.1 and 3.3).

3.7.1 Gamma Normal Conjugate Prior

The conjugate prior of the joint distribution (Equation 3.21a) is the gamma normal distribution ($\mathcal{GN}$). It is defined as follows and gives the following results (Murphy (2012))[17]:

$$\left.\begin{aligned}(x,\lambda_z) &\sim \mathcal{GN}(x,\lambda_z;\mu_0,\kappa_0,\alpha_0,\beta_0)\\ &= \mathcal{N}(x;\mu_0,(\kappa_0\lambda_z)^{-1})\mathcal{G}(\lambda_z;\alpha_0,\beta_0)\end{aligned}\right\} \implies \left\{\begin{aligned} x &\sim \mathcal{T}\\ z &\sim \mathcal{T}\\ \lambda_z &\sim \mathcal{G}\\ x|z^{[1:n]} = \check{z}^{[1:n]} &\sim \mathcal{T}\\ z|z^{[1:n]} = \check{z}^{[1:n]} &\sim \mathcal{T}\\ \lambda_z|z^{[1:n]} = \check{z}^{[1:n]} &\sim \mathcal{G}\end{aligned}\right.$$

Figure 3.5 – If $p(x,\lambda_z)$ is gamma normally distributed then, $p(x)$ is t-distributed, $p(z)$ is t-distributed, $p(\lambda_z)$ is gamma distributed, $p(x|\check{z}^{[1:n]})$ is t-distributed, $p(z|\check{z}^{[1:n]})$ is t-distributed and $p(\lambda_z|z^{[1:n]})$ is gamma distributed.

[17] Section 3

This is an appealing choice as it gives closed-form solutions to all the desired distributions. However, upon closer inspection its structure is problematic in the context of inferring noisy measurements of a scalar. The gamma normal prior requires that:

$$(x, \lambda_z) \sim \mathcal{GN}(x, \lambda_z; \mu_0, \kappa_0, \alpha_0, \beta_0) \implies \text{Var}[x] = \kappa \text{Var}[z] \tag{3.23}$$

Thus, the precision of the likelihood function must be proportional to the precision about x. This is necessary for the marginal distributions to integrate to a t distributions (Murphy (2007))[18].

In practice, this would imply the sensor noise influences uncertainty about x even before any observations are made. Further, in the multidimensional case it implies the respective covariance matrices are proportional to each other. In some contexts there is potential for this to be justified or salvaged, however, that is beyond the scope of this thesis.

3.7.2 Gamma Normal - Semi Conjugate Prior

The semi conjugate prior[19] of the normal distribution is defined in a way that allows for the marginal priors of x and $\lambda_{z\tau}$ to be independent (Murphy (2012))[20]:

$$\begin{aligned} p(x, \lambda_z) &= p(x)p(\lambda_z) \\ &= \mathcal{N}(x; \mu_0, \lambda_0)\mathcal{G}(\lambda_z, \alpha_0, \beta_0) \end{aligned} \tag{3.24}$$

Note the marginal priors $p(x)$ and $p(\lambda_z)$ both follow their respective conjugate priors, hence the name semi-conjugate. However, their marginal posterior distributions are tantalizingly[21] intractable. When applying the semi conjugate prior to stochastic processes in Section 5.4, their distributions are approximated using mixture models and Monte-Carlo methods (Figure 5.7).

[18] Equations 69-72.
[19] Also referred to as the conditionally conjugate prior
[20] Section 4.6.3.2.
[21] possessing a quality that arouses or stimulates desire or interest. Also mockingly or teasingly out of reach Merriam-Webster (2004).

In an informative path planning context it will be shown how the Gamma normal semi conjugate prior may be used to reduce uncertainty about the covariance of the prior and the QOI (Section 5.4).

Chapter 4

Gaussian Processes with Noisy Observations

This section will introduce the Gaussian processes and discuss their Bayesian update step. Continuing in the pursuit of analytic solutions applicable to informative path planning, the conjugate prior to the mean of a multivariate normal will be used to update the model given noisy observations. Further, the informativeness and mutual information of observations will be discussed allowing for easy integration into IPPAs in later chapters.

Gaussian processes (GPs) are a stochastic process used across many disciplines. The models discussed in this chapter are based on the definitions in Rasmussen et al. (2006) and Murphy (2012)[1]. GPs are considered a non parametric model, thus the model will always fit the data. To emphasise this point, the parameters of a GP's kernel function are often referred to has hyperparameters.

There are many aspects of GPs that are researched including kernel function selection, hyperparameter training, applications to classification, approximations for large data sets and more.

This thesis will build off the non-parametric characteristics of GPs by analysing noisy

[1] Section 4.3.

observations and the covariance matrix from a Bayesian perspective. This chapter is largely a review of canonical GPs in preparation for Chapter 5 which puts a prior over the covariance matrix and Chapter 6 which applies these models to informative path planning.

4.1 Gaussian Processes With Exact Observations

4.1.1 Definition

As GPs will be used for informative path planning in later chapters, they will be defined in the context of mapping spatial data. The following notation builds off Rasmussen et al. (2006):

Variable Definitions

- Let t_i represent the i^{th} location in the environment and the prior information known at that location.
- Let x_i represent the QOI at t_i.
- let z_i represent the space of possible observation of x_i.
- Let $\check{z}_i$ represent an observed value at t_i.
- Let T be a continuous space of t_is.
- Let X be a continuous space of x_is indexed by T.
- Let subscript $_q$ of a set denote the $(1:q)$ elements of the set (ie. $X_q = \{x_1, ..., x_q\}$).
- Let subscript $_n$ of a set denote $(1:n)$ elements of the set such that $n \leq q$ (ie. $X_n = \{x_1, ..., x_n\} \subset X_q$).
- Let subscript $_*$ of a set denote the $(n+1:q)$ elements of the set (ie. $X_* = \{x_{n+1}, ..., x_q\}$).

Kernel Function Definition

The kernel function (k) defines the covariance between x_is as a function of their prior information:

$$k(t_i, t_j) = \text{cov}(x_i, x_j) \tag{4.1}$$

Building of the set notation defined earlier, this paper will use the following notation to denote the blocks of a kernel matrix [2].

$$K_{qq} = \begin{bmatrix} K_{nn} & K_{n*} \\ K_{*n} & K_{**} \end{bmatrix} = \left[\begin{array}{ccc|ccc} k(t_1,t_1), & \dots & k(t_1,t_n) & k(t_1,t_{n+1}), & \dots & k(t_1,t_q) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ k(t_n,t_1), & \dots & k(t_n,t_n) & k(t_n,t_{n+1}), & \dots & k(t_n,t_q) \\ \hline k(t_{n+1},t_1), & \dots & k(t_{n+1},t_n) & k(t_{n+1},t_{n+1}), & \dots & k(t_{n+1},t_q) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ k(t_q,t_1), & \dots & k(t_q,t_n) & k(t_q,t_{n+1}), & \dots & k(t_q,t_q) \end{array}\right] \tag{4.2}$$

The kernel function is primarily what defines the behaviour and performance of a GP. The examples in this chapter will use the radial basis function (RBF) kernel and t_i will be a location on the horizontal-axis with no additional prior information. It is a popular kernel due to the smooth model beliefs it gives:

$$k(t_i, t_j) = r^2 e^{-\frac{\|t_i - t_j\|^2}{2l^2}} \tag{4.3}$$

Where $\|.\|$ represents the geometric distance and r and l are the radius and length scale parameters respectively.

Gaussian Process Prior

- Let $\phi : T \to X \;\; st. \;\; \phi(t_i) = \text{E}[x_i]$ which vectorizes as $\phi(T) = \{\text{E}[x_1], ..., \text{E}[x_q]]\}$

With T, X, ϕ and k expressed as a function of T; a GP may be defined. GPs assume every subset $X_q \subset X$ follows a q dimensional Gaussian distribution with mean $\phi(T)$

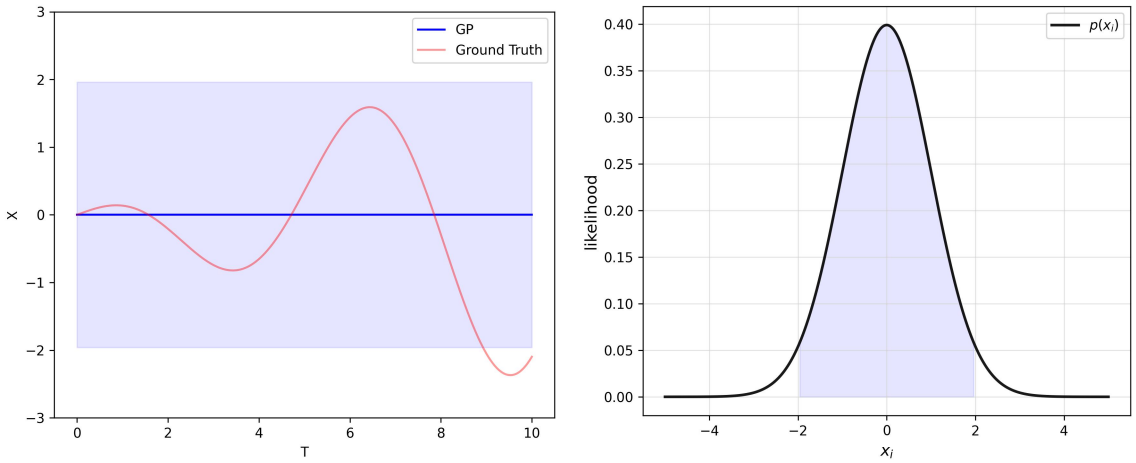
[2] Also referred to as a Gram matrix in this context (Murphy (2012))

and covariance K_{qq}. Although multivariate Gaussian distributions are defined over vectors of a finite dimension, what makes GPs a continuous stochastic process over X is that X_q may be defined so that any $x_i \in X$ may also be in X_q. With all the above, the Gaussian prior over X is (Figure 4.1a):

$$p(X) = \mathcal{N}(X; \phi(T), K_{qq}) \tag{4.4}$$

By taking the marginal distributions of x_i from $P(X)$ (Equation 4.4), the prior for x_i is (Figure 4.1b):

$$p(x_i) = \mathcal{N}(x_i; \phi(t_i), k(t_i, t_i)) \tag{4.5}$$



(a) GP of X over T.

(b) Marginal distribution of x_i at some point t_i (vertical slice at t_i).

Figure 4.1 – GP using an RBF kernel with hyper parameters $r = l = 1$ and prior expectation $\phi(x_i) = 0$. The blue line shows E$[X]$ and the blue shading shows the 95% confidence interval. It can be seen the GP appears continuous over T.

4.1.2 Posterior of Gaussian Process With Exact Observations

Following exact observations of X_n denoted $\check{Z}_n$ ($X_n = \check{Z}_n$), the GP may be updated giving posterior distribution $P(X|\check{Z}_n)$. To solve for this, it is useful to fist express

the prior as:

$$p(X) = \mathcal{N}\left(\begin{bmatrix} X_n \\ X_* \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_*), \end{bmatrix}, \begin{bmatrix} K_{nn} & K_{n*} \\ K_{*n} & K_{**} \end{bmatrix}\right) \tag{4.6a}$$

Because the observations in this case are exact, following observations $\check{Z}_n$, X_n is known and no longer follows a distribution. Thus, $p(X|\check{Z}_n)$ can be expressed as the conditional of the Gaussian Prior $p(X_*|X_n = \check{Z}_n)$ giving (Murphy (2012))[3]:

$$p(X_*|X_n = \check{Z}_n) = \mathcal{N}(X_*; \tilde{\mu}_*, \tilde{\Sigma}_*) \tag{4.6b}$$

where

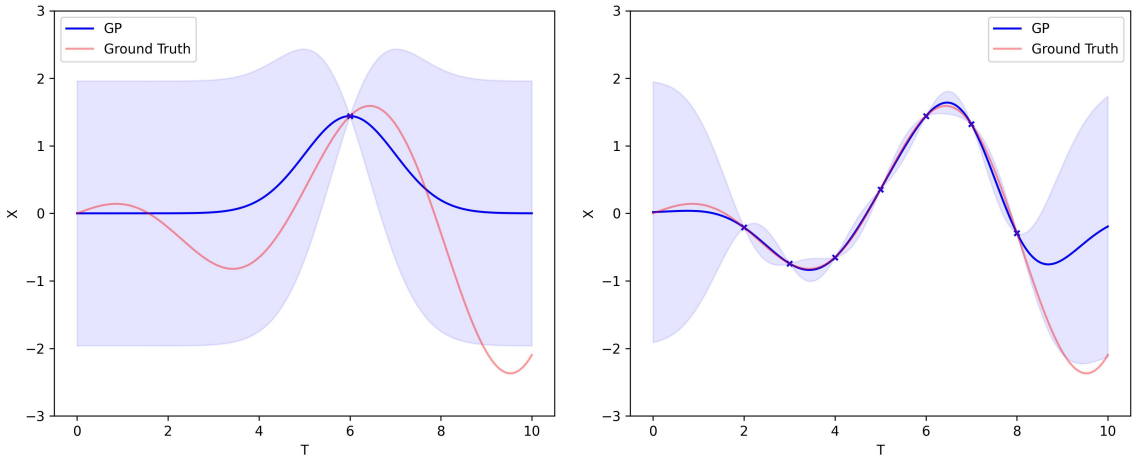
$$\tilde{\mu}_* = \phi(T_*) + K_{*n}K_{nn}^{-1}(\check{Z}_n - \phi(T_n)) \tag{4.6c}$$

$$\tilde{\Sigma}_* = K_{**} - K_{*n}K_{nn}^{-1}K_{n*} \tag{4.6d}$$

Note, $\tilde{\Sigma}_*$ is a Schur complement of K_{qq} with respect to K_{nn} (denoted $K_{q/n}$). Figure 4.2 demonstrates a GP conditioned on exact observations.

The time complexity required to compute the posterior distribution is dominated by obtaining inverse matrix K_{nn}^{-1} requiring time complexity $\mathcal{O}(n^3)$ (Rasmussen et al. (2006))[4].

[3] Equations 15.6 - 15.9
[4] Section 1.2



(a) GP with one point observed at $t = 7$.

(b) GP with points $2 : 9$ observed.

Figure 4.2 – Example of GP conditioned on exact observations (same parameters as figure 4.1).

4.1.3 Informativeness of Exact Observations

To integrate GPs into an IPPA, it is necessary to quantify the informatic properties of the stochastic process.

Uncertainty of a Gaussian Process

There are two approaches to calculating the entropy of a variable that follows a stochastic process, both applicable to GPs.

- Approximation by discretization. This involves simplifying the GP to a multivariate Gaussian Distribution by fixing q. In this case, the entropy of X is the differential entropy of the q-dimensional Gaussian distribution. This can be taken from Arellano Valle et al. (2013)[5] as:

$$\begin{aligned} H(X) &\approx \int \mathcal{N}(X;\mu,\Sigma)\ln(\mathcal{N}(X;\mu,\Sigma))\,dX \\ &= \frac{q}{2}\ln(2\pi e) + \frac{1}{2}\ln(|\Sigma|) \end{aligned} \tag{4.7}$$

[5]Equation 6

Where μ and Σ are the mean and covariance of X's multivariate normal distribution.

- Entropy Rate may also be used to quantify uncertainty across the continuous domain of a GP. This is essentially the limit of the Entropy approximation by discretization as the discretization approaches continuity (Equation 2.3). The Entropy Rate of a GP is given by Feutrill and Roughan (2021)[6] as:

$$H(X) = \frac{1}{2} + \frac{1}{4\pi} \int_{-\pi}^{\pi} \ln(f(\lambda))\, d\lambda \tag{4.8}$$

Where $f(\lambda)$ is the spectral density of the GP. Although entropy rate is a more complete definition of uncertainty, the spectral density of the GP must still be approximated.

Of the two methods, the first which approximates $H(X)$ by $H(X_q)$ is more convenient to work with and sufficiently approximate for the purposes of most IPPAs. The rest of this thesis will use this method to quantify the entropy of a stochastic process over X.

Informativeness and Mutual Information

Unlike the uncertainty of X, the Informativeness of an observation can be calculated exactly. As shown below (Equation 4.9), the part of the equation that involves

[6] Equation 1.

approximation by fixing q cancels out (Arellano Valle et al. (2013))[7]:

$$\begin{aligned} I(X; Z_n) &= H(X) - H(X; Z_n) \\ &\approx \frac{q}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{qq}|) - \left(\frac{q-n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|\tilde{\Sigma}_x|)\right) \\ &= \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln\left(\frac{|K_{qq}|}{|\tilde{\Sigma}_x|}\right) \\ &= \frac{n}{2}\ln(2\pi e) + \ln\left(\frac{|K_{nn}||K_{q/n}|}{|K_{q/n}|}\right) \\ &= \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{nn}|) \end{aligned} \tag{4.9}$$

Where $K_{q/n}$ is the Schur complement of K_{qq} with respect to K_{nn} as given by $\tilde{\Sigma}_*$ (Equation 4.6d). Decomposing $|K_{qq}|$ into these components (Horn and Zhang (2005))[8] results in the desired simplification cancelling out the terms that must be approximated.

Note the actual values of $\check{Z}_n$ do not effect their informativeness about X, thus $I(X; Z_n = \check{Z}_n) = I(X; Z_n)$.

4.2 Application of Conjugate Priors for Noisy Observations

This section will use the conjugate prior to the mean of a multivariate Gaussian distribution to solve for the posterior of a GP given noisy observations. As in Section 3.4, it is assumed observations are direct measurements of the QOI subject to Gaussian white noise.

- Let $\psi(t_i) = \text{Var}(z_i|x_i)$ return the variance of the sensor model at t_i. Note, it is assumed the variance of the sensor at t_i is known.

[7] Section 2.3.
[8] Equation 1.1.4.

As Gaussian white noise implies independence, the multivariate distribution of observations given the true value of X_n is:

$$p(Z_n|X_n) = \mathcal{N}(Z_n; X_n, \Sigma_z) \tag{4.10a}$$

where

$$\Sigma_z = \begin{bmatrix} \psi(t_1) & 0 & ... & 0 \\ 0 & \psi(t_2) & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \psi(t_n) \end{bmatrix} \tag{4.10b}$$

Σ_z is diagonal because it is assumed noise is independent. ψ does not affect the Prior of the GP. However, it will affect the posterior. In most informative path planning applications, independent noise is an appropriate assumption, as prior information about the QOI is assumed to be independent of sensor properties.

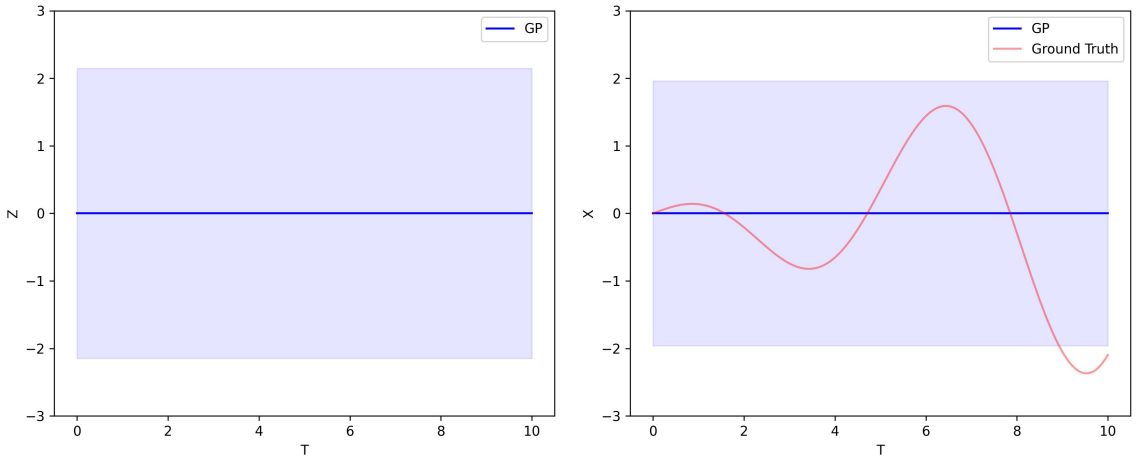
Prior of Gaussian Process With Noisy Observations

Following from Section 3.4 the posterior predictive of Z can be calculated as (Murphy (2007))[9]:

$$\begin{aligned} p(Z_n) &= \int p(Z_n|X_n)p(X_n)dX \\ &= \int \mathcal{N}(Z_n; X_n, \Sigma_Z)\mathcal{N}(X_n; \phi(T_n), K_{nn})dX \\ &= \mathcal{N}(Z; \phi(T_n), k_{nn} + \Sigma_z) \end{aligned} \tag{4.11}$$

Thus, Z is modelled by a GP with the same mean as X, just with wider variance. Z can conveniently be modelled by a GP because the noise function was chosen to be normal, and uniquely the conjugate prior to the mean of a normal distribution is the normal distribution. See the the respective priors in Figure 4.3.

[9] Equation 213.



(a) GP Prior over possible observations. **(b)** GP Prior over QOI.

Figure 4.3 – Comparison of GP prior over noisy observations (left) and the QOI (right). Note the labels on the y-axis and the slightly wider confidence. This was using the same kernel and prior as in Figure 4.1 and constant sensor noise $\psi(t_i) = 0.2$ (This gives 95% confidence interval of errors between $(-0.86, 0.86)$).

Posterior of Gaussian Process With Noisy Observations

To solve the posterior of the of the GP with noisy observations the joint distribution of observations and the QOI is constructed:

$$p\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}; \mathrm{E}[\begin{bmatrix} Z_n \\ X_q \end{bmatrix}], \begin{bmatrix} \mathrm{cov}(Z_n, Z_n) & \mathrm{cov}(Z_n, X_q) \\ \mathrm{cov}(X_q, Z_n) & \mathrm{cov}(X_q, X_q) \end{bmatrix}\right) \tag{4.12}$$

Where the covariance of two vectors is defined to return the covariance matrix comprising of their elements. $\mathrm{E}[Z_n], \mathrm{E}[X_q], \mathrm{cov}(Z_n, Z_n)$ and $\mathrm{cov}(X_q, X_q)$ can be taken form their respective marginal distributions Equations 4.11 and 4.4 respectively. The other blocks of the covariance matrix can be simplified using additive properties of covariance (Equation A.1).

Thus given Equations 4.12, 4.11 and 4.4 the joint prior can be expressed as:

$$p\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_q) \end{bmatrix}, \begin{bmatrix} K_{nn} + \Sigma_z & K_{nq} \\ K_{qn} & K_{qq} \end{bmatrix}\right) \tag{4.13a}$$

Note, X_q is used rather than X_* when observations are note exact. This is because X_n is not given following observations $\check{Z}_n$, thus they must be included in the posterior. As in the exact case (Equation 4.6), the posterior distribution is solved for as as the conditional of the joint distribution:

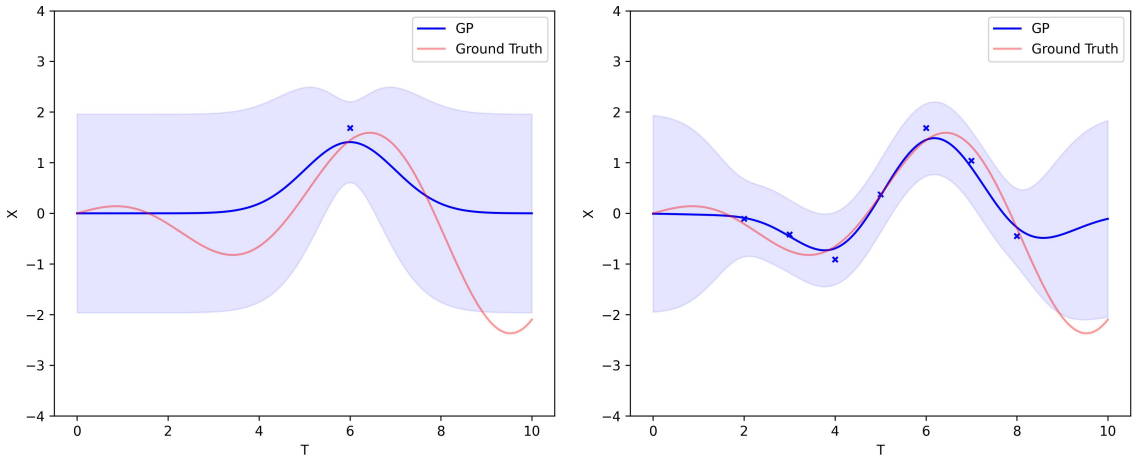
$$p(X_q|Z_n = \check{Z}_n) = \mathcal{N}(X_q; \tilde{\mu}_q, \tilde{\Sigma}_q) \tag{4.14a}$$

where

$$\tilde{\mu}_q = \phi(T_q) + K_{*n}(K_{nn} + \Sigma_z)^{-1}(\check{Z}_n - \phi(T_n)) \tag{4.14b}$$

$$\tilde{\Sigma}_* = K_{**} - K_{*n}(K_{nn} + \Sigma_z)^{-1}K_{n*} \tag{4.14c}$$

See Figure 4.4 demonstrating the posterior of an GP given noisy observations.



(a) Posterior of GP following a noisy observations at point 6.

(b) Posterior of GP following noisy observations at points $1:8$.

Figure 4.4 – GP with noisy observations (same parameters as Figure 4.3).

4.2.1 Informativeness of Noisy Observations

As in Section 4.1.3, uncertainty about X will be approximated by fixing q. As discussed in relation to Equation 4.13a, when using noisy observations, $p(X_n|Z_n = \check{Z}_n) \sim \mathcal{N}$. Thus X_n must also be is included in the entropy of the posterior. For this case it

is more clear to decompose entropy as follows:

$$H(X) \approx \frac{q}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{qq}|) \tag{4.15a}$$

$$H(Z_n) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{nn} + \Sigma_z|) \tag{4.15b}$$

$$H(X, Z_n) \approx \frac{q+n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|\Sigma_J|) \tag{4.15c}$$

Where Σ_J is the covariance matrix of the joint prior used in Equation 4.13a. Using the above equations (Equation 4.15), the mutual information can be written as (Arellano Valle et al. (2013))[10]:

$$\begin{aligned} I(X; Z_n) &= H(X) + H(Z_n) - H(X, Z_n) \\ &\approx \frac{1}{2}\ln\left(\frac{|K_{qq}||K_{nn} + \Sigma_z|}{|\Sigma_J|}\right) \\ &= \frac{1}{2}\ln\left(\frac{|K_{qq}||K_{nn} + \Sigma_z|}{|K_{nn} + \Sigma_z||\tilde{\Sigma}_{qq}|}\right) \\ &= \frac{1}{2}\ln\left(\frac{|K_{qq}|}{|\tilde{\Sigma}_{qq}|}\right) \\ &= \frac{1}{2}\ln(|K_{qq}|) - \frac{1}{2}\ln(|\tilde{\Sigma}_{qq}|) \end{aligned} \tag{4.15d}$$

Where $\tilde{\Sigma}_*$ is the Schur complement of the joint covariance matrix between X_q and Z_n (Equation 4.14c). Because the Z has a different covariance matrix than X, the Schur complements cannot cancel out, thus the choice of q will have an effect on $I(X; Z_n = \check{Z}_n)$.

As in the precise case (Section 4.1.3) the observed values $\check{Z}_n$ do not effect their informativeness. Thus, again the informativeness of $\check{Z}_n$ is $I(X; Z_n = \check{Z}_n) = I(X; Z_n)$.

When integrating noisy GPs into IPPAs, it is inconvenient that the mutual information must be approximated and further involved computationally expensive operations, especially for large q. Section 6.3.1 will outline an approach to minimize this computational overhead.

[10] Equation 3.

4.3 Model Selection and Training

Bayesian probabilistic methods are the focus of this thesis. This entails defining priors, observing values in the observation space and using Bayesian updates to infer posterior distributions. However, other aspects of statistics and machine learning are also important when modelling data in the context of informative path planning. This section will briefly cover the application of parametric optimization to Bayesian models particularly GPs.

GPs offer a non parametric Bayesian approach to modelling the QOI. However, the covariance matrix of the GP still relies on the parameters of the kernel function[11]. Learning the kernel function of data is an extensively researched problem with many approaches. It is often referred to as training the GP (Rasmussen et al. (2006))[12].

To determine how well the GP fits the observed data, loss functions are evaluated using likelihood methods or cross validation. Kernel parameters can then be adjusted to optimized these fit metrics using techniques such as back propagation, grid search or partial derivatives. Learning rates, parameter bounds and other machine learning principals may also be used to prevent overconfidence and over-fitting (Rasmussen et al. (2006))[13].

Bayesian Model Selection

Selecting the parametric-form[14] of the Kernel function is a separate problem. MacKay (1992) outlies a hierarchical Bayesian Framework to select the most suitable kernel function from a finite list. At the lowest level of the hierarchy are *hyperparameters* which define priors over the next hierarchical variable *kernel-parameters*. lastly, a finite set of *model-structures* define a set of parametric kernel functions ($\mathcal{H}$).

Following observations, the posterior predictive likelihoods of $\check{Z}_n$ using different *model-structures* are compared $p(\check{Z}_n|\mathcal{H}_i)$. This involves approximating the integrals over the

[11] Sometimes referred to as hyperparameters
[12] Section 2.3.
[13] Chapter 2.
[14] Also referred to as a model structure

kernel functions' parameters which are often intractable. The *model-structure* that gives the greatest likelihood of $\check{Z}_n$ using priors over the kernel parameters is chosen to best model the data. One interesting property of this method is it naturally encourages model simplicity, reducing over-fitting problems (Rasmussen et al. (2006))[15].

Although this is termed Bayesian Model selection, often this is where the Bayesian methods end. Rarely are posterior distributions over the kernel-parameters calculated resulting in a Bayesian update step about the kernel function.

Kernel Parameter Training

To train the hyper parameters of a kernel function to fit the data, typically the log-likelihood of the observed points (sometimes referred to as the evidence) is maximized. This becomes an arg-max problem (Equation 4.16) that can be approached using standard techniques.

$$\max_{\theta} \ln\Big(p(\check{Z}_n)\Big) \tag{4.16}$$

Where θ are the kernel parameters being optimized.

Algorithm 2.1. from Rasmussen et al. (2006) outlines the most computationally efficient way to approach this, leveraging the Cholesky factorization of $\text{cov}(Z_n, Z_n)$. Building off this, the derivative of the log likelihood may also be computed to leverage gradient based optimizers.

To implement Algorithm 2.1. from Rasmussen et al. (2006), first consider:

$$\text{Let } L := \text{Cholesky}(K_{nn} + \Sigma_z) \tag{4.17a}$$

$$\text{Let } \alpha := L^T \backslash \Big(L \backslash \Big(\check{Z}_n - \phi(T_n) \Big) \Big) \tag{4.17b}$$

[15] Section 5.2.

Then, the log likelihood of observed values can be expressed as (Equation A.2f):

$$\begin{aligned}\ln\Big(p(\check{Z}_n)\Big) &= -\frac{1}{2}(\check{Z}_n - \phi(T_n))^T(K_{nn}+\Sigma_z)^{-1}(\check{Z}_n - \phi(T_n)) - \frac{1}{2}\ln(|K_{nn}+\Sigma_z|) - \frac{n}{2}\ln(2\pi) \\ &= \frac{1}{2}(\check{Z}_n - \phi(T_n))\alpha - \sum_{i=1}^{n}\Big(\ln(l_{ii})\Big) - \frac{n}{2}\ln(2\pi)\end{aligned} \tag{4.17c}$$

Where l_{ii} is the i^{th} the diagonal of L. Building off this, Its derivative may be expressed as (Equation A.2g):

$$\frac{\partial}{\partial\theta}\ln\Big(p(\check{Z}_n)\Big) = \frac{1}{2}\mathrm{tr}\Big((\alpha\alpha^T - (LL^T)^{-1})\frac{\partial(K_{nn}+\Sigma_z)}{\partial\theta}\Big) \tag{4.17d}$$

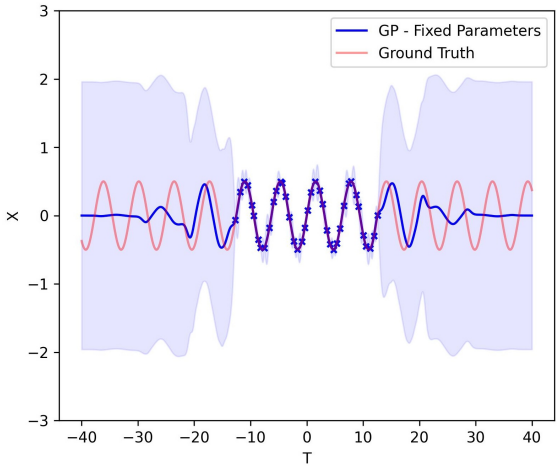
Note Σ_z is included in the derivative giving the option to optimize for the noise parameters as well, deeming it a white kernel.

The time complexity of computing the log likelihood and its derivative is dominated by the need to invert the matrix, or alternatively compute the Cholesky decomposition. Computing the Cholesky decomposition (Equation 4.17a) has complexity $\mathcal{O}(\frac{n^3}{6})$. Subsequently, solving the triangular system (Equation 4.17b) has complexity $\mathcal{O}(\frac{n^2}{2})$. Further computing the derivative has complexity $\mathcal{O}(n^2)$. Thus, there is little additional computational demand when using gradient based optimizers (Rasmussen et al. (2006))[16].

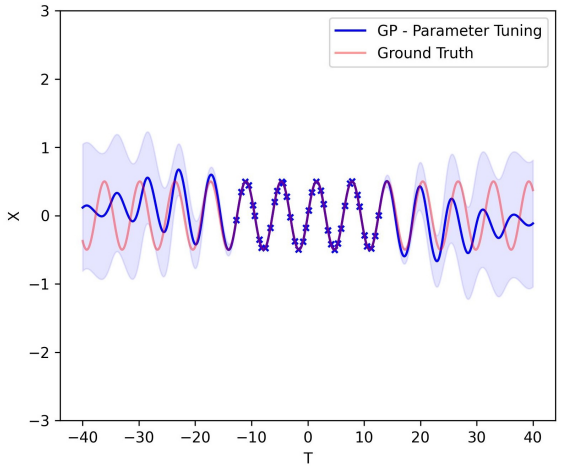
Given Equations 4.17c and 4.17d, the arg-max algorithm used in this thesis, is the BFGS optimizer from the machine learning package Scikit-Learn (Pedregosa et al. (2011)). Further, 5 random kernel parameter values are chosen as starting points. The optimizer is then rerun using each starting to point to reduce the chance of returning suboptimal local maximas.

Figure 4.5 demonstrates how kernel parameter training can be used to fit a GP to a signal. It can be seen the trained kernel parameters using exact or noisy observations yield GPs that better fits the signal.

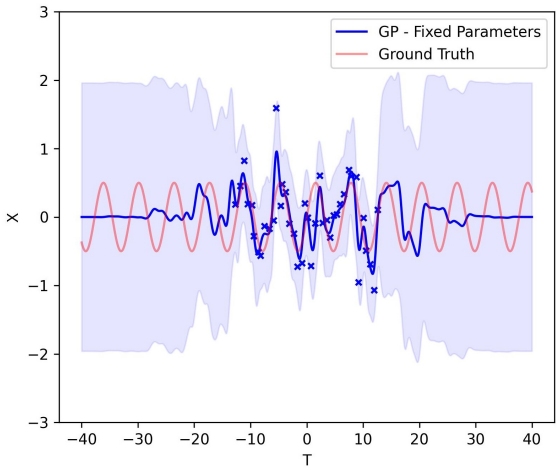
[16]Section 5.4.1. and Algorithm 2.1.



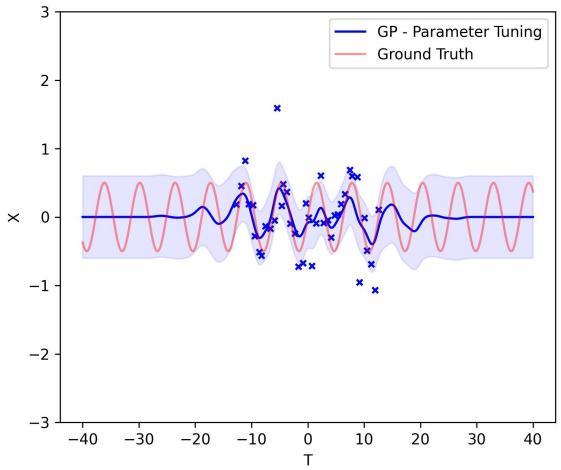
(a) Posterior of GP using incorrect kernel parameters.



(b) Posterior of GP following kernel parameter training



(c) Posterior of GP using incorrect kernel parameters with Noisy observations.



(d) Posterior of GP following kernel parameter training using noisy observations

Figure 4.5 – GP with exact observations and noisy observations with $\sigma_z^2 = 0.2$. Duvenaud (2014) - Chapter 2's locally-periodic kernel was used with parameters $(r = 1, l_1 = 0.5, p = 8, l_2 = 8)$. The left figures do not use any kernel parameter training, while the right figures learn the periodicity and variance of the data. Note, a cap was placed on the RBF length scale (l_2) which determines how local the periodicity of the data is. This limits how much trust is placed on the periodic kernel to limit potential over-fitting. Parameter training was done using Rasmussen et al. (2006) - Algorithm 2.1. with python packages scikit-learn Pedregosa et al. (2011) and Scipy Virtanen et al. (2020). It can be seen the figures on the right appear to better model the ground truth.

Discussion of Kernel Parameter Training Applied to Informative Path Planning

Training kernel parameters can be a useful tool when modelling the QOI across an environment. Depending on the kernel structure, it may change the relative uncertainty about X across the environment effecting which observation actions are expected to be most informative. The expected informativeness of actions may be particularly impacted when parameters are trained to correlate the QOI with known features in the environment (Harrison et al. (2024)).

However, parameter training is not Bayesian. Particularly with few observations, parameters training may lead to volatile sometimes idiosyncratic predictions. Of particular harm to IPPAs are over-confident predictions which may lead to suboptimal observation actions during a research mission with constrained resources.

The parameters of the kernel function should reflect the properties of the QOI and their relationship to features in the environment. With the proper parametric form and training techniques, these relationships can be autonomously learned in field environments (Jackson (2023))[17].

Generally, some element of locality is included to limit the amount of confidence the GP has about unobserved areas. This ensures the IPPA has somewhere to observe and reinforces the intuitive assumption that observations should be somewhat spatially correlated.

Chapter 6 will use the material discussed in this chapter including parameter training to develop an IPPA that uses GPs to model the QOI across an environment. In this application, bounds are placed on the length scale parameter of the kernel function, limiting the undesirable properties of kernel parameter training discussed.

[17] Section 2.7.

Chapter 5

t Processes With Noisy Observations

Motivation and History of t Processes

The performance of canonical GPs are strongly tied to their Kernel functions. Without training the kernel parameters of a GP, the value of observations do not effect the covariance of the posterior distribution, thus limiting how much the model can adapt to observed data. Further, training the kernel parameters poses the disadvantages discussed in Section 4.3.

To address this in a Bayesian rather than a parametric way, a t process (TP)[1] is derived by placing a prior over the covariance matrix of the GP $p(\Sigma_x)$. Following observations, the posterior of this distribution $p(\Sigma_x|\check{Z}_n)$ is updated using a multivariate analogues of the gamma distribution, the distribution used to solve for the precision of a sensor model in Section 3.6.1. When taking a Bayesian approach to updating the covariance matrix, the pitfalls of kernel parameter training such as over-fitting and over confidence are avoided. Further, confidence about the covariance matrix is given and uncertainty about the covariance permeates as uncertainty about X (Figure 5.1).

[1] Sometimes referred to as a student t Process (STP) in other works

TPs where initially dismissed by Rasmussen et al. (2006)[2] because of their inability to analytically handle noisy observations. Rasmussen et al. (2006)'s framework[3] also treats the degrees of freedom as a hyperparameter, limiting $p(\Sigma_x)$'s interpretability as a Bayesian prior.

Zhang and Yeung (2010) rebuts this by adding noise to the diagonals of the kernel function (Section 5.3.2). Later however, Shah et al. (2014) points out it was incorrectly assumed by Zhang and Yeung (2010) that the noise was independent of the TPs degrees of freedom, among other issues. Despite this, Shah et al. (2014) uses this model showing how it acts as a GP that is more robust to outliers. Robustness against outliers is a consistent theme across TP research (Roetzer (2020) [4]; Papadimitriou and Sojoudi (2022); Andrade (2023)).

Wang et al. (2017) builds on the above by scaling the covariance matrix by an inverse gamma prior as opposed to a inverse chi-squared distribution. This gives what they term an *Extended t Process* that allows for more specific priors over Σ_x. The marginals of this process are referred to as *extended t* distributions.

Some approaches have been taken to approximate the posterior distribution of a TP with independent noisy observations or adjacent problems. These will be discussed in Section 5.3 of this chapter.

Chapter Outline

This chapter will begin by deriving the multivariate t distribution using conjugate Bayesian statistics. TP are then derived and their properties interpreted from this perspective. TPs subject to noisy observations are then explored covering some of related work. This thesis then proposes an efficient Monte Carlo sampling method to approximate TPs subject to independent noisy observations. The ideas in this chapter will be tied back to informative path planning an modelling the QOI throughout.

[2] Section 9.9

[3] Rasmussen et al. (2006) scales the covariance by a constant proportional to the inverse chi squared distribution, resulting in a t posterior distribution.

[4] Section 2.2.3.

5.1 Deriving the Multivariate t Distribution

The inverse Wishart Distribution is the conjugate prior to the covariance matrix of a normal distribution (Murphy (2007))[5]. It is analogous to a multidimensional inverse chi squared distribution. Suppose:

$$X \in \mathbb{R}^q \;\; st. \;\; X \sim \mathcal{N}(X; \mu, \Sigma_x) \tag{5.1}$$

$\Sigma_x \in \mathbb{R}^{q \times q}$ is said to follow an inverse Wishart distribution with scale parameter Ψ and degrees of freedom v if:

$$\Sigma \sim \mathcal{IW}(\Sigma_x; \Psi, v) = \frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu q}{2}} \Gamma_q\left(\frac{\nu}{2}\right)} |\Sigma_x|^{-\left(\frac{\nu+q+1}{2}\right)} \exp\left(-\frac{1}{2}\text{tr}(\Psi \Sigma_x^{-1})\right) \tag{5.2}$$

Where $v \geq n-1$, Γ_q is the multivariate gamma function and Ψ is an n dimensional positive semidefinite matrix.

Eaton (2007)[6] shows the diagonals of the inverse Wishart distributions, ie. the variance of $x_i \in X$ follow an inverse-Chi-squared distribution ($\mathcal{IX}$). The inverse-Chi-squared distribution is a special case of the inverse-Gamma distribution such that (Murphy (2007))[7]:

$$\mathcal{IG}\left(x; \alpha = \frac{v}{2}, \beta = \frac{\tau^2 v}{2}\right) = \mathcal{IX}(x; \tau, v) \tag{5.3}$$

Where $\mathcal{IG}$ is the inverse gamma distribution parametrized in terms of scale (α) and rate (β); and the inverse-Chi-squared distribution is parametrized in terms of shape (τ) and degrees of freedom (v).

In the one dimensional case, the more general inverse-Gamma distribution can be used to model more specific priors over variance (Section 3.6). However, the multidimensional analogues of the inverse-Gamma, the non central inverse Wishart is more difficult to work with and is less researched. Most notably, when it is used as the prior to the covariance matrix of a normal distribution, the posterior distribution of

[5] Equation 245
[6] Chapter 8.
[7] Equation 286.

x is not a t distribution Hillier and Kan (2022)[8].

Posterior distribution of x using the Inverse-Wishart Prior

Suppose $X \in \mathbb{R}^q \sim \mathcal{N}(X; \mu, \Sigma_x)$ $st.$ $\Sigma_x \sim \mathcal{IW}(\Sigma_x; \Psi, v)$. Then, the posterior of X can be derived as follows (Murphy (2007))[9]:

$$\begin{aligned} p(X) &= \int p(X|\Sigma_x)p(\Sigma_x)d\Sigma_x \\ &= \int \mathcal{N}(X; \mu, \Sigma_x)\mathcal{IW}(\Sigma_x; \Psi, v)d\Sigma_x \\ &= \mathcal{T}\Big(X; \mu, \frac{\Psi}{v-q+1}, v-q+1\Big) \end{aligned} \tag{5.4}$$

Note, $v \geq q-1 \implies v-q+1 \geq 0$. This ensures the t distribution has a valid parametrization. Also note, the Inverse-Wishart distribution only has mean and covariance when $v > 2$ (Alvarez et al. (2016)[10]), thus for the purposes of constructing a TP the degrees of freedom are parametrized in a way such that $v > 2$.

Also note, the shape parameter of the T distribution is not the covariance of X. This becomes an important detail when parametrizing TPs in terms of the kernel function.

$$X \sim \mathcal{T}(X; \mu, \Psi, v) \implies \text{cov}(X) = \frac{v}{v-2}\Psi \tag{5.5}$$

[8]Equation 16.
[9]Equation 272.
[10]Section 2.1 Paragraph 1

5.2 Deriving the t Process

5.2.1 Deriving the t Process Prior

The definitions of $X, T, Z, \check{Z}, K, k(.,.), \phi(.)$ and $\psi(.)$ hold from the definition of GPs (Section 4.1.1). Degrees of freedom are the only new parameter, parametrized as:

- Let v be the inverse Wishart prior's degrees of freedom.
- Let $v_0 > 2$ be the surplus degrees of freedom of the inverse Wishart prior. $v_0 > 2$ to ensure the distribution has stable mean and mode.
- Let Σ_x continue to be the covariance matrix over X, such that its dimensions are implied by X or the Wishart's shape parameter.

Because the shape parameter of a t distribution is proportional to the covariance of X (Equation 5.5), the inverse Wishart prior must be parametrized to preserve the property $\text{cov}(X) = ker(T, T)$. Giving (Tracey and Wolpert (2018))[11]:

$$\Sigma_x \sim \mathcal{IW}(\Sigma_x; (v_0 - 2)K_{qq}, q - 1 + v_0) \tag{5.6}$$

Note, the degrees of freedom are now parametrized in terms of the dimension of X and the surplus degrees of freedom. Given the prior over Σ_x (Equation 5.6), the posterior $p(X)$ is given as (by Equation 5.4):

$$\begin{aligned} p(X) &= \int \mathcal{N}(X; \phi(T), \Sigma_x)\mathcal{IW}(\Sigma_x; (v_0 - 2)K_{qq}, q - 1 + v_0)d\Sigma_x \\ &= \mathcal{T}\Big(X; \phi(T), \frac{v_0 - 2}{v_0}K_{qq}, v_0\Big) \end{aligned} \tag{5.7}$$

Due to the parametrization of the shape parameter of $p(\Sigma_x)$ and the degrees of freedom, the desired outcome of $\text{cov}(X) = K_{qq}$ is obtained. To solve for the Marginal distribution $p(x_i)$, the marginals of the inverse Wishart and Normal are used. This

[11] Equation 13.

is found to be (Alvarez et al. (2016))[12]:

$$\begin{aligned} p(\Sigma_x) &= \mathcal{IW}(\Sigma_x; (v_0 - 2)K_{qq}, q - 1 + v_0) \\ \implies p(\sigma_i^2) &= \mathcal{IX}\Big(\sigma_i^2; \frac{(v_0 - 2)}{v_0} k(t_i, t_i), v_0\Big) \end{aligned} \tag{5.8a}$$

Where σ_i^2 is the variance of x_i thus the i^{th} diagonal element of Σ_x. By applying the marginal of Σ_x to the marginal of $X|\Sigma_x$ the posterior of x_i is given as (Murphy (2007))[13]:

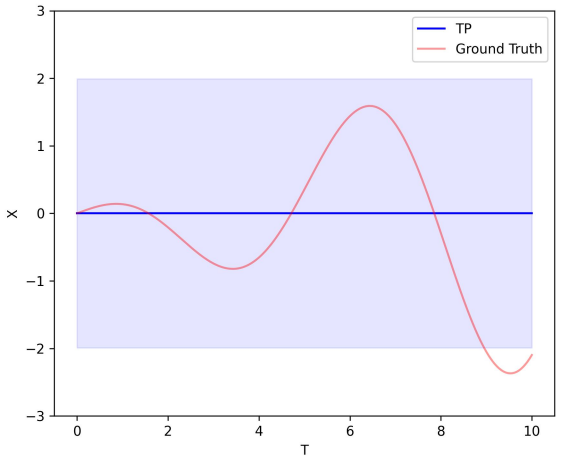
$$\begin{aligned} p(x_i) &= \int p(x_i|\sigma_i^2)p(\sigma_i^2)\, d\sigma_i^2 \\ &= \int \mathcal{N}(x_i; \phi(t_i), \sigma_i^2)\mathcal{IX}\Big(\sigma_i^2; \frac{(v_0 - 2)}{v_0} k(t_i, t_i), v_0\Big) \\ &= \mathcal{T}\Big(x_i; \phi(t_i), \frac{(v_0 - 2)}{v_0} k(t_i, t_i), v_0\Big) \end{aligned} \tag{5.8b}$$

This also maintains the desired property $\text{var}(x_i) = k(t_i, t_i)$ (Equation 5.5). Alternatively, Equation 5.8b can be derived as the marginal of the multivariate t distribution derived in Equation 5.7.

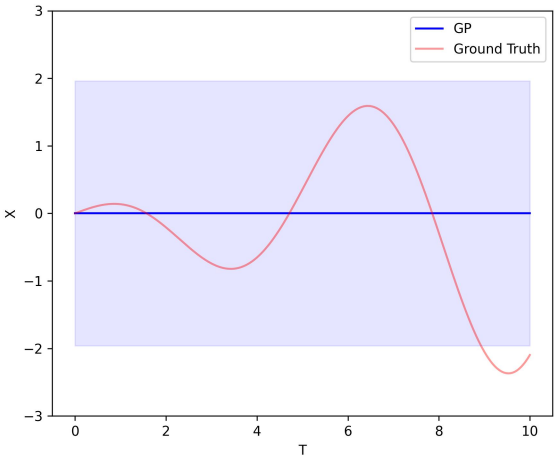
Note, although q affects the parametrization of the latent covariance prior, it does not affect the marginal distributions of x_i. This is a necessary condition when defining any stochastic process the same way as a GP. The properties of the inverse Wishart distribution that make it a conjugate prior to the covariance of the multivariate normal, are the same properties that give the TP its tractable forms. As the degrees of freedom of a TP approach infinity, the prior over variance becomes more exact and the TP approaches a GP (Shah et al. (2014)).

Figure 5.1 demonstrates the difference in TP and GP priors. Note the confidence interval over σ_x^2 corresponding to the TP and the slightly different shape given by the TP's marginals.

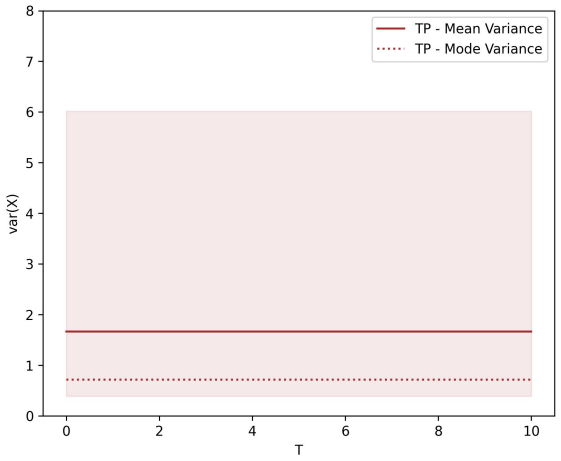
[12] Section 2.1 Paragraph 4.
[13] Equation 167.



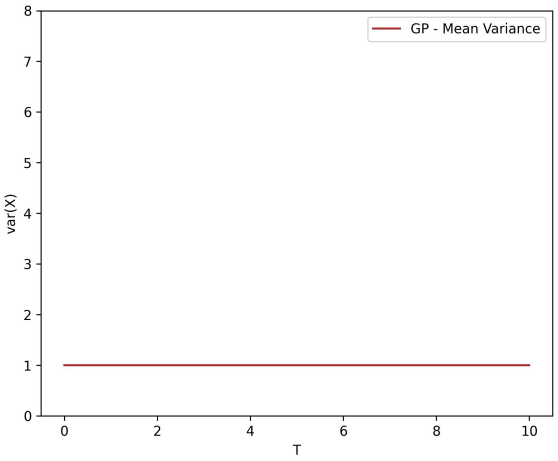
(a) TP prior over x



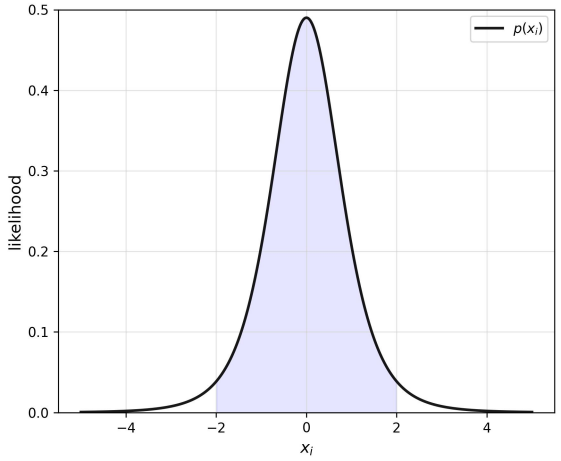
(b) GP prior over x



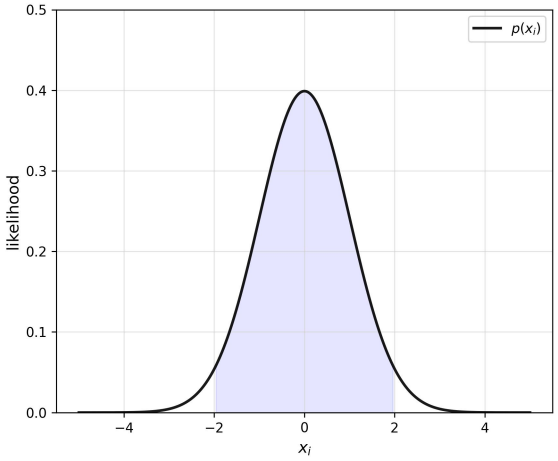
(c) TP Prior over σ_x^2



(d) GP's given variance



(e) TP's marginal distribution at t_i



(f) GP's marginal distribution at t_i

Figure 5.1 – Comparison of TP prior (left) and GP prior (right). The most notable difference is in the second row where TPs have a prior over σ_x^2 and GPs assume σ_x^2 is given. The GP was constructed using RBF kernel with $r = l = 1$. The TP used the same kernel function and $\nu_0 = 5$.

5.2.2 t Process Posterior With Exact Observations

As in the Gaussian case it is beneficial to first express the TP prior (Equation 5.7) as:

$$p(X) = \mathcal{T}\left(\begin{bmatrix} X_n \\ X_* \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_*) \end{bmatrix}, \frac{v_0 - 2}{v_0} \begin{bmatrix} K_{nn} & K_{n*} \\ K_* n & K_{**} \end{bmatrix}, v_0\right) \tag{5.9}$$

When observations are exact, the posterior can be expressed as the conditional of the t distribution $p(X_*|X_n = \check{Z}_n)$ (Ding (2016))[14] giving:

$$\text{Let: } \Psi_{qq} = \frac{v_0 - 2}{v_0} K_{qq} \tag{5.10a}$$

$$p(X|Z_n) = \mathcal{T}(X_*; \tilde{\mu}, \tilde{\Psi}, \tilde{v}) \tag{5.10b}$$

st.

$$\tilde{\mu} = \phi(T_*) + \Psi_{*n}\Psi_{nn}^{-1}(\check{Z}_n - \phi(T_n)) \tag{5.10c}$$

$$\tilde{\Psi} = \frac{S_n + v_0}{n + v_0}(\Psi_{**} - \Psi_{*n}\Psi_{nn}^{-1}\Psi_{n*}) \tag{5.10d}$$

$$\tilde{v} = n + v_0 \tag{5.10e}$$

Where S_n is the Mahalanobis distance with respect to the t prior's shape parameter:

$$S_n = (\phi(T_n) - \check{Z}_n)^T \Psi_{nn}^{-1}(\phi(T_n) - \check{Z}_n) \tag{5.10f}$$

This gives marginal distribution about x_i (by Equation 5.8b):

$$p(x_i|X_n = \check{Z}_n) = \mathcal{T}(x_i; \tilde{\mu}_i, \tilde{\Psi}_{ii}, \tilde{v}) \quad st. \quad x_i \notin X_n \tag{5.11}$$

As with GPs, the time complexity is of conditioning TPs is dominated by obtaining the inverse matrix Ψ_{nn}^{-1} requiring time complexity $\mathcal{O}(n^3)$. Thus, minimal additional compute is required when conditioning TPs. An alternative way of expressing the above Equations (5.10) in terms of K_{qq} can be found in Tracey and Wolpert (2018)[15].

[14] Equation 4.
[15] Equations 14-16.

5.2.3 Scaling Covariance by Mahalanobis Distance

When using exact observations, the mean of X's posterior ($\mathrm{E}[X_*|X_n = \check{Z}_n]$ Equation 5.10c) is the same as when using a GP. However, the covariance of a TP's posterior distribution is different to that of a GP's. Further, the marginals of a GP are normally distributed and the marginals of a TP are t distributed.

The covariance of a TP is proportional to the covariance of a GP (compare Equations 5.10d and 4.6d). This drives the primary difference when modelling data with each stochastic process. When using a TP, the scale of the covariance matrix it is fit using the Mahalanobis distance of the observations with respect to Ψ_{nn} and $\psi(T_n)$ (Equation 5.10f). This is essentially a multivariate z-score. Yu et al. (2007) refers to it as the *explainability* of the observed data. It quantifies by how much the observed data deviates from their prior.

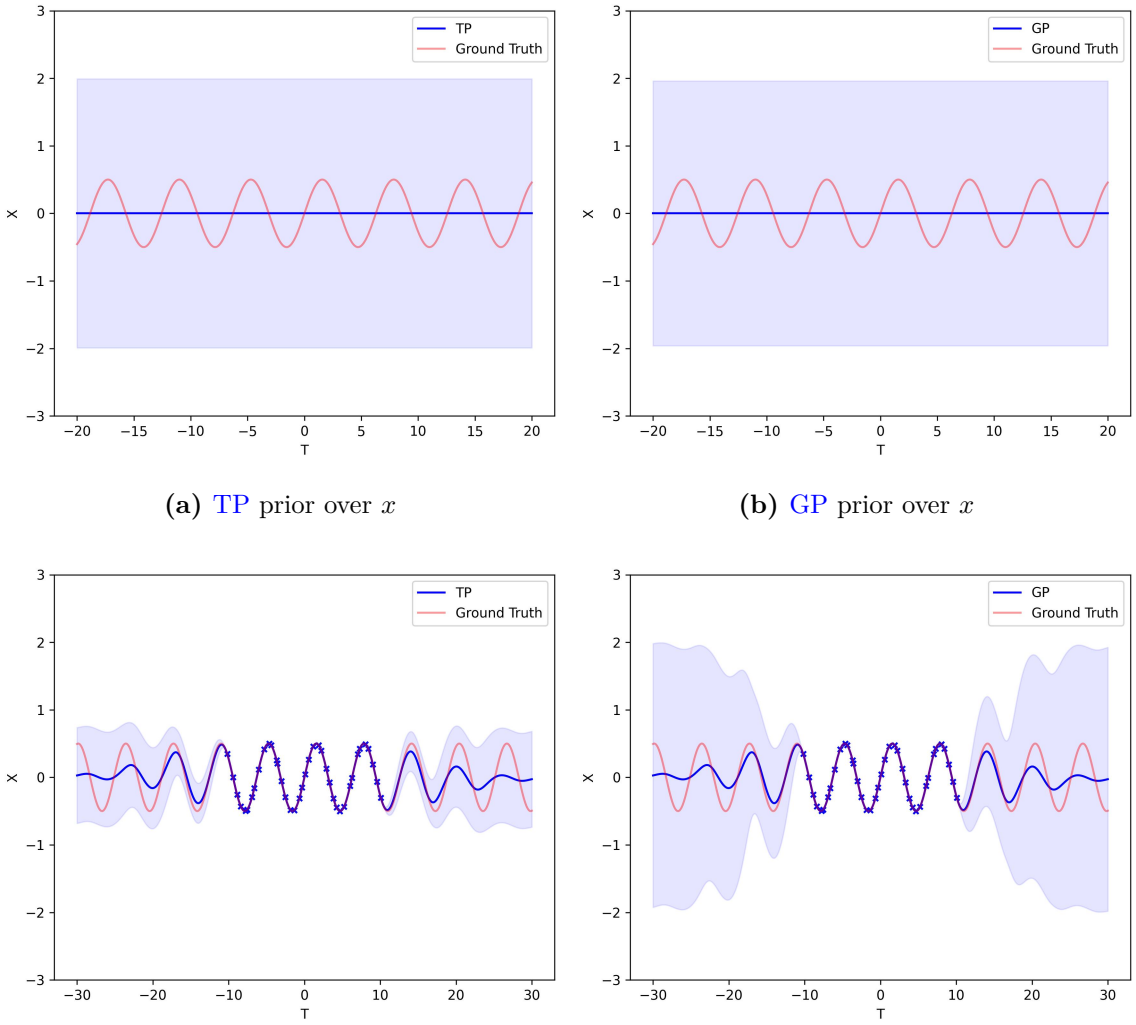
The expected Mahalanobis distance of $X \in \mathbb{R}^n \sim \mathcal{T}(\mu, \Psi, v)$ is n (Tracey and Wolpert (2018)). If the observed Mahalanobis distance is greater than n, that suggests the data was sampled from a distribution with a greater covariance than parametrized by the prior. In response, the TP update step scales up the shape parameter of the posterior t distribution (Figure 5.2). Conversely, if the Mahalanobis is less than n, the shape parameter of the posterior t distribution is scaled down(Figure 5.3) (Shah et al. (2014)).

Lastly remark, the spread of observed values are compared against the prior. Thus TPs are not completely robust to incorrect kernel functions. However, approximately correct kernels and a reasonable amount of observations were shown to still scale the posterior covariance appropriately.

Relevance to Informative Path Planning

This scaling property is a useful tool when applied to informative path planning. The confidence of the prior in many informative path planning cases is difficult to gauge. However, it is a crucial aspect of the problem as it quantifies the uncertainty of

unobserved areas, driving the objective function. TPs partially relieve this sensitivity by scaling the variance of the prior in a Bayesian way. Thus throughout a research mission, not only is the QOI learned, but the confidence of the regression model is also learned. Further, this is done with almost no additional compute.



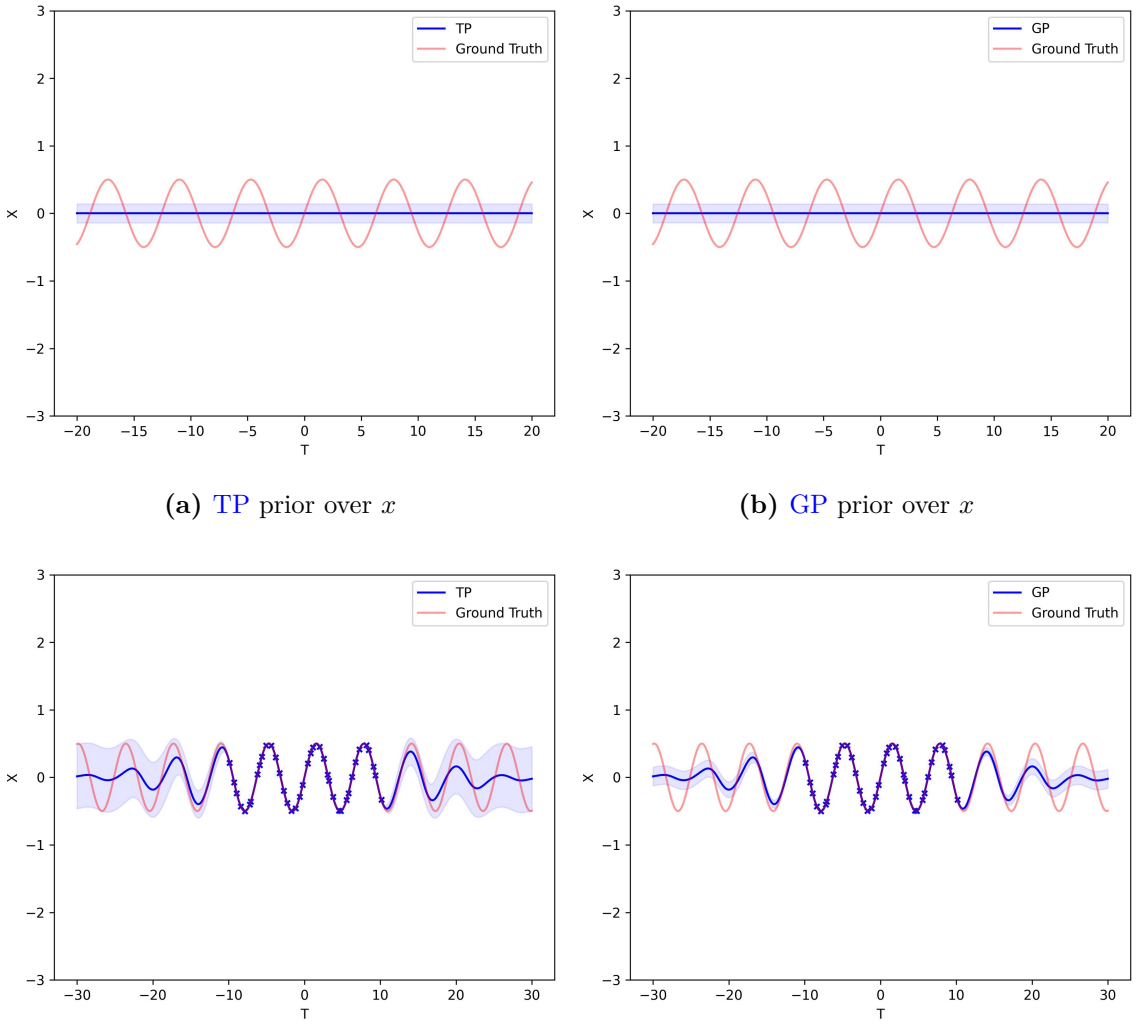
(a) TP prior over x

(b) GP prior over x

(c) TP Posterior over x

(d) GP Posterior over x

Figure 5.2 – Comparison of TP (left) and GP (right) demonstrating the TPs ability to scale the covariance of X. In this case, the prior variance was too large. It can be seen however, the TP adjusts the variance of the model given observed data, making the posterior more confident and giving a better fit to the ground truth. The GP does not scale the covariance, resulting in ultimately the same confidence as the prior away from observed points. Both models use a Duvenaud (2014)'s Local Periodic Kernel with appropriate parameters ($r = 1, l_1 = 1, p = 6, l_2 = 8$) and the TP starts with $v_0 = 5$.



(a) TP prior over x

(b) GP prior over x

(c) TP Posterior over x

(d) GP Posterior over x

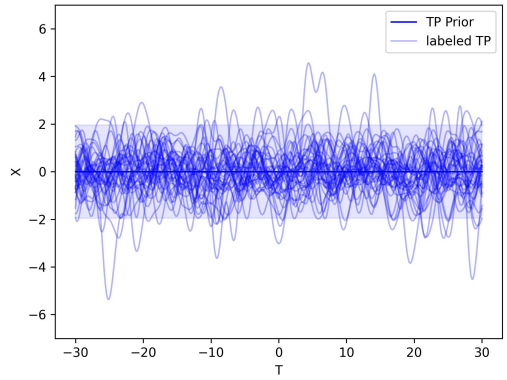
Figure 5.3 – Comparison of TP (left) and GP (right) demonstrating the TPs ability to scale the covariance of X. In this case, the prior variance was too small. It can be seen however, the TP adjusts the variance of the model given observed data, making the posterior less confident and giving a better fit to the ground truth. The GP does not scale the covariance, resulting in ultimately the same confidence as the prior away from observed points. Both models use a Duvenaud (2014)'s Local Periodic Kernel with appropriate parameters ($r = 0.005, l_1 = 1, p = 6, l_2 = 8$) and the TP starts with $v_0 = 5$.

5.2.4 Discussion of t Process Shape

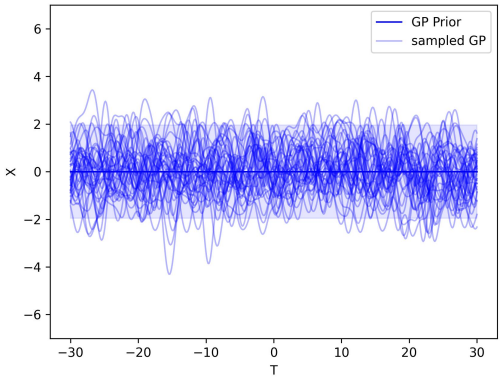
An important property of the resulting marginal t distributions are their robustness to outliers. This is a consequence of t distributions having heavier tails (Shah et al. (2014); Tracey and Wolpert (2018)) thus giving outliers greater likelihoods with respect upper confidence bounds. The effect this has on the shape of the belief model can be visualized by comparing sampled regressions from TP and GP priors (Tracey and Wolpert (2018)). This is done by sampling a q dimensional point from the respective q dimensional prior. Figure 5.4 compares sampled regressions from each prior and sampled points from each marginal distribution. It can be seen that although the priors have the same variance, the regressions and points sampled from the t shaped priors are more volatile and include more outliers.

Relevance to Informative Path Planning

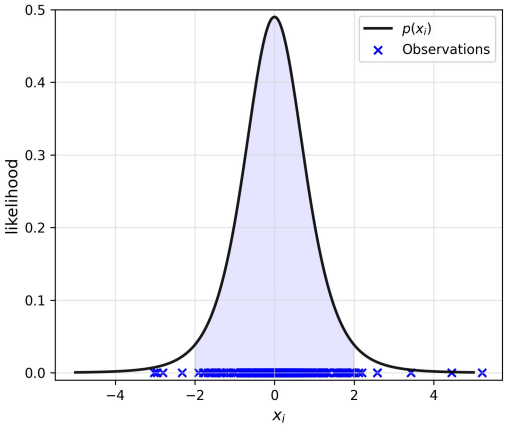
In many informative path planning applications, the QOI does not follow a smooth regression over the latent variable, and further has environmental features resulting in outliers and sharp changes in the data. In these cases, it is advantageous to model the QOI using a TP, as they allocate more likelihood to outliers and irregular regressions given the same kernel function.



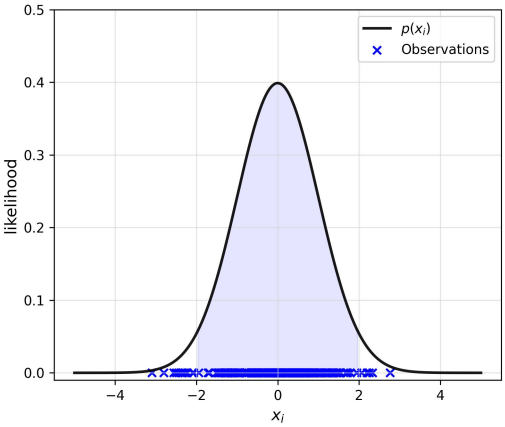
(a) 40 regressions sampled from a TP prior.



(b) 40 regressions sampled from a GP prior.



(c) 300 points sampled from the prior's marginal t distribution.



(d) 300 points sampled from the prior's marginal normal distribution.

Figure 5.4 – Random regressions sampled from a TP prior and GP prior respectively. Both have the same covariance generated from an RBF kernel with $r = l = 1$ and the TP is given degrees of freedom $v_0 = 5$. It can be seen that regressions sampled from the TP prior veer further from the mean as a consequence of the heavier tailed multivariate t distribution.
Similarly, 300 points are sampled from the priors' marginal distribution. It can also be seen that more outliers are sampled from the marginal t distribution.

5.2.5 Informativeness of Exact Observations

Uncertainty about $\boldsymbol{X}$

As with GPs (Section 4.1.3) the uncertainty about X may be approximated by the differential entropy of X_q. In this case the discretized data follows a q dimensional t distribution. Expressing the shape parameter in terms of Ψ (Equation 5.10a) the uncertainty can be approximated as:

$$H(X) \approx \frac{1}{2}\ln(|\Psi|) + H(\mathcal{T}_{q,v_0}) \tag{5.12}$$

Where $\mathcal{T}_{q,v_0}$ is the q dimensional unit t distribution[16] with v_0 degrees of freedom. This uses the property that $H(AX) = |A| + H(X)$ (Cover and Thomas (2006))[17] and $H(\mathcal{T}_{q,v_0})$ can be taken from Arellano Valle et al. (2013)[18].

Informativeness and Mutual Information

Unlike the Gaussian case (Equation 4.9) the values of the observed data will effect the informativeness. This is a consequence of how the Mahalanobis scales the covariance. Further, the informativeness of observed values must be approximated as it is dependant on q:

$$\begin{aligned}
I(X; Z_n = \check{Z}_n) &= H(X) - H(X|Z = \check{Z}_n) \\
&= \frac{1}{2}\ln(|\Psi|) + H(\mathcal{T}_{q,v_0}) - (\frac{1}{2}\ln(|\tilde{\Psi}|) + H(\mathcal{T}_{q-n,v_0+n})) \\
&= \frac{1}{2}\ln\Big(\frac{|\Psi|}{|\tilde{\Psi}|}\Big) + H(\mathcal{T}_{q,v_0}) - H(\mathcal{T}_{q-n,v_0+n}) \\
&= \frac{1}{2}\ln\Big(\frac{|\Psi_{nn}||\Psi_{T/n}|}{c|\Psi_{T/2}|}\Big) + H(\mathcal{T}_{q,v_0}) - H(\mathcal{T}_{q-n,v_0+n}) \\
&= \frac{1}{2}\ln\Big(\frac{|\Psi_{nn}|}{c}) + H(\mathcal{T}_{q,v_0}\Big) - H(\mathcal{T}_{q-n,v_0+n})
\end{aligned} \tag{5.13}$$

[16] Multivariate t distribution with mean the 0 vector and shape the identity matrix.
[17] Theorem 8.6.4.
[18] Equation 7.

Where $H(\mathcal{T}_{q,v_0}) - H(\mathcal{T}_{q-n,v_0+n})$ can be simplified as in Arellano Valle et al. (2013)[19]. and $c = \frac{S_n+v_0}{n+v_0}$ is the scaling factor of the posterior t-distribution's shape parameter (Equation 5.10d).

Unlike informativeness however, the mutual information of exact observations can be solved for in closed form:

$$\begin{aligned} I(X; Z_n) &= H(X) + H(Z_n) - H(X, Z_n) \\ &= H(X) + H(Z_n) - H(X) && \text{By } Z_n \in X \text{ when observations are exact.} \\ &= H(Z_n) \\ &= \frac{1}{2}\ln(|\Psi_{nn}|) + H(\mathcal{T}_{n,v_0}) \end{aligned} \tag{5.14}$$

To be consistent with the entropy of a GP it can be seen that: $v_0 \to \infty \implies H(\mathcal{T}_{n,v_0}) \to \frac{n}{2}\ln(2\pi e)$.

[19] Discussion of Equation 7. ($I_{XY}^{T_{n+m}}$)

5.3 t Processes with Noisy Observations

In Section 4.2 it was shown how the conjugate prior to the mean of a Normal Distribution can be used to construct a GP with noisy observations. In the previous section (5.2) it was shown how the conjugate prior to the covariance of a Normal Distribution can be used to construct a TP. This section will explore how these concepts can be combined to yield a TP with noisy observations.

In Section 3.7 the normal's conjugate prior and semi-conjugate prior were discussed with regards to a Bayesian sensor model that can simultaneously learn the QOI and the precision of the sensor model. Modelling TPs with noisy observations is a multivariate analogous of this. However, uncertainty is added to the covariance of the prior, rather than the likelihood function.

The limiting factor when deriving this, is the algebraic form required to solve the posterior $p(X_q|Z_n = \check{Z}_n)$ in closed-form. Further, the t distribution is not part of the exponential family and does not have a conjugate prior.

5.3.1 Intractable Posterior Distributions

This section will briefly discuss conceptually compelling implementations of TPs with noisy observations. However, each of these cases have tractability issues.

Inverse Wishart Prior Over Σ_z With Given Σ_z

If an inverse Wishart prior is given to Σ_x as in the TP case, and additionally the observation noise is given Σ_z, the joint distribution $p(Z_n, X_q)$ takes the form:

$$p(Z_n, X_q|\Sigma_z) = \mathcal{T}\Big(X_q; \phi(T_q), \frac{v_0 - 2}{v_0} K_{qq}, v_0\Big)\mathcal{N}(Z_n; X_n, \Sigma_z) \tag{5.15}$$

Although $p(Z_n, X_q)$ has an analytic solution, $p(X_q|Z_n = \check{Z}_n)$ is intractable. This is because of the sensitivity of obtaining a gamma function when integrating out Σ_x.

Approximations and related work pertaining to this process are covered in Section 5.4. Note, if conversely Σ_x is given and $\Sigma_z \sim \mathcal{IW}$, the process essentially has the same form as expressed in Equation 5.15.

Inverse Wishart Prior Over Σ_z and Σ_x

If an inverse Wishart prior is given over both Σ_x and Σ_z, the joint distribution $p(Z_n, X_q)$ takes the form:

$$p(Z_n, X_q) = \mathcal{T}\Big(X_q; \phi(T), \frac{v_0 - 2}{v_0} K_{qq}, v_0\Big)\mathcal{T}\Big(Z_n; X_n, \frac{w_0 - 2}{w_0} I_q \psi(T), w_0\Big) \tag{5.16}$$

Where w_0 are the degrees of freedom of used to parametrize $p(\Sigma_z)$. Again, although $p(Z_n, X_q)$ has an anaclitic solution, $p(X_q|Z_n = \check{Z}_n)$ is intractable. This is because the mean of a t distribution does not have a conjugate prior. Yu et al. (2007)[20] discusses the potential of this model describing it as "double robust control" making reference to robust regression (Gelman et al. (2013))[21]. Tang et al. (2017) later approximated this model using the second order Taylor series.

Normal Inverse Wishart Conjugate Prior

The inverse Wishart conjugate prior over $p(Z_n, X_q)$ takes form:

$$p(Z_n, X_q) = \int \mathcal{N}\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_q), \end{bmatrix}, \begin{bmatrix} \kappa\Sigma_{nn} & \Sigma_{nq} \\ \Sigma_{qn} & \Sigma_{qq} \end{bmatrix}\right) \mathcal{IW}\left(\Sigma; \begin{bmatrix} \Psi_{nn} & \Psi_{nq} \\ \Psi_{qn} & \Psi_{qq} \end{bmatrix}, v\right) \tag{5.17}$$

Where κ is a scalar, Σ is a latent covariance variable and subscripts of Σ denote block matrices. In this case $p(X_q|Z_n = \check{Z}_n)$ is intractable for the same reasons as Equation 5.15. Note If $n = q$ the distribution would be tractable giving it its property as a conjugate prior. Lastly, this model requires $\Sigma_z = \kappa\Sigma_x$ which does not hold in most applications.

[20] Section 5.
[21] Section 17.5.

5.3.2 Tractable Solution - The Canonical Method

To the authors knowledge the only tractable approach to constructing a TP with noisy observations, is using the conjugate prior to the covariance of the joint distribution $p(Z_n, X_q)$ as outlined in Shah et al. (2014). This will be referred to as the canonical method in this thesis.

Recall, the joint distribution used by the GP (no prior over Σ_x) is (Equation 4.13a):

$$p\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix} | \Sigma_x, \Sigma_z\right) = \mathcal{N}\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_q) \end{bmatrix}, \begin{bmatrix} K_{nn} + \Sigma_z & K_{nq} \\ K_{qn} & K_{qq} \end{bmatrix}\right) \tag{5.18}$$

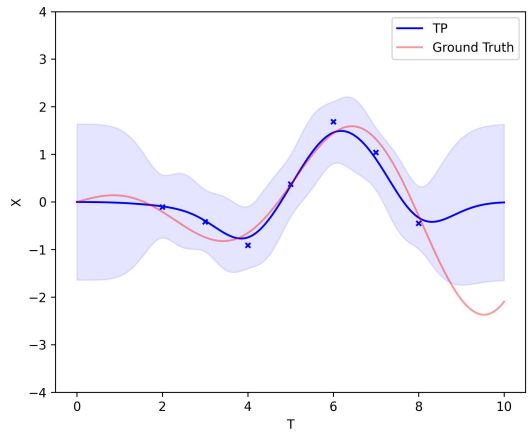
Shah et al. (2014) place an inverse Wishart prior over the joint covariance matrix of $p(Z_n, X_q)$ giving:

$$p(\Sigma) = \mathcal{IW}\left(\Sigma; (v_0 - 2)\begin{bmatrix} K_{nn} + \Sigma_z & K_{nq} \\ K_{qn} & K_{qq} \end{bmatrix}, q - 1 + v_0\right) \tag{5.19}$$

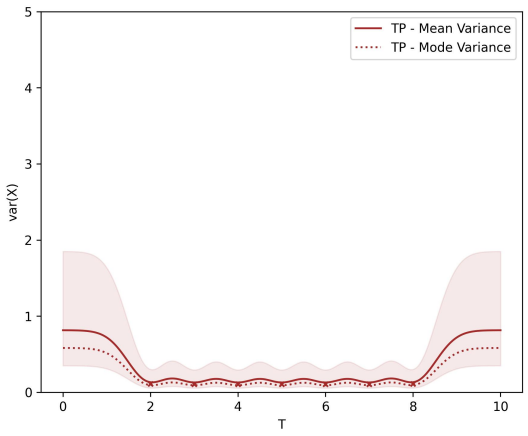
Where Σ is the latent covariance matrix of the joint distribution. This gives joint posterior distribution (by Equation 5.4):

$$p\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}\right) = \mathcal{T}\left(\begin{bmatrix} Z_n \\ X_q \end{bmatrix}; \begin{bmatrix} \phi(T_n) \\ \phi(T_q) \end{bmatrix}, \frac{v_0 - 2}{v_0}\begin{bmatrix} K_{nn} + \Sigma_z & K_{nq} \\ K_{qn} & K_{qq} \end{bmatrix}, v_0\right) \tag{5.20}$$

Under this definition, the same principals of a regular TP follow (Equation 5.10) with the detail that the posterior conditioned on Z_n includes X_n. Figure 5.5 demonstrates the posterior of a TP given noisy noisy observations and the confidence of the posterior's variance.



(a) Posterior of TP following Noisy observations at points $1:8$.



(b) Posterior of covariance of TP following Noisy observations at points $1:8$.

Figure 5.5 – TP using an inverse Wishart prior over the covariance of the joint distribution $p(Z_n, X_q)$ with noisy observations. This was using an RBF kernel with hyper parameters $r = \gamma = 1$, prior expectation $\phi(x_i) = 0$, $\psi(t_i) = 0.2$ and $v_0 = 5$. The same mean, kernel parameters and ground truth are used as in Figure 4.4. This figure can be compared with the GP that model the same data in Figure 4.4. It can be seen the shape of the TP appears to reasonably follow that of a TP subject to noisy observations and the posterior marginal variance (right) is in line with the confidence of the regression model.

Limitation of Prior over Joint Covariance Matrix

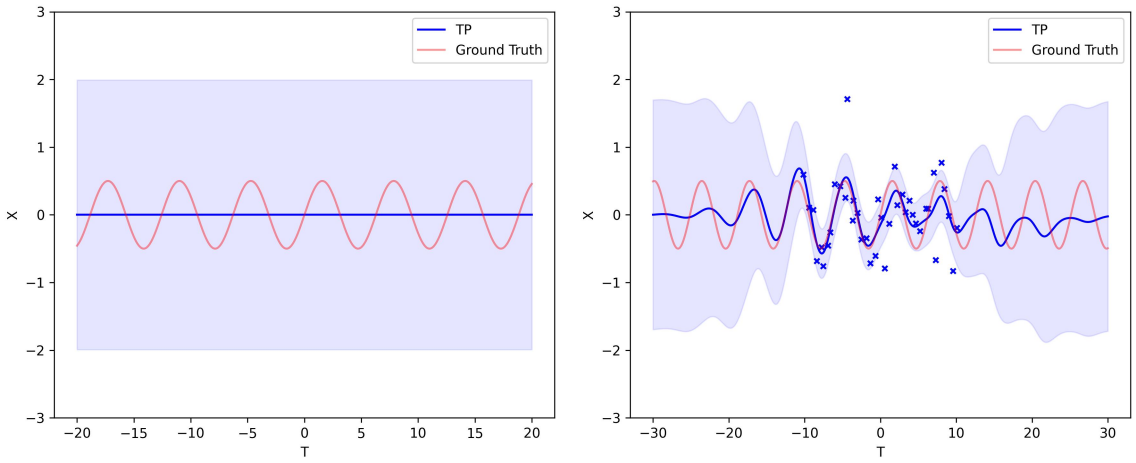
The only data available in the Bayesian update step are noisy observation values ($\check{Z}_n$) and their prior $p(Z_n) = \mathcal{T}(Z_n; \phi(T_n), \frac{v_0-2}{v_0}(K_{nn} + \Sigma_z))$. Observed values are used to calculated the Mahalanobis distance as:

$$S_n = (\phi(T_n) - \check{Z}_n)^T (K_{nn} + \Sigma_z)^{-1} (\phi(T_n) - \check{Z}_n) \tag{5.21}$$

This quantifies how different the observed values are with respect to their prior. However, because there is no distinction in confidence between the priors for Σ_z and Σ_x, there is no way to determine if Σ_z or Σ_x is driving the Mahalanobis distance. Rasmussen et al. (2006)[22] refers to this as *noise entanglement*. Note, it cannot be

[22] Section 9.9.

assumed Σ_z is known otherwise the Bayesian structure of $p(X, Z)$ would take the intractable form described in Equation 5.15.



(a) Prior $p(x)$ showcasing Σ_x

(b) Posterior $p(X|Z = \check{Z}_{40})$ showcasing $\text{cov}(X)|\check{Z}_{40}$.

Figure 5.6 – This plot shows the canonical TPs limited ability to learn the covariance matrix with noisy observations. 40 observations were made with given Σ_z, however, $\text{cov}(x)$ changed very little following the observations. The same kernel was used as in figure 5.2 and $\sigma_z^2 = 0.2$ was used.

Even if Σ_z is correct, it was found scaling the covariance matrix using the Mahalanobis distance of noisy observation will result in a larger posterior covariance. This results in overall less confidence in the posterior distribution (Figure 5.6).

Limitations when applied to Informative Path Planning

When integrating this model into IPPAs, the confidence of the model can scale in unpredictable ways. This can result in poor observation actions as was demonstrated in Section 6.5.

Informativeness of Observation

The informativeness of an observation takes similar form as when using exact observations (Equation 5.13). However, the Mutual Information of observations $\check{Z}_n$ and X can no longer be solved for in closed form. The mutual information can be taken from Arellano Valle et al. (2013)[23] as the mutual information between sets of elements of a multivariate t-distribution:

$$\begin{aligned} I(X; Z_n) &= \frac{1}{2} \ln\left(\frac{|\Psi z + \Psi_{nn}||\Psi_{qq}|}{|\Psi_J|}\right) + f(v, n, q) \\ &= \frac{1}{2} \ln\left(\frac{|\Psi_{qq}|}{|\Psi_{J/(\Psi_z + \Psi_{nn})}|}\right) + f(v, n, q) \end{aligned} \tag{5.22}$$

Where Ψ_J is the shape parameter in Equation 5.20, $\Psi_z = \frac{v_0 - 2}{v_0}\Sigma_z$ and $f(v, n, q)$ is a lengthy function of v, n and q taken from Arellano Valle et al. (2013). Note however, $f(v, n, q)$ will cancel out when comparing an observation sequence sequences of the same length. Also note, the first term takes the same form as the mutual information of a noisy Gaussian process.

[23] Equation 7.

5.4 Approximation of t Process With Independent Noise

To address the limitations of the canonical noisy TP outlined in Section 5.3.2, this section outlines a method of approximating a less conventional stochastic process that preserves the independence of noise.

As discussed in Section 5.3.1 (Equation 5.15) if Σ_z is given, and a prior is given over Σ_x the joint posterior $p(Z_n, X_q)$ does not follow a conventional distribution or have tractable posterior distribution $p(X_q|Z_n = \check{Z}_n)$. The joint distribution given Σ_z can be expressed as:

$$\begin{aligned} p(Z_n, X_q|\Sigma_z) &= \int p(\Sigma_x|\Sigma_z)p(X_q|\Sigma_x, \Sigma_z)p(Z_n|\Sigma_x, \Sigma_z, X_q)\, d\Sigma_x \\ &= \int p(\Sigma_x)p(X_q|\Sigma_x)p(Z_n|X_n, \Sigma_z)\, d\Sigma_x \\ &= \int \mathcal{IW}(\Sigma_x; (v_0-2)K_{qq}, q-1+v_0)\mathcal{N}(X_q; \phi(T), \Sigma_x)\mathcal{N}(Z_n; X_n; \Sigma_z)\, d\Sigma_x \\ &= \mathcal{T}\Big(X_q; \phi(T), \frac{v_0-2}{v_0}K_{qq}, v_0\Big)\mathcal{N}(Z_n; X_n, \Sigma_z) \end{aligned} \tag{5.23}$$

This assumes sensor noise is conditionally independent of the covariance of the QOI ie. $p(\Sigma_x|\Sigma_z) = p(\Sigma_x)$. This process will de termed a t process subject to independent noise in this thesis.

Related Work

Yu et al. (2007)[24] addresses this intractability by expressing the posterior as a mixture of GPs. In cases when the sample number is sufficiently large, the gamma scaling factor used in their parametrization has small variance and is hence approximated by it's mode.

Roetzer (2020) approaches this as a special case of what he terms generalized student t process regression. Under this framework t Laplace approximation and variational

[24] Section 4.3.

student t approximation were used.

This thesis will use a similar method to Yu et al. (2007). However, the gamma distribution will not be approximated by its mode and rather than using a mixture of GPs, samples from the marginal distributions will be used for faster compute.

5.4.1 Approximation by Sampling

Although $p(X_q|\Sigma_z, Z_n = \check{Z}_n)$ does not have a closed form solution, values may be sampled from $p(X|\Sigma_z, Z_n = \check{Z}_n)$. This is done by first sampling a covariance matrix, $\hat{\Sigma}_x \sim p(\Sigma_x)$ then sampling from $p(X|\Sigma_z, \Sigma_x = \hat{\Sigma}_x, Z_n = \check{Z}_n)$. Monte Carlo sampling methods may then be used to approximate $p(X|\Sigma_z, Z_n = \check{Z}_n)$ by repeating this process many times:

$$X \sim p(X|\Sigma_z, Z_n = \check{Z}_n) \implies X \sim p(X|\Sigma_z, \Sigma_x = \hat{\Sigma}_x, Z_n = \check{Z}_n) \;\; st. \;\; \hat{\Sigma}_x \sim p(\Sigma_x) \tag{5.24}$$

$p(X_q|\Sigma_z, \Sigma_x = \hat{\Sigma}_x, Z_n = \check{Z}_n)$ is calculated as the posterior distribution of a GP as Σ_x is given (Equation 4.14). This process is similar to the sequential sampling techniques used in Gibbs sapling (Murphy (2012))[25] to approximate multivariate distributions. The sampling method is verified by showing the samples match that of a t distribution when $\Sigma_z = 0$ (Figure 5.7a).

Given a sufficiently large number of samples, the first two moments of $p(X|\Sigma_z, Z_n = \check{Z}_n)$ may be approximated. A parametric distribution is then used to approximate $p(x_i|\Sigma_z, Z_n = \check{Z}_n)$ giving the marginal PDFs of the stochastic process. The chosen distribution is a mixture model of a t distribution and a normal distribution weighted by the kernel function and the noise at t_i.

$$p(x_i) \approx \frac{\sigma_{zi}^2}{\sigma_{zi}^2 + k(t_i, t_i)} \mathcal{N}(x_i; \bar{x}_i, \bar{\sigma}_i^2) + \frac{k(t_i, t_i)}{\sigma_{zi}^2 + k(t_i, t_i)} \mathcal{T}\Big(x; \bar{x}_i, \bar{\sigma}_i^2 \frac{v-2}{v}, v_0\Big) \tag{5.25}$$

Where $\bar{x}$ and $\bar{\sigma}^2$ are the arithmetic mean and variance of the sampled points respectively. To justify the choice of this mixture model consider the two extreme cases:

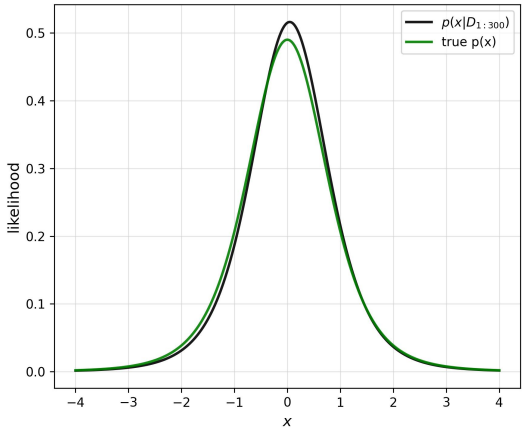
[25] Section 24.2.1.

First, suppose $\Sigma_z = 0$, then the stochastic process defines a TP with exact observations (Section 5.2.2). In this case the first term of the the mixture model (Equation 5.25) is zero leaving just the t distribution with v_0 degrees of freedom, matching the marginals of a TP with exact observations (Equation 5.8b).

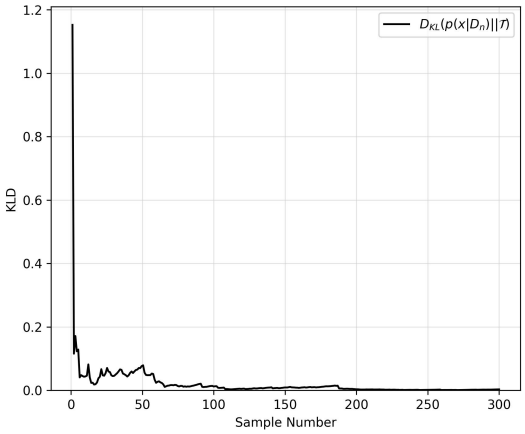
Second, suppose Σ_z is very large, and $K_{nn} = \delta I_n$, where I_n is the n-dimensional identity matrix and δ is a very small regularization value (to ensure K_{nn} is invertible). Under this prior, the Σ_x's sampled from the Inverse-Wishart prior will be near zero matrices, having a negligible impact on $\Sigma_x + \Sigma_z$. Thus, sampling x_is using Equation 5.24 is practically the same as sampling $x_i \sim \mathcal{N}(x_i; \bar{x}_i, \sigma_{zi}^2)$. This is reflected in the mixture model (Equation 5.25) as the term with the t-distribution being near zero leaving only the term with the normal distribution.

Between these cases, where neither Σ_x nor Σ_z are negligible, the weighted average of both cases are taken with respect to $\sigma_{z_i}^2$ and $k(t_i, t_i)$ (Equation 5.25). see figure Figure 5.7 showing the approximation of $p(x_i|\Sigma_z)$ in such a case.

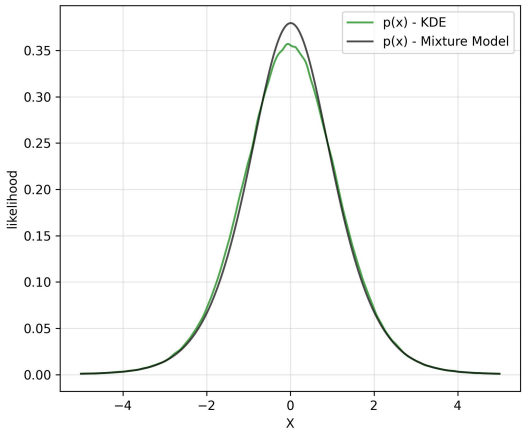
Further, as more observations about X are made, the degrees of freedom of the t distribution in the second term will increase. Thus the two terms will become more similar, elevating errors that come as a result of suboptimal weights used by the mixture model (Equation 5.25).



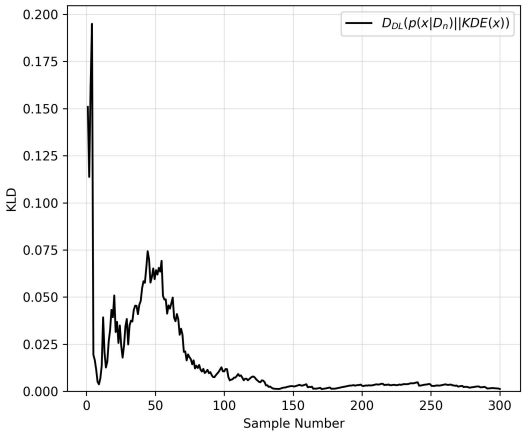
(a) T-Distribution approximated by mean and variance of 300 samples.



(b) Kullback–Leibler divergence (KLD) between true T-Distribution and approximation by number of samples.



(c) Approximation of $p(x_i|\Sigma_z)$ by kernel density estimation using 10^6 samples (acting as ground truth) versus approximation by Monte-Carlo Sampling and the two term mixture model using 300 samples.



(d) KLD between approximation of $p(x_i|\Sigma_z)$ by kernel density estimation using 10^6 samples (acting as ground truth) and approximation by Monte-Carlo Sampling and the two term mixture model using 300 samples.

Figure 5.7 – Demonstration that a mixture model of a Normal distribution and t-distribution (Equation 5.25) is an appropriate approximation for $p(x_i|\Sigma_z)$ with as little as 300 samples. for these charts a noisy case of $\Sigma_z = 0.5$ and $\Sigma_x \sim \mathcal{IW}(\Sigma_x; 1, 4)$ were used.

5.4.2 Implementation and Computational Constraints

Using Monte Carlo sampling methods to approximate a posterior distribution raises some computational concerns:

1. A sufficient number of covariance priors must be sampled from the inverse Wishart prior.
2. The Posterior distribution must be computed with time complexity $\mathcal{O}(n^3)$ for each sampled covariance.
3. Values must be sampled from the posterior distribution to approximate the first two moments of the marginal distributions $p(x_i|Z_n = \check{Z}_n)$

Firstly, The Inverse Wishart sampling function form scikit-learn (Pedregosa et al. (2011)) is capable of sampling a high number of large covariance matrices at a sufficient speed. Second, by leveraging the vectorization and einstein functions of numpy (Harris et al. (2020)), the matrix computations used to condition the posterior distributions run at a sufficient speed.

Lastly, sampling values directly from each posterior distribution took the most amount of time of all three steps. The best performing out of the box tool was found to be the multivariate normal sampling tool from numpy[26]. However, the sampling method still relies on singular value decomposition which in this case has has time complexity $\mathcal{O}(q^3)$ (Vasudevan and Ramakrishna (2017)). Further, this must be repeated for every sampled covariance matrix. This was found to take an unpractical amount of time considering most on demand applications.

To make this step more efficient, it must be first be recognized that each marginal distribution ($p(x_i|Z_n = \check{Z}_n)$) may be approximated independently from each other. Thus each sampled posterior covariance matrix is made diagonal (Algorithm 5.1 line 12).

[26] numpy/random/mtrand.pyx

This however, does not address the issue of sampling from multiple normal distributions. To address this, a q dimensional standard normal distribution is sampled from m times (Algorithm 5.1 line 5) only once (rather than for each sampled covariance). The standard deviation and mean of each posterior distribution calculated from a sampled covariance matrix is then used to scale the random samples from the standard normal distribution. This gives the same result as sampling from each sampled multivariate distribution with covariance made diagonal m times. This is not a heuristic, however, the marginal distributions are approximated independently of each other resulting in more jagged approximations along the T axis given small sample sizes.

The outline of the algorithm is as follows:

Algorithm 5.1: Monte Carlo Approximation of Posterior Marginal Distributions

1: Let l be the number of covariance matrices sampled.
2: Let m be the number of values to sample from each posterior distribution.
3: **Input:** $p(\Sigma_x)$, Σ_z, $\check{Z}_n$
4: **Output:** posterior samples
5: **Initialization:** standard_samples $\leftarrow$ sample($\mathcal{N}_q$, m)
6: **Step 1: Sample Prior Covariances**
7: prior_covs $\leftarrow$ sample($p(\Sigma_x)$, l)
8: **Step 2: Compute Posterior Means and Covariances**
9: posterior_means $\leftarrow$ $\{\mathrm{E}[p(X_q)|\Sigma_z, \Sigma_x = \Sigma_i, Z_n = \check{Z}_n] \mid \Sigma_i \in \text{prior\_covs}\}$
10: posterior_covs $\leftarrow$ $\{\mathrm{Cov}[p(X_q)|\Sigma_z, \Sigma_x = \Sigma_i, Z_n = \check{Z}_n] \mid \Sigma_i \in \text{prior\_covs}\}$
11: **Step 3: Extract Marginal Posterior Variances**
12: marginal_posterior_variances $\leftarrow$ $\{\mathrm{diag}(\mathrm{Cov}_i) \mid \mathrm{Cov}_i \in \text{posterior\_covs}\}$
13: **Step 4: Generate Posterior Samples**
14: posterior_samples $\leftarrow$ $\{\text{standard\_samples} \times s_i + \mu_i \mid (\mu_i, s_i^2) \in (\text{posterior\_means}, \text{marginal\_posterior\_variances})\}$
15: **return** posterior_samples

Where $\mathcal{N}_q$ is the q dimensional standard normal distribution.

The compute times on a MacBook-m1 are as follows:

Table 5.1 – Time it takes to approximate the posterior distribution using a varying number of samples. Each sample used $q = 200$ and $n = 50$.

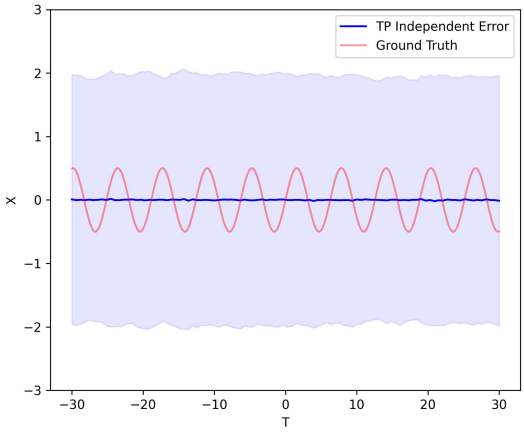
l \ m	**50**	**100**	**200**	**500**
50	1.39	1.63	2.09	3.42
100	1.61	1.89	2.39	3.99
200	1.72	2.16	2.91	5.75

When running the above, it appears the majority of the runtime is taken inverting the sampled covariance matrices of size K_{nn}, then sampling from the inverse Wishart prior. For the use cases of this thesis, It was found 200 covariance samples (l) and 50 standard normal samples (m) gave a sufficiently precise estimation.

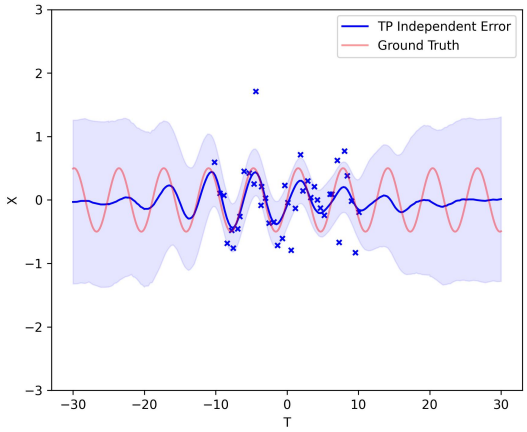
5.4.3 Performance of t Process Using Independent Noise

As discussed in Section 5.2.2 a TP with exact observations is capable of scaling the covariance of the posterior to match the data (Figure 5.2 and Figure 5.3). However, when using the canonical method (Section 5.3.2) to address noisy observations, the TP fails to scale the covariance of the posterior as the Mahalanobis distance is trained on noisy observations with respect to $\Sigma_x + \Sigma_z$.

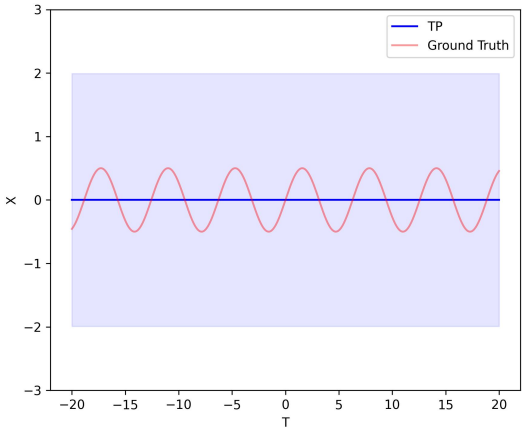
It was found the proposed model using independent noise could scale the covariance matrix while using noisy observations (Figure 5.8). Further, given the correct kernel function, the scaled posterior variance using independent noise appears to converge to the scaled posterior variance of the TP using exact observations (Figure 5.6). The expected values given by the model's posterior using independent noise was also slightly different than those of the canonical method. To the authors knowledge this is the first comparison of this kind.



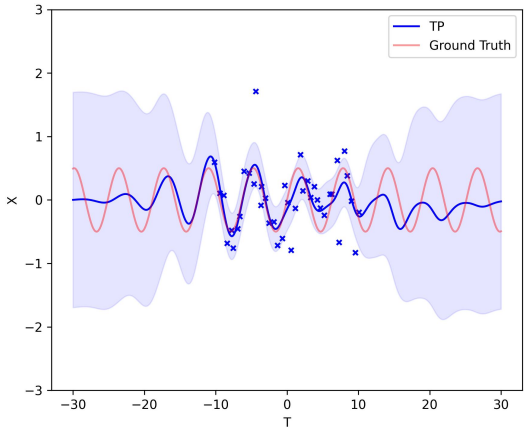
(a) Prior $p(x)$ showcasing Σ_x and the *jitter* in the confidence interval caused by the numeric approximation.



(b) Posterior $p(x|\check{Z}_{40}, \Sigma_z)$ showcasing the more accurate confidence over unobserved points.



(c) Prior $p(x)$ plotted as the closed-form TP



(d) Canonical method posterior $p(x|\check{Z}_{40}, \Sigma_z)$ showcasing the less accurate confidence over unobserved areas.

Figure 5.8 – The proposed TP subject to independent noisy observations (top). The canonical noisy TP (bottom). Though in both cases, the covariance scaled more conservatively then when exact observations were used, this is to be expected as noisy observations still regard information from the prior. Most notably the proposed model adjusts the covariance more than the canonical TP using noisy observations. The same parameters are used as in Figure 5.6.

Figure 5.9 compares: TPs with exact observations; GPs with exact observations; canonical TPs with noisy observations; and the proposed TP with independent noise. To analyse how well the models scale the covariance matrix, the model's predictive variance about x_i at $t_i = 1,000$ is tracked as more observations are made along the T axis. The choice of large t_i creates distance from the observed points making the effect of their local covariances negligible. Thus, capturing only the effect of the covariance scaling by the TPs. The same ground truth and parameters were used as in Figure 5.8. Note the kernel function is scaled to the variance of the ground truth in this case because the prior expected value is correct, and the kernel function is approximately correct besides for the scaling parameter.

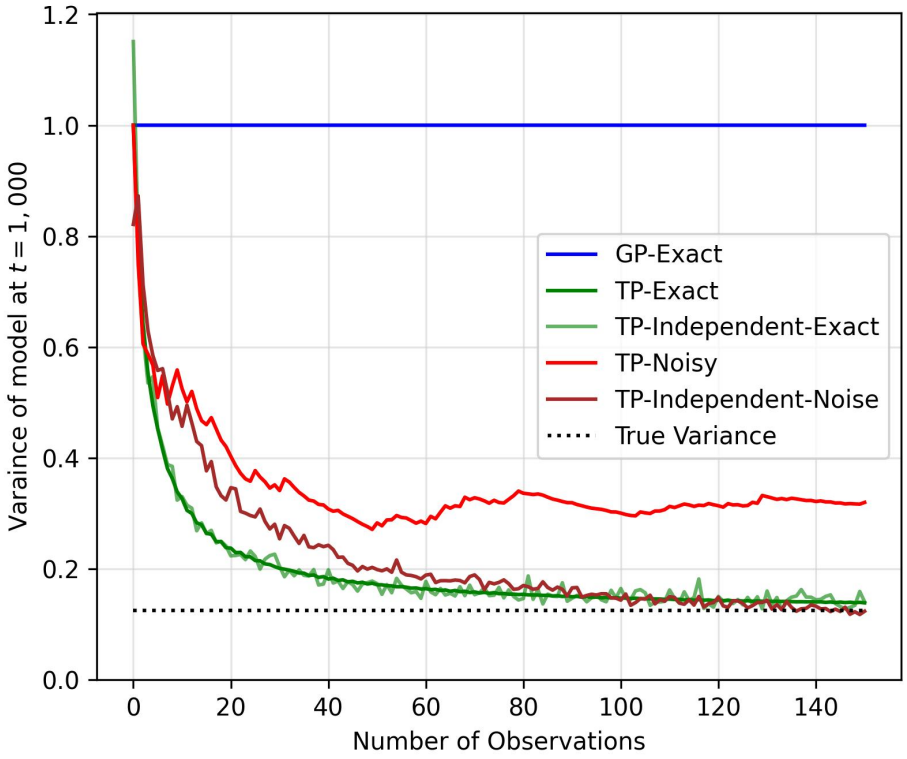


Figure 5.9 – Comparison of stochastic processes scaling their covariance to a sin function. All the models use the locally periodic kernel, prior expectations and ground truth function used in Figure 5.6. The numeric approximations were made with 1,000 samples.

Firstly, Figure 5.9 shows the covariance of the GP (GP-Exact) is fixed, predicting x_i within a constant variance. The covariance of the TP using exact observations (TP-Exact) converges to the variance of the ground truth function. The sampling method 5.24 using exact observations (TP-Independent-Exact) deviates very little from the true TP validating the sampling method. The canonical TP using noisy observations (TP-Noisy) is volatile, and does not converge to the true variance. The proposed TP using independent noise (TP-Independent-Noisy) is more stable than the canonical model and converges to the true variance while using noisy observations.

Application to Informative Path Planning

In most informative path planning applications observations will be subject to sensor noise. The proposed t process subject to independent noise preserves the scaling properties of TPs using exact observations while still accommodating noisy observations, making it a practical and informative option.

One disadvantage of applying the proposed method to IPPAs is that the mutual information of an observation cannot be easily approximated. When the method is applied in this thesis, the mutual information of a GP will be used as a proxy. The other disadvantage of this method is the additional compute time required.

5.5 Model Selection and Parameter Training

At a high level, the methods of model selection and kernel parameter training used by GPs (Section 4.3) may be applied to TPs. When training TPs however, the kernel function has moved from the covariance of the normal distribution, to the shape parameter of the Inverse Wishart prior.

As in the Gaussian case, kernel parameter values are optimized by maximizing the log likelihood of the observed values ($\check{Z}_n$) (Equation 4.16). One difference however, is that there is a strong argument to not train for the variance of the data as this will be scaled already by the Bayesian properties of the TP.

5.5.1 Parameter Training Over Canonical *t*-Processes

When using exact observations or the canonical noise model, the log likelihood of observed values is is calculated by their joint multivariate t distribution, taking the form (Tracey and Wolpert (2018))[27]:

$$\begin{aligned}\ln(p(\check{Z}_n)) &= \Gamma(\frac{v_0+n}{2}) - \Gamma(\frac{v_0}{2}) - \frac{n}{2}\ln(v_0\pi) - \frac{1}{2}\ln(|\Psi_{nn}|) - \frac{v_0+n}{2}\ln(1+\frac{S_n}{v_0}) \\ &= g(v_0,n) - \frac{1}{2}\ln(|\Psi_{nn}|) - \frac{v_0+n}{2}\ln(1+\frac{S_n}{v_0})\end{aligned} \tag{5.26}$$

Where $g(n, v_0)$ is a convenience function[28] encapsulating the terms that are independent of the kernel function; Ψ_{nn} may include canonical additive noise (Equation 5.20); and S_n is the Mahalanobis distance of the observed values with respect to Ψ_{nn} (Equation 5.10f).

Note, v_0 degrees of freedom are used rather than v_0+n as Equation 5.26 measures the log-likelihood over the prior. Though note, this is a design choice, optimizing using

[27] Equation 27

[28] $g(n, v_0) = \Gamma(\frac{v_0+n}{2}) - \Gamma(\frac{v_0}{2}) - \frac{n}{2}\ln(v_0\pi)$

the likelihood of a t-distribution with $v_0 + n$ degrees of freedom may also be justified (Yu et al. (2007)).

Building off of Algorithm 2.1 of Rasmussen et al. (2006), this thesis outlines a novel decomposition of the multivariate t distribution's log likelihood function and derivative to minimize the time complexity of the training algorithm. Analogous to Equations 4.17, first consider:

$$\text{Let } L := \text{Cholesky}(\Psi_{nn}) \tag{5.27a}$$

$$\text{Let } \alpha := L^T \backslash \Big(L \backslash \Psi_{nn} \Big) \tag{5.27b}$$

It follows (Equation A.3b):

$$\ln(p(\check{Z}_n)) = g(v_0, n) - \sum_{i=1}^{n} \Big(\ln(l_{ii}) \Big) - \frac{v_0 + n}{2} \ln(1 + \frac{(\phi(T_n) - \check{Z}_n)^T \alpha}{v_0}) \tag{5.27c}$$

The derivative with respect to kernel hyperparameter θ may also be written as (Equation A.3c):

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln\Big(p(\check{Z})\Big) &= \frac{\partial}{\partial \theta} \Big((-\frac{1}{2} \ln(|\Psi_{nn}|) \Big) \\ &\quad + \frac{\partial}{\partial \theta} \Big(-\frac{v_0 + n}{2} \ln\Big(1 + \frac{(\check{Z}_n - \phi(T_n)^T)^T \Psi_{nn}^{-1} (\check{Z}_n - \phi(T_n)^T)}{v_0}\Big)\Big) \\ &= \frac{1}{2} \text{tr}\Big(\Big(\frac{v_0 + n}{v_0 + (\check{Z}_n - \phi(T_n)^T)\alpha} \times \alpha\alpha^T - (LL^T)^{-1} \Big) \frac{\partial \Psi_{nn}}{\partial \theta} \Big) \end{aligned} \tag{5.27d}$$

As with the time complexity of training GPs, the complexity of training TPs is dominated by solving for the Cholesky decomposition and the subsequent linear system, $\mathcal{O}(\frac{n^3}{6})$ and $\mathcal{O}(\frac{n^2}{2})$ respectively. Again as in the GP case, further computing the derivative has complexity $\mathcal{O}(\frac{n^2}{2})$ and thus requires limited additional computational demand. Thus the time complexity of training TPs and GPs using gradient based optimizers are virtually the same.

To search for the maximum likelihood of the multivariate TP over the kernel param-

eter space (Equation 4.16), equations 5.27c and 5.27d are given to the same BFGS optimizer with 5 random start points as was used to train GPs (Virtanen et al. (2020)).

Figures 5.10b and 5.10d show the results of training the TP over 40 exact and canonical noisy observations. The optimal kernel parameters were found almost instantly when run on a M1 processor. It can be seen the TP is better fit to the signal following kernel parameter training in both cases.

5.5.2 Parameter Training Using Independent Noisy Observations

When training the kernel parameters of the proposed TP model, the likelihood of observations subject to independent noise must be approximated. This is done using the Monte-Carlo mixture model method described in Section 5.4.1. Unfortunately, this approximation must be repeated to evaluate each new set of kernel parameters throughout the optimization algorithm. Further, the derivative is not readily available thus rendering gradient based optimizers unavailable.

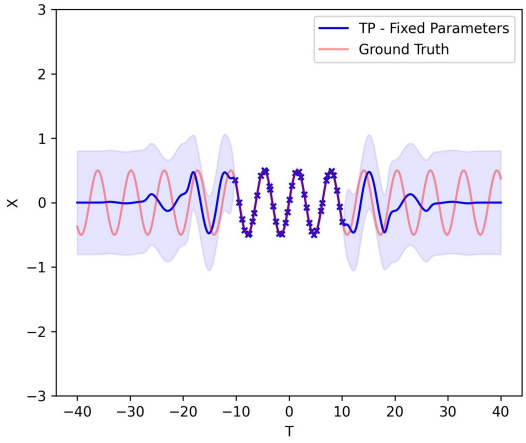
This results in more computationally demanding optimization algorithms. Training the proposed method using 40 observations subject to independent noise (Figure 5.10f) took 35 seconds on an M1 processor, whereas the canonical TP trained instantly. It was also found that the 'COBYLA' optimizer from Scipy (Virtanen et al. (2020)) was most roboust when optimizing over approximations of the log likelihood.

Although training the TP using independent noise takes more time, the results are superior to the trained TP using the canonical noise model. When comparing Figure 5.10d to Figure 5.10f, it can be seen the proposed method is better able to learn the periodicity of the signal.

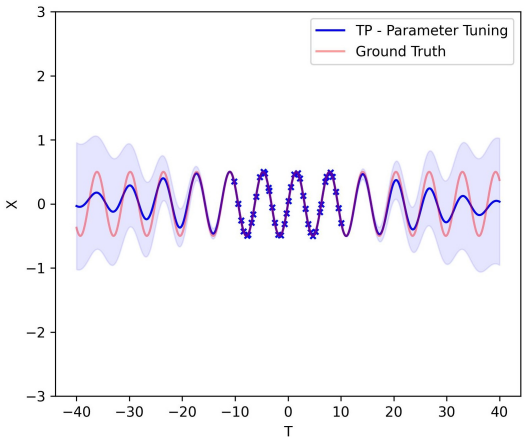
Further, comparing the plots subject to noisy observations prior to training (Figure 5.10c and Figure 5.10e), it can be seen the proposed model better fits the data

when using a suboptimal kernel function. Again, this because the covariance of the proposed model better fits to the observed values using Bayesian methods.

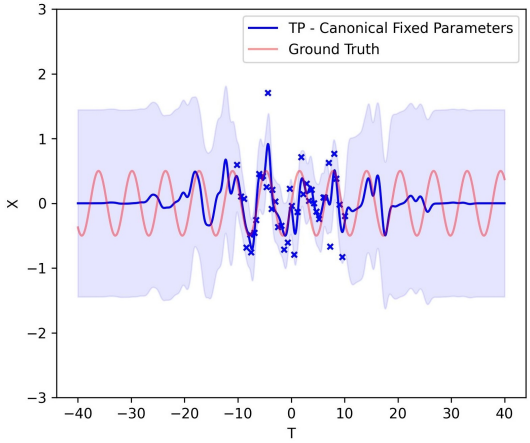
More robust and computationally efficient parameter training algorithms is a promising avenue of future work.



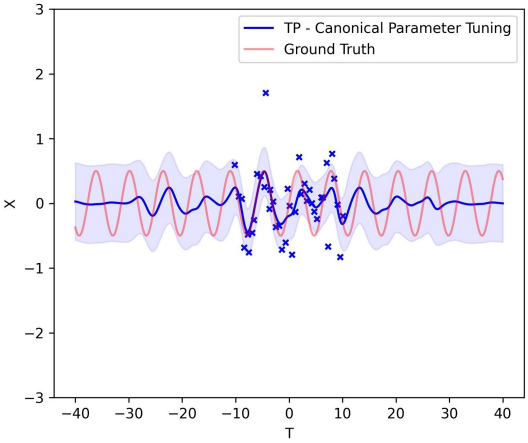
(a) TP of signal with exact observations using inaccurate kernel parameters.



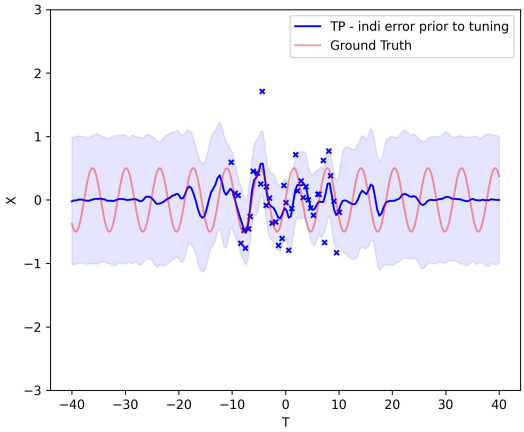
(b) TP of signal with exact observations following kernel parameter training.



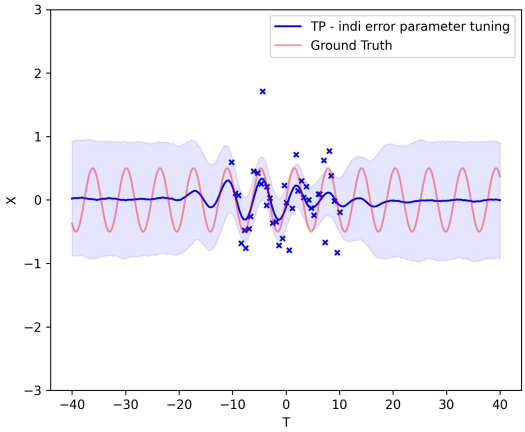
(c) TP of signal with canonical noisy observations using inaccurate kernel parameters.



(d) TP of signal with canonical noisy observations using kernel parameter training.



(e) Proposed TP using independent noise of signal using inaccurate kernel parameters



(f) Proposed TP using independent noise of signal using kernel parameter training.

Figure 5.10 – Comparing TP training on signal with and without parameter training. The same parameters were used as in the GP case (Figure 4.5) across all models.

Chapter 6

Application of Stochastic Processes to Informative Path Planning

This section will combine ideas from Chapters 2, 4 and 5 to apply the stochastic processes discussed in this thesis to the informative path planning constraint problem.

Firstly, simple but representative IPPAs are defined for each model. They are then evaluated on a potassium density data set, using a variety of metrics to showcase their strengths and weaknesses.

It was found when TPs were applied to this informative path planning problem, they were able to successfully scale the prior, generally resulting in an better fit. The improvement in fit was most noticeable given an under-confident prior, as the Bayesian update step was able to make more pronounced changes. Given an over-confident prior, the quality fit of the data was largely the same. The prediction error was slightly greater, but made up for by the scaled variances.

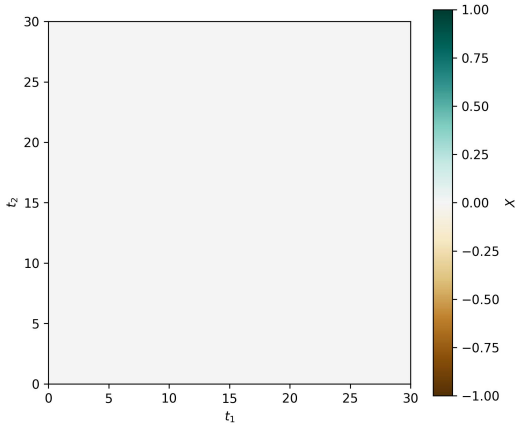
A high level review of the application of GPs to informative path planning, a discussion of related work and the proposed benefits of TPs in this context is covered in Section 2.4.1.

6.1 Stochastic Processes in Two Dimensional Space

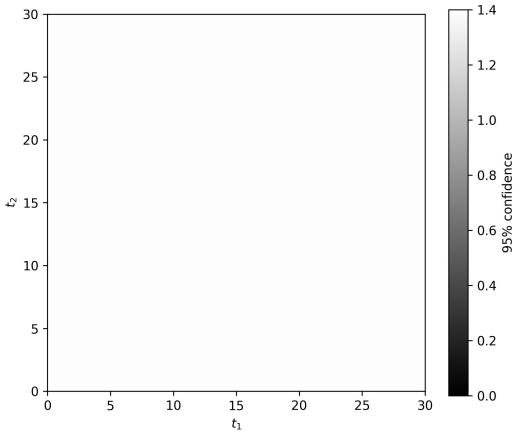
Thus far GPs and TPs have been defined and exemplified in one dimensional space ($T \subseteq \mathbb{R}$). However, most physical environments occupy two or three dimensional space. The examples in this chapter will assume the environment can be represented in two dimensional space ($T \subseteq \mathbb{R}^2$). In this context, the following variable domains and function signatures change:

- $t_i \in \mathbb{R}^2$
- $k : (\mathbb{R}^2, \mathbb{R}^2) \to \mathbb{R}$
- $\phi : \mathbb{R}^2 \to \mathbb{R}$
- $\psi : \mathbb{R}^2 \to \mathbb{R}^+$

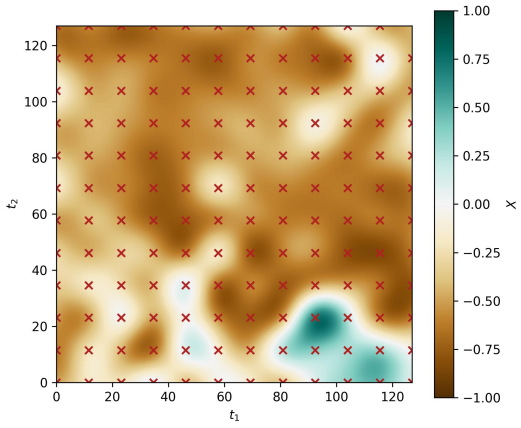
Figure 6.1 demonstrates a GP modelling two dimensional geographic data. The values of the raster data correspond to the potassium density over a region in New South Wales, Australia. The data set will be discussed in more detail in Section 6.3.



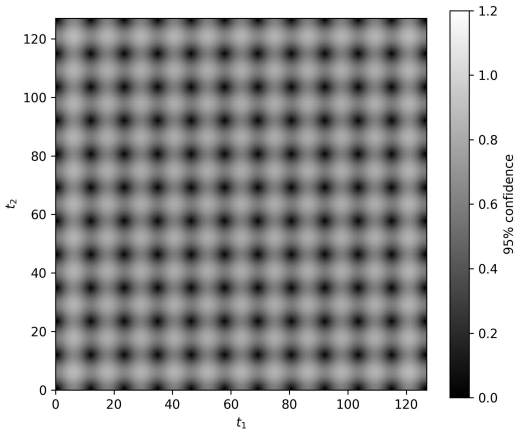
(a) Prior expected values of X.



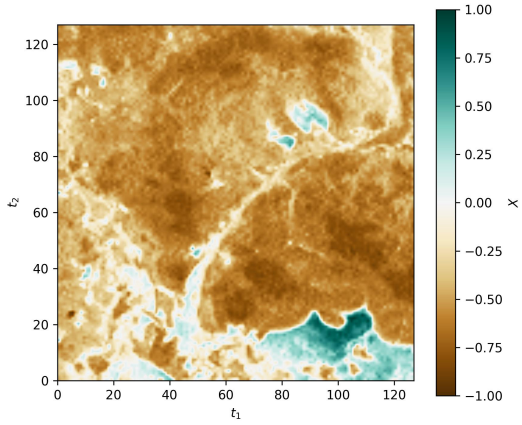
(b) 95% confidence interval of prior.



(c) Posterior expected values of X following 12^2 observations.



(d) 95% confidence interval of posterior following 12^2 observations.



(e) Ground Truth of X

Figure 6.1 – GP modelling two dimensional data. These figures demonstrate how expected values and confidence disperse across the environment. The GP used RBF kernel with $r = 0.5$ and $l = 7$.

It can be seen the GP can only roughly approximate the ground truth even when using 12^2 exact observations. This is largely due to the unpredictability and coarseness of the data set. None the less, the GP captures the general trend of the data and captures the uncertainty between observed locations.

6.2 Application of Stochastic Processes to Informative Path Planning

6.2.1 Problem Definition

This section outlines several IPPAs, each driven by the different stochastic processes discussed in this thesis. In each case the environment will be a bounded, convex, two dimensional space, and the agent will be subject to a travel distance constraint.

- Let $T \subset \mathbb{R}^2 \;\; st. \;\; T$ be a convex set[1].
- Let $t_i = (t_{i_0}, t_{i_1}) \in T$ represent the current location of the agent.
- Let $c(a_i) = \|t_j - t_i\| = \sqrt{(t_{j_0} - t_{i_0})^2 + (t_{j_1} - t_{i_1})^2}$ be the cost of an observation action. This is the distance the agent must travel from t_j to t_i in order to make an observation about x_i.
- Let $c(A_{1:n}) = \sum_{i=1}^{n} \|t_{i-1} - t_i\|$ be the cost of an action sequence. This is the geometric distance of the path across all points.
- Let $B_0 \in \mathbb{R}^+$ be the starting budget of the research mission.

The definitions of X, Z, A, t_i, x_i, z_i and a_i follow directly from Chapter 2.

Following from Section 2.3, IPPAs will be defined as solutions to the following optimization constraint problem:

$$\max_{C(A_n)<B} I(X; Z_n) \tag{6.1}$$

[1]In a convex set, any two points in the set can be connected by a straight line contained in the set. Defining T to be convex simplifies the cost function while maintaining generality.

6.2.2 Stochastic Processes Using Exact Observations

When using GPs or TPs with exact observations to model the QOI, the mutual information between observations and the QOI ($I(X; Z_n)$) has a closed form solution. This allows for more straight forward IPPAs.

Using Gaussian Processes With Exact Observations

When using GPs with exact observations the IPPA is derived from:

- $p(X_q) = \mathcal{N}(x_i; \phi(t_i), K_{qq})$ derived in Equation 4.5.
- $p(X_q|Z_n) = \mathcal{N}(X; \tilde{\mu}, \tilde{\Sigma})$ derived in Equation 4.6.
- $I(X; Z_n) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{nn}|)$ derived in Equation 4.9.

Giving:

$$\begin{aligned} &\max_{C(A_{1:n})<B} I(X; Z_n) \\ &= \max_{\sum_{i=1}^{n} \|t_{i-1}-t_i\|<B} \Big(\frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln(|K_{nn}|)\Big) \end{aligned} \tag{6.2}$$

Using t Processes With Exact Observations

When using TPs with exact observations the IPPA is derived from:

- $p(\Sigma_x) \sim \mathcal{IW}(\Sigma_x; (v_0 - 2)K_{qq}, q - 1 + v_0)$ derived in Equation 5.6.
- $p(X_q) = \mathcal{T}(X; \phi(T), \frac{v_0-2}{v_0}K_{qq}, v_0)$ derived in Equation 5.7
- $p(X_q|Z_n) = \mathcal{T}(X_*; \tilde{\mu}, \tilde{\Psi}, \tilde{v})$ derived in Equation 5.10.
- $I(X_q; Z_n) = \frac{1}{2}\ln(|\Psi_{nn}|) + H(\mathcal{T}_{n,v_0})$ derived in Equation 5.14.

Giving:

$$\begin{aligned} &\max_{C(A_{1:n})<B} I(X; Z_n) \\ &= \max_{\sum_{i=1}^{n} \|t_{i-1}-t_i\|<B} \Big(\frac{1}{2}\ln(|\Psi_{nn}|) + H(\mathcal{T}_{n,v_0})\Big) \end{aligned} \tag{6.3}$$

6.2.3 Informative Path Planning Tree Search Given Exact Observations

When using exact observations, Equation 6.2 and Equation 6.3 define the optimization constraint problem driving the IPPA when GPs and TPs are used to model the QOI respectively. The solution to these constraint problem must now be found or approximated by searching the action space (A). This will give the next observation action to exercise. Once the action is exercised and the belief map is updated, the next observation action is searched for again.

When using GPs or TPs, finding the optimal action sequence is an NP-hard problem, especially if T is treated as a continuous space (Hitz et al. (2017)). To make the problem more approachable, the continuous space T is discretized into a grid. Under this discretization, the space of possible action sequences is finite. This is a common simplification among IPPAs (Jakkala and Akella (2024)). Although finite, in practice the action space is still far too large to search through completely, this motivates analytic solutions, approximations or heuristic tree search techniques.

In many applications, the kernel function and occupancy map is given a priori. In this context, the optimal action sequence may be derived prior to the mission, relieving the need for online IPPAs. However, when the uncertainty of a belief model changes in a spatially inconsistent way following observations, online IPPAs are optimal.

When using GPs or TPs, the kernel function determines the most informative action sequence. In cases where the kernel function is trained in an anisotropic or spatially inconsistent way, the most informative action sequence will change depending on observed values. Thus, justifying an online algorithm.

It was found in other works with similar applications, that Monte-Carlo Tree Search (MCTS) is a suitable technique to approximate the most informative action sequence of an online IPPA (Arora (2018); Garza (2021)). The search algorithm is anytime and well suited for approximating large search spaces. A thorough explanation of its implementation into IPPAs can be found in Arora (2018)[2].

[2]Section 3.2.3.

To limit the search space, the available actions at each iteration in the tree search are limited to a surrounding grid of available points (Figure 6.2). This also guides the search in a reasonable direction.

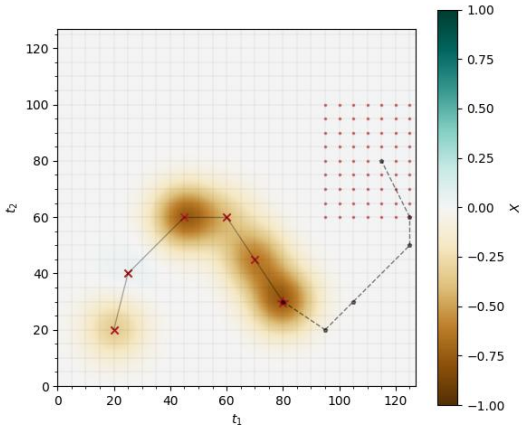


Figure 6.2 – Example of 6 observations, followed by an action sequence in the search space going 5 observations deep. The star shows the agent's current location; the dotted line connects potential observation actions and the red grid shows the search space of a 6^{th} observation action.

6.3 Stochastic Processes Using Noisy Observations

When modelling the QOI using GPs or TPs subject to noisy observations, components of the IPPA must be approximated. Using these stochastic processes requires a more involved IPPA.

Gaussian Processes Using Noisy Observations

When using GPs with noisy observations, the mutual information of observations must be approximated. The IPPA is hence approximated using:

- $p(X_q) = \mathcal{N}(x_i; \phi(t_i), K_{qq})$ derived in Equation 4.5.
- $p(X_q|Z_n) = \mathcal{N}(X; \tilde{\mu}, \tilde{\Sigma})$ derived in Equation 4.14.

- $I(X;Z_n) \approx I(X_q;Z_n) = \frac{1}{2}\ln(|K_{qq}|) - \frac{1}{2}\ln(|\tilde{\Sigma}_{qq}|)$ derived in Equation 4.15d.

Giving:

$$\begin{aligned} &\max_{C(A_{1:n})<B} I(X;Z_n) \\ \approx &\max_{\sum_{i=1}^{n}\|t_{i-1}-t_i\|<B} \Big(\frac{1}{2}\ln(|K_{qq}|) - \frac{1}{2}\ln(|\tilde{\Sigma}_{qq}|)\Big) \end{aligned} \tag{6.4}$$

Canonical *t* Processes Using Noisy Observations

As with the noisy GP case (Equation 6.4), when using canonical TPs with noisy observations, the mutual information of observations must be approximated. The IPPA is hence approximated using :

- $p(\Sigma_x) \sim \mathcal{IW}(\Sigma_x;(v_0-2)K_{qq}, q-1+v_0)$ derived from joint prior Equation 5.19.
- $p(X_q) = \mathcal{T}(X;\phi(T), \frac{v_0-2}{v_0}K_{qq}, v_0)$ derived from joint prior Equation 5.20
- $p(X_q|Z_n) = \mathcal{T}(X_*;\tilde{\mu},\tilde{\Psi},\tilde{v})$ derived from joint prior using Equation 5.10.
- $I(X;Z_n) \approx I(X_q;Z_n) = \frac{1}{2}\ln\Big(\frac{|\Psi_{qq}|}{|\Psi_{J/(\Psi_z+\Psi_{nn})}|}\Big)+f(v,n,q)$ derived in Equation 5.22.

Giving:

$$\begin{aligned} &\max_{C(A_{1:n})<B} I(X;Z_n) \\ \approx &\max_{\sum_{i=1}^{n}\|t_{i-1}-t_i\|<B} \left(\frac{1}{2}\ln\Big(\frac{|\Psi_{qq}|}{|\Psi_{J/(\Psi_z+\Psi_{nn})}|}\Big)+f(v,n,q)\right) \end{aligned} \tag{6.5}$$

t Processes Using Independent Noisy Observations

t Processes using independent noisy observations as described in Section 5.4 do not have closed form solutions for $p(X|Z_n)$, let alone $I(X;Z_n)$ or $I(X_q;Z_n)$. Thus for simplicity, action sequences in this case will be chosen using the same methods as when using a GP.

6.3.1 Informative Path Planning Tree Search Using Noisy Observations

As described in Equation 6.4 and Equation 6.5, when using GPs or TPs subject to noisy observations to model the QOI, the mutual information of observations must be approximated. Thus, the objective function of the IPPA's tree search algorithm must be approximated. This is done by discretizing the environment into a finite number of points $\Big(I(X; Z_n) \approx I(X_q; Z_n)\Big)$.

The larger q is, the more representative the approximation is of the continuous environment. However, the computational demands of a large q must also be considered, especially when integrated into an anytime heuristic tree search algorithm.

To choose what points to include in this approximation, the shape of the kernel function driving the mutual information is considered. When there is a length component to the kernel function, the mutual information between points decreases with distance:

$$\|t_i - t_j\| \uparrow \implies I(x_i; z_j) \downarrow \tag{6.6}$$

Thus, points closer to the potential observation points (Z_n) should be prioritized when approximating $I(X; Z_n)$.

When evaluating the mutual information of an action sequence, a grid of points is cast around each considered observation location Figure 6.3. This results in a collection of points that are representative of how the entropy of X will change following observations Z_n.

To avoid large numeric issues when approximating mutual information, the budget of the agent is temporarily reduced before running the MCTS algorithm. This limits the depth of the search and thus the number of points on the comparison grid.

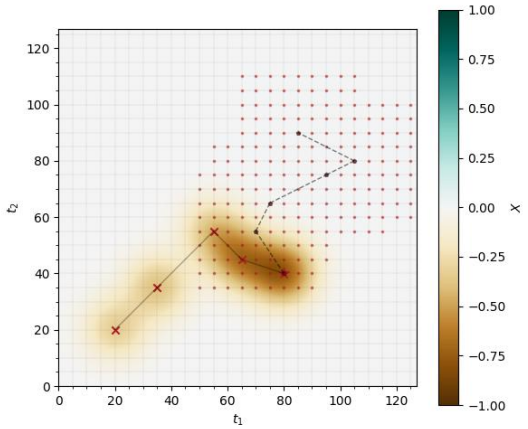


Figure 6.3 – Example of 5 observations, followed by an action sequence in the search space going 5 observations deep. The red dots show what points are evaluated when approximating the mutual information of the observation sequence.

6.4 New South Whales Potassium Mapping

This section will apply the abstract IPPAs of each stochastic process to real geologic data. The strengths and weaknesses of each IPPA will then be demonstrated using a variety of metrics.

6.4.1 New South Wales Potassium Data Set

To evaluate the proposed IPPAs, a series of research missions will be run on a dataset of potassium density over NSW Australia (Figure 6.4). The potassium density measured as a percentage, is determined by the mineral composition of the subject rocks or soil. The dataset was generated by the Mining Exploration and Geoscience Division of Regional NSW (Matthews (2023)). It was compiled by merging a series of airborne radiometric surveys.

The resolution of the map is approximately 400m/pixel, varying slightly by region. Potassium % values in the original data set range from -1.36% to 6.85%.

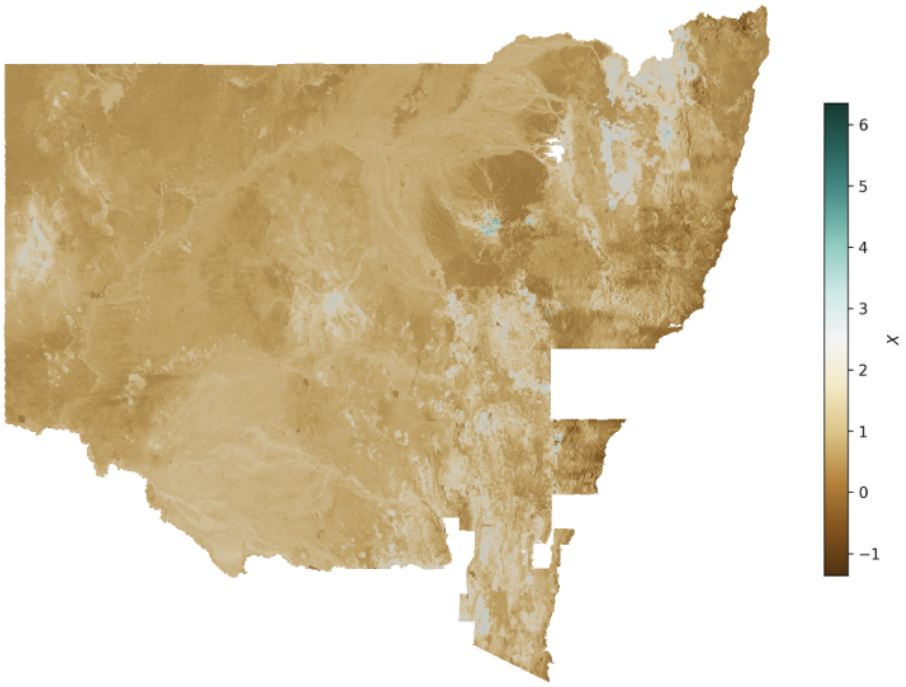
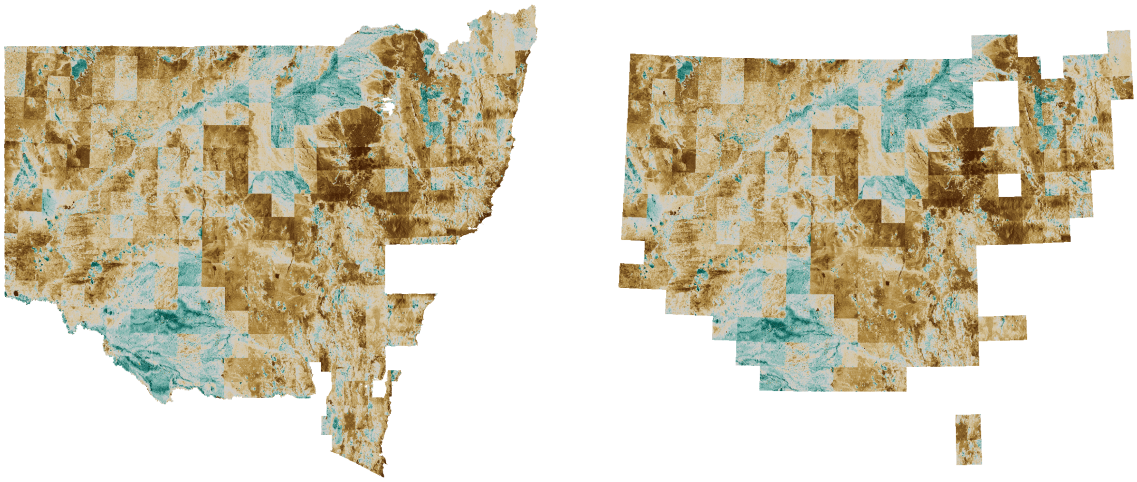


Figure 6.4 – Potassium density map over NSW Australia.

Data Preprocessing

To generate a collection of comparable data sets to evaluate the proposed IPPAs on, the NSW Potassium dataset is subdivided into 500 square regions. The potassium density values in each region are then normalized to be between -1 and 1. Regions with null values[3] are also ignored to control for size.



(a) NSW potassium data subdivided into regions and normalized.

(b) NSW normalized potassium data, excluding regions with null values.

Figure 6.5 – NSW data preprocessing steps.

This leaves 236 regions with the following properties:

- size: 50km × 50km
- resolution: 125 × 125
- value rage: $[-1, 1]$

These preprocessing steps provide a collection of coarse scalar fields to test the proposed IPPAs on. Their consistent size and value range make for relatively comparable error metrics reducing outliers in performance.

[3] Typically regions including bodies of water, cities or borders.

6.4.2 New South Wales Potassium Informative Path Planning Algorithms

This section will make the abstract IPPAs outlined in Section 6.2 specific to the potassium density data. In this simulation, the QOI is the normalized value of potassium density. Thus, the objective of the IPPAs is to reduce the uncertainty about this value.

This simulates the informative path planning constraint problem of an agent travelling across the region making in-situ potassium density observations, giving physical meaning to the variables:

- $q = (125 \times 125)$ measured in 400m$\times$ 400m pixels.
- $\phi(t) = 0$ is the prior's expected value.
- $X \subset \mathbb{R}^q$ is the potassium density (normalized to be between -1, 1)
- $B_0 = 500$ (400m pixels) enough budget to circle the region's perimeter.

When observations are specified as noise, they will be subject to Gaussian white noise with standard deviation 0.1, otherwise observations will be exact:

- When noisy: $p(z_i|x_i) = \mathcal{N}(z_i; x_i, 0.1^2)$
- When exact: $z_i = x_i$

The informative path planning optimization constraint problem specific to the potassium density simulation may be formulated as:

$$\max_{C(A_n)<500} I(X_{125^2}; Z_n) \tag{6.7}$$

Potassium Density Kernel Function

Each stochastic process used to model the QOI will be instantiated with a diagonally anisotropic Matérn kernel function (Equation 6.8a) (Rasmussen et al. (2006))[4]. This

[4]Equation 4.1.4 and Section 4.2.

is a stationary kernel function that spatially correlates the presence of potassium in the region and allows the the length scale parameters to be trained in both axes.

$$k(t_i, t_j) = r^2 \times \frac{1}{\Gamma(u)2^{u-1}}\Big(\sqrt{2u}(t_i - t_j)^{\mathrm{T}} M^{-1}(t_i - t_j)\Big)^u \mathcal{K}_u\Big((t_i - t_j)^T M^{-1}(t_i - t_j)\Big) \tag{6.8a}$$

where $\mathcal{K}_v$ is the modified Bessel function and:

$$M = \begin{bmatrix} l_1 & 0 \\ 0 & l_2 \end{bmatrix} \tag{6.8b}$$

The Matérn kernel function is a generalization of the RBF and exponential kernel function (Genton (2002)). The u parameter will be fixed to a low value of 1.5, giving relatively high likelihoods to less smooth regressions. This helps the model adapt to the coarseness and apparent randomness of the data set's ground truth. As the u parameter approaches infinity, the Matérn kernel approaches the RBF kernel.

When the kernel parameters are not being trained for, a fixed length scale of 12 will be used equating to 4.8km. Otherwise they will be trained subject to bounds $l \in (3, 20)$. The bounds prevent obscure behaviour in the event that the model cannot fit to the data or in the event that the best fit goes against fixed prior beliefs about the distribution of potassium.

In each case the radius parameter r will be fixed. Section 6.5.1 and Section 6.5.2 will fix r to be too high and low respectively, to evaluate the performance of the IPPAs under these challenging conditions.

Computational Constraints

The MCTS algorithm used to approximate the subsequent best observation action, is given two seconds of compute time. Further, to encourage breadth first search, the MCTS algorithm will not consider states that exceed a travel distance of 200 pixels. To approximate the TP subject to independent noise, 200 covariance matrices are sampled.

6.4.3 Performance Metrics

Several metrics will be compared to evaluate the performance of each IPPA. To improve the runtime of evaluating models, rather than evaluate the models using the full resolution available (125^2), the models will be evaluated using a resolution of 32^2. This had negligible impact on the result trends.

The metrics at each point will be computed individually, rather than considering the multidimensional distribution over X_q. This is because the proposed TP subject to independent noise does not approximate a multidimensional distribution, and further, marginal distributions over x_i are more important in many applications.

To define the performance metrics, the following notation is used:

- Let τ_i be the true value of x at t_i.

Mean Absolute Error

The mean absolute error quantifies the error of expected values given by the stochastic process:

$$\text{Mean Absolute Error} = \frac{1}{n}\sum_{i=1}^{q}\Big(\mathrm{E}[x_i|Z_n = \check{Z}_n] - \tau_i\Big) \tag{6.9}$$

Mean Squared Error

The mean squared error (MSE) quantifies the error of expected values given by the stochastic process, penalizing errors quadratically:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{q}\Big(\mathrm{E}[x_i|Z_n = \check{Z}_n] - \tau_i\Big)^2 \tag{6.10}$$

Average Variance

The average variance is not strictly a performance metric, however is useful when contextualizing the other results.

$$\text{Average Variance} = \frac{1}{n}\sum_{i=1}^{q}\text{Var}[x_i|Z_n = \check{Z}_n] \tag{6.11}$$

In this section, average variance also serves as a proxy for entropy as the entropy of mixture models is not easily approximated.

Mean Squared Prediction Error

The mean squared prediction error (MSPE) quantifies the expected MSE, given a distribution over the QOI at a point t_i:

$$\begin{aligned}\text{MPSE} &= \frac{1}{n}\sum_{i=1}^{q}\text{E}[(x_i - \tau_i)^2|Z_n = \check{Z}_n] \\ &= \frac{1}{n}\sum_{i=1}^{q}\Big(\text{Var}[x_i] + (\text{E}[x_i|Z_n = \check{Z}_n] - \tau_i)^2\Big) \\ &= \text{MSE} + \frac{1}{n}\sum_{i=1}^{q}\text{Var}[x_i]\end{aligned} \tag{6.12}$$

As can be seen, this is effectively MSE (Equation 6.10) that additionally penalizes variance. Hindley (2017)[5] refers to the terms in the second line of Equation 6.12 as the process variance (variance of the stochastic process's predictor) and estimation variance respectively. See Equation A.4 for further decomposition of Equation 6.12.

Average Likelihood

The average likelihood of the ground truth τ_i given by the probability distribution at that point $p(x_i|Z_n = \check{Z}_n)$ quantifies how well the distribution models the ground

[5]Equation A.7.

truth. This simultaneously evaluates the error and confidence of a stochastic process.

$$\text{Average Likelihood} = \frac{1}{n} \sum_{i=1}^{q} p(x_i = \tau_i | Z_n = \check{Z}_n) \tag{6.13}$$

Average Log Likelihood

The average log likelihood of the ground truth given the probability distribution at a point, quantifies how well a distribution models the ground truth, while penalising low likelihood outliers exponentially.

$$\text{Average Log Likelihood} = \frac{1}{n} \sum_{i=1}^{q} \ln(p(x_i = \tau_i | Z_n = \check{Z}_n)) \tag{6.14}$$

6.5 Results

To consolidate the topics discussed in this chapter, first, abstract IPPAs were defined for each stochastic process discussed in this thesis (Section 6.2). Next, the NSW potassium density data set was introduced (Section 6.4.1), and problem specific kernel functions, operational constraints and computational constraints were set. Lastly, a series of metrics were defined to quantify the performance of each IPPA given the ground truth.

These models will be tested given an over-confident prior and an under-confident prior, meaning the variance of the prior is less then and larger than that of the data respectively [6]. The presented results train the anisotropic length scale kernel parameters after each observation, justifying the online MCTS policy. largely equivalent results without kernel parameter training can be found in Section A.5.

Each IPPA is tested across 100 randomly selected regions from the NSW potassium data set. To compare the results of each IPPA, each metric is averaged across the tested regions and plotted with respect to the agents budget with a 1 standard deviation buffer.

[6] One dimensional analogies of this this can be seen in Figure 5.2 and Figure 5.3.

6.5.1 Results Given Under Confident Prior

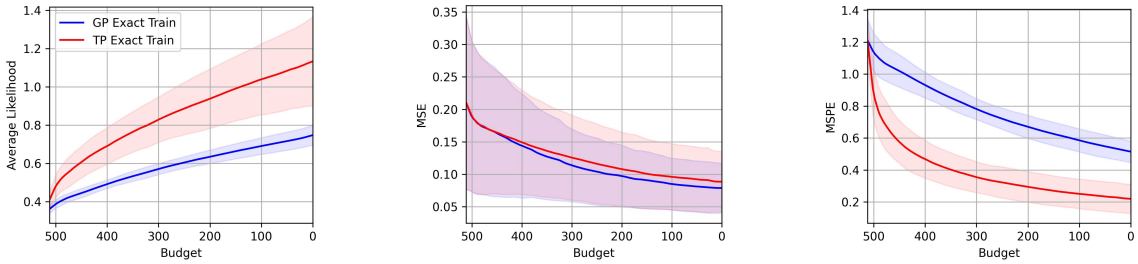
This subsection will fix the radius parameter r of the kernel function (Equation 6.8a) to be 1. This gives an under-confident prior over the distribution of potassium with respect to the ground truth. Example runs of each IPPA given this prior can be found in Figure A.1.

Figure 6.6 compares the performance of the IPPAs given the under-confident prior and exact observations. Firstly, it is observed the MSE and Mean absolute error of each IPPA decrease at a similar rate as the distance budget is exhausted (Figure 6.6b and Figure 6.6e). This is to be expected as the stochastic processes are conditioned on observations and the expected value of TPs and GPs are algebraically the same[7]. On average, the IPPAs driven by TPs have slightly more error when compared to the IPPAs driven by GPs. This is likely because the TPs prioritizes different points to observe than the GP, especially under the time pressure of the MCTS algorithm.

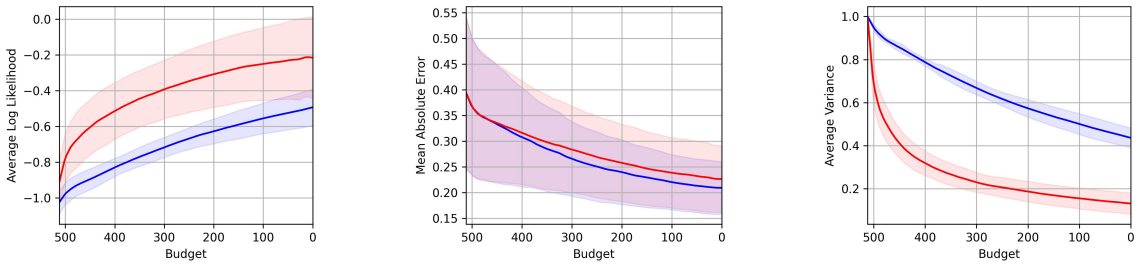
Where the TP stands out, is their ability to recognize the prior is under-confident and consequently scale down the covariance of the model (Figure 6.6f). Thus, even though the TPs error is slightly greater, its MSPE is significantly smaller (Figure 6.6c). Further, the lower variance is justified, as the likelihood of the ground truth is greater when modelled using TPs (Figure 6.6a Figure 6.6d).

Figure 6.6 compares the results of the IPPAs subject to Gaussian white noise given the under-confident prior. The IPPAs considered model potassium using noisy GPs, canonical noisy TPs and TPs subject to independent noise. Overall, these results are similar to the results using exact observations (Figure 6.6). However, it was shown TPs using independent noise were capable of scaling the variance of the posterior distribution just as the canonical TP did. Further, in part because the IPPA uses the objective function given by a GP, errors were also smaller. Overall using the proposed TP subject to independent noise resulted in a better fit with greater likelihood.

[7]Compare Equation 5.10c and Equation 4.6c

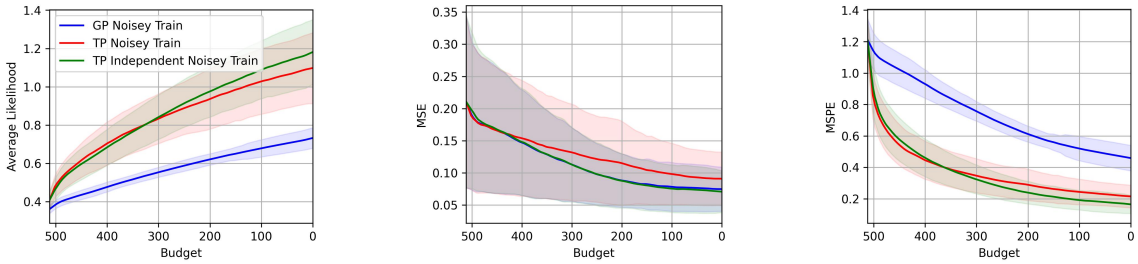


(a) Likelihood **(b)** MSE **(c)** MSPE

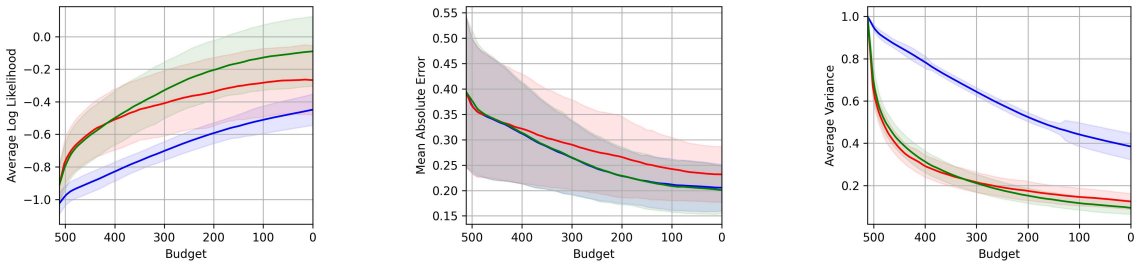


(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure 6.6 – IPPA performance given **exact** observations and under-confident prior.



(a) Likelihood **(b)** MSE **(c)** MSPE



(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure 6.7 – IPPA performance given **noisy** observations and under-confident prior.

6.5.2 Results Given Over-Confident Prior

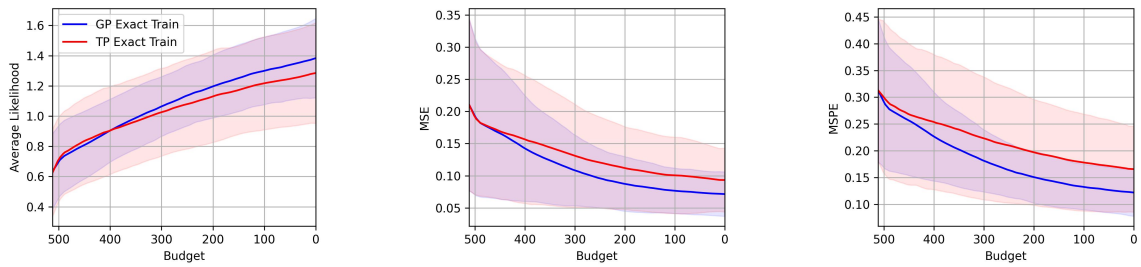
This section will fix the radius parameter r of the kernel function (Equation 6.8a) to be 0.32. This gives an over-confident prior with respect to the ground truth. Example runs of each IPPA given this prior can be found in Figure A.2.

Figure 6.8 compares the results of the IPPAs driven by GPs and TPs given the over-confident prior and exact observations. As with the under-confident case, the error of the IPPAs driven by TPs is slightly greater than those using GPs.

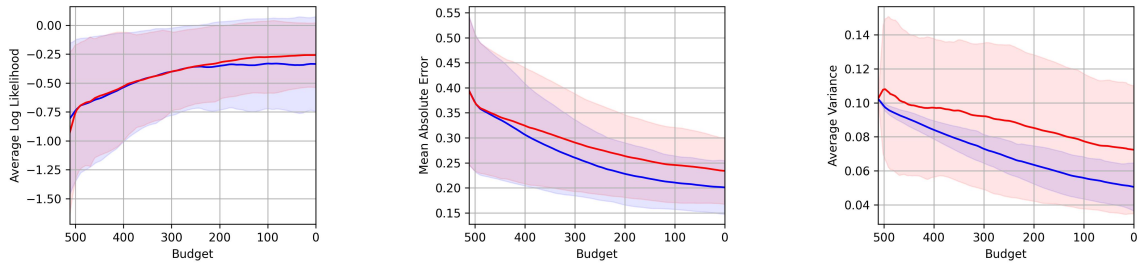
Again, the notable difference in the mode performance lies in how variance is scaled. In this case, the variance given by the prior is too small. Thus, at first, the TPs increase the variance as observations are collected. After several observations, the scale of the covariance begins to stabilize and the variance begins to decrease as observations are used to better model the QOI.

This increase in variance has a negative impact on the likelihood of well represented data as the peaks of the marginal distributions ($p(\tau_i|Z_n = \check{Z}_n)$) are lower. However, poorly modelled data is given a greater likelihood as the tails of the marginal distributions are wider. When using the log likelihood, this effect is made more intense as outliers are penalized more severely. The likelihood of the ground truth when modelled using TPs was slightly lower, likely due to the slightly greater absolute error. However, the log likelihood was slightly greater as the TPs better captured outliers due to their scaled variance and heavy tailed shape.

Figure 6.6 compares the results of the IPPAs subject to Gaussian white noise given the over-confident prior. In this case, each IPPA fit the data with similar likelihood. The GP and canonical TP scaled the variance in a similar way as when using exact observations. However, the variance of the TP using independent noisy observations scaled the covariance differently. The variance increased at first recognizing the under-confident prior, but then gradually decreased as observations were made. Interestingly, this yielded the same likelihood as the other models, likely due to the model's greater absolute error as a consequence of conditioning the posterior closer to the over-confident prior's expectation of $\phi(t) = 0$.

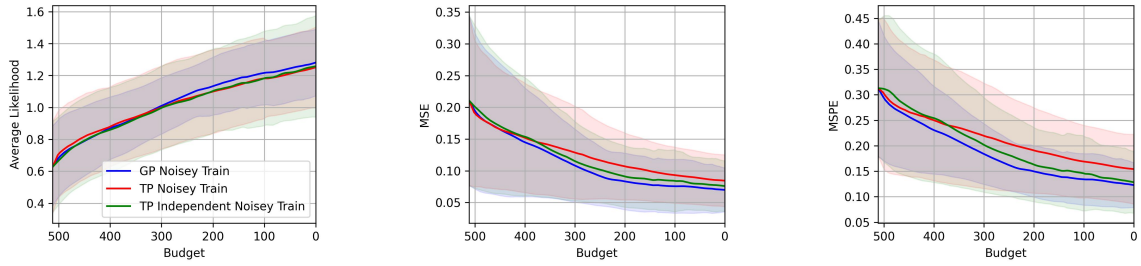


(a) Likelihood **(b)** MSE **(c)** MSPE

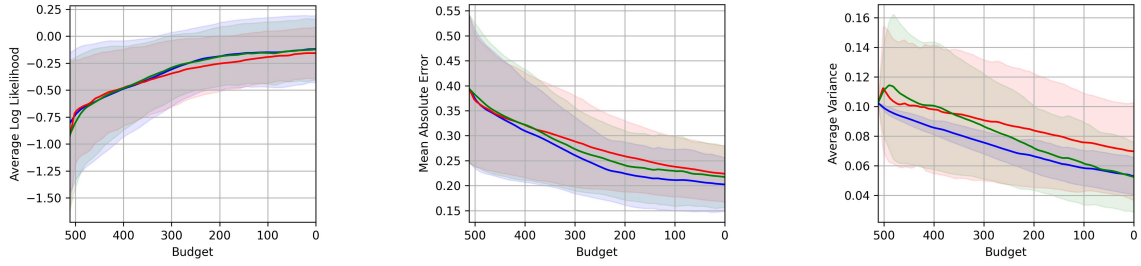


(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure 6.8 – IPPA performance given **exact** observations and over-confident prior.



(a) Likelihood **(b)** MSE **(c)** MSPE



(d) Log Likelihood **(e)** Mean absolute error **(f)** Average variance

Figure 6.9 – IPPA performance given **noisy** observations and over-confident prior.

6.5.3 Discussion

When TPs were applied to informative path planning in this simulation, they scaled the variance in each scenario encompassing under-confident priors, over-confident priors, exact observations and noisy observations.

In every scenario, canonical TPs gave slightly more absolute error than GPs. This is because the TPs mutual information was found to prioritize less effective points. However, due to the other properties of TPs, they generally modelled the ground truth with equal or better likelihood. As the informativeness of an observation using TPs depends on the values observed, the mutual information in this case is prone to giving incorrect expectations. This can easily be addressed by experimenting with different objective functions. Note that, this is not a consequence of TPs giving poorer expected values, as the expected values of TPs and GPs are the same when conditioned on the same points. Also note, this is not a consequence of TPs training less effectively as the same behaviour can be seen when the kernel parameters are fixed (Section A.5).

Generally it was found TPs outperformed the GPs most when given an under-confident prior. This is because the under-confident prior is more susceptible to change in a Bayesian framework. This property is rooted in the the fundamental Bayesian sensor models outlined in (Chapter 3).

In this application, TPs subject to independent error performed comparably to the canonical noisy TPs. When using the under-confident prior, the likelihood was greater in part due to the shape of the marginal t distributions and slightly lower variance, but also as a consequence of using the GP's objective function.

When the TPs subject to independent error were given an over-confident prior, the variance of the data was scaled in a much different way to the canonical TP (Figure 6.9f). The variance increased to start, accounting for the overconfident prior, but then decreased more rapidly eventually being inline with the variance of the GP. Nonetheless, this gave approximately the same likelihood as the other models.

Discussion of Prior Expectation and Kernel Functions

In this simulation, the true mean of the QOI is not given, and thus the prior expectation is assumed to be $\psi(t) = 0$. In this context, the TP does not solve for the variance of the QOI as could be done in more theoretical examples (Figure 5.9). Rather, the TP scales the covariance of the posterior with respect to how different observed values are from the prior's expected value and kernel function. This can be though of as scaling the confidence of the prior.

The geographic data in this simulation also does not have a true kernel function as could be derived in more theoretical examples. Thus the covariance matrix is scaled with respect to an approximate kernel function. This is also partially addressed in the kernel parameter training step.

Although there cannot be performance guarantees in less controlled environments, the scaling and wide tailed properties of TPs were still found to generally maintain or improve performance while providing useful insight into properties of the prior.

Chapter 7

Conclusion

7.1 Thesis Summary

This thesis begins by introducing the relationship between the fields of statistics and classical robotics. In particular, this thesis focuses on the robotic problem of informative path planning, building of Bayesian statistics and information theory.

There are many topics of research in the field of informative path planning, including tailored objective functions, problem specific perception, integration into robotic systems and more. This thesis approaches the problem head on by taking an analytic Bayesian approach.

This thesis first investigates Bayesian sensor models from the perspective of conjugate priors, highlighting the importance of the exponential family in deriving analytic solutions. The informative properties of these models are also discussed and later applied to stochastic processes subject to noisy observations and informative path planning.

In situations where the QOI can be mathematically described as a scalar field, GPs are by far the most popular belief model used by IPPAs. The related work and applications of these are specifically discussed. This thesis provides a detailed technical review of GPs and the conjugate Bayesian properties that allow for the integration of

normal sensor models. This chapter also provides the necessary background to derive and analyse TPs and the proposed TP subject to independent noise. The informative properties of the models and their ability to scale confidence in a Bayesian way are explored in great detail. A novel approach to training the kernel parameters of TPs is also outlined building of the work of Rasmussen et al. (2006) and Thrun et al. (2005).

Lastly, each model was integrated into a simple but representative IPPA. They were then evaluated on a geographic data set of potassium density over NSW Australia. In a novel finding, TPs and TPs subject to independent noise were able to equally or better model the data due to their ability to scale the confidence of the prior and capture outliers.

7.2 Contributions

Firstly, this thesis gives a complete overview of analytic informative path planning, starting from principals of information theory, Bayesian sensor models and conjugate priors, progressing all the way through to an application of informative path planning on real data.

Using this approach, the limitations of canonical noisy TPs were explored and addressed by proposing a model that builds of the normal distribution's semi conjugate prior. This model has been discussed in few works, and must be approximated due to its intractable form. This thesis outlines a novel Monte Carlo sampling method to approximate the model's posterior, outlines a kernel parameter training method and showcases its benefits when compared against canonical noisy TPs.

Lastly, this thesis showcased a novel application of TPs to informative path planning where their unique Bayesian properties were demonstrated.

7.3 Future Work

This section will go into specific areas of this thesis and related work outlining potential areas of future work and and their potential contribution to informative path planning.

Chapter 5 (Equation 5.16) describe a process where separate priors are placed over Σ_x and Σ_z giving what Yu et al. (2007) describes as double robust control, and what Tang et al. (2017) approximate using the second order Taylor series. This model has the potential to scale the covariance of the data and the noise of observation making a *double robust* model. This would allow researchers to place a prior over sensor noise and the covariance of the QOI allowing the agent to scale these parameters in a Bayesian way as observations are collected. This has the potential to make IPPAs less sensitive to these parameter initializations and further correctly scale them.

In Chapter 5 (Figure 5.1) it was shown how using TPs can give a confidence interval over the covariance of the QOI. Appling this to multi-output GPs, it would be interesting to explore how the confidence of different outputs and their correlation may vary. Confidence over correlations may help researchers evaluate hypotheses and could help inform an agent of the next best observation action depending on the agents objective.

Chapter 5 (Section 5.5) discusses and demonstrates kernel parameter training for TPs. Exploring different optimizer to make searching the kernel parament space more robust would be a practical improvement that builds of well studied fields. This would generally stabilize the IPPAs using TPs.

In Chapter 5 (Equation 5.24) Monte-Carlo sampling methods were used to approximate the posterior distribution of a TP with independent noise. This sampling method has the potential to be sped up by exploiting the representation of the multivariate t-distribution shown in Ding (2016)[1]. This would be advantageous for online IPPAs using TPs subject to independent noise.

[1] Equation 2.

In Chapter 6 it was shown the mutual information given by a TP results in a slightly less effective objective function then what is given by a GP. Experimenting with different objective functions would build of a strong body of related work to potentially improve the performance of IPPAs using TPs.

More generally, the multivariate normal distribution has many uses in engineering, planning, robotics and more. It would be interesting to explore how placing a prior over their covariance would affect the performance of these models in their applications.

7.3.1 Robotic Application

The most pertinent future work would be to test TP driven IPPAs on a physical robotic informative path planning problem. Testing the proposed models given physical sensor noise and physically observed values would inform how to best implement the model and how its properties are effected.

Building off the potassium density example (Section 6.4), one potential application would be to mount a Geiger counter to a robotic agent tasked with making radiometric observations in the field. In this context, a Bayesian sensor model would be used to relate radiometric observations and the presence of potassium.

7.4 Concluding Remarks

Analytic Bayesian statistics is an essential tool for building computationally efficient systems robust to uncertainty. Informative path planning is no exception with standard approaches relying on stochastic processes, Bayesian sensor models and information theory. Building of these fundamentals has made for a fruitful avenue of research with practical applications and promising future work.

Appendix A

Supplementary Equations and Figures

A.1 Covariance Equations

Equation A.1 demonstrates how the additive properties of covariance can be used to simplify the covariance of the QOI and an observation subject to Gaussian white noise. This builds off the definitions of Section 4.2.

$$\begin{aligned} \operatorname{cov}(z_i, x_j) &= \operatorname{cov}(x_i + \epsilon_i, x_j) && \text{where } \epsilon_i \sim \mathcal{N}(\epsilon_i; 0, \psi(t_i)) \\ &= \operatorname{cov}(x_i, \epsilon_i) + \operatorname{cov}(x_i, x_j) \\ &= \operatorname{cov}(x_i, x_j) && \text{by noise being independent of the QOI} \\ &= k(t_i, t_j) \end{aligned} \tag{A.1}$$

A.2 Log Likelihood Equations

A.2.1 Multivariate Normal Distribution Log Likelihood

Equation A.2f shows how the log likelihood of a multivariate normal distribution can be rearranged for computational efficiency (Rasmussen et al. (2006) Algorithm 2.1.):

$$\text{Let } K := K_{nn} + \Sigma_z \tag{A.2a}$$

$$\text{Let } Y := \check{Z}_n - \phi(T_n) \tag{A.2b}$$

$$\text{Let } L := \text{Cholesky}(K) \tag{A.2c}$$

$$\text{Let } l_{ii} := \text{the } i^{th} \text{ diagonal of } L \tag{A.2d}$$

$$\begin{aligned} \text{Let } \alpha :&= L^T \backslash (L \backslash Y) \\ &= K^{-1} Y \end{aligned} \tag{A.2e}$$

Then

$$\begin{aligned} \ln(p(\check{Z}_n)) &= -\frac{1}{2} Y^T K^{-1} Y - \frac{1}{2} \ln(|K|) - \frac{n}{2} \ln(2\pi) \\ &= -\frac{1}{2} Y^T \alpha - \frac{1}{2} \ln(|LL^T|) - \frac{n}{2} \ln(2\pi) \\ &= \frac{1}{2} Y\alpha - \sum_{i=1}^{n} \Big(\ln(l_{ii}) \Big) - \frac{n}{2} \ln(2\pi) \end{aligned} \tag{A.2f}$$

The derivative may be calculated as (Rasmussen et al. (2006) Equation 5.9.):

$$\begin{aligned} \frac{\partial}{\partial \theta_1} \ln(p(\check{Z}_n)) &= \frac{1}{2} Y^T K^{-1} \frac{\partial K}{\partial \theta_1} K^{-1} Y - \frac{1}{2} \text{tr}(K^{-1} \frac{\partial K}{\partial \theta_1}) \\ &= \frac{1}{2} \alpha^T \frac{\partial K}{\partial \theta_1} \alpha - \frac{1}{2} \text{tr}(K^{-1} \frac{\partial K}{\partial \theta_1}) \\ &= \frac{1}{2} (\text{tr}\Big(\alpha\alpha^T \frac{\partial K}{\partial \theta_1} \Big) - \text{tr}\Big(K^{-1} \frac{\partial K}{\partial \theta_1} \Big)) \\ &= \frac{1}{2} \text{tr}\Big(\alpha\alpha^T \frac{\partial K}{\partial \theta_1} - K^{-1} \frac{\partial K}{\partial \theta_1} \Big) \\ &= \frac{1}{2} \text{tr}\Big((\alpha\alpha^T - K^{-1}) \frac{\partial K}{\partial \theta_1} \Big) \end{aligned} \tag{A.2g}$$

A.2.2 Multivariate t-Distribution Log Likelihood

Equation Equation A.3 shows how the log likelihood of a multivariate t distribution may be rearranged for computational efficiency:

Let the convenience Equations A.2b A.2c A.2 and A.2 hold except using:

$$\begin{aligned} \text{Let } K &:= \frac{v_0 - 2}{v_0}(K_{nn} + \Sigma_z) \\ &^* = \Psi_{nn} \end{aligned} \tag{A.3a}$$

$$\begin{aligned} \ln(p(\check{Z}_n)) &= g(v_0, n) - \frac{1}{2}\ln(|\Psi_{nn}|) - \frac{v_0 + n}{2}\ln(1 + \frac{(\phi(T_n) - \check{Z}_n)^T \Psi_{nn}^{-1}(\phi(T_n) - \check{Z}_n)}{v_0}) \\ &= g(v_0, n) - \frac{1}{2}\ln(|K|) - \frac{v_0 + n}{2}\ln(1 + \frac{Y^T K^{-1} Y}{v_0}) \\ &= g(v_0, n) - \frac{1}{2}\ln(|LL^T|) - \frac{v_0 + n}{2}\ln(1 + \frac{Y^T \alpha}{v_0}) \\ &= g(v_0, n) - \sum_{i=1}^{n}\Big(\ln(l_{ii})\Big) - \frac{v_0 + n}{2}\ln(1 + \frac{Y^T \alpha}{v_0}) \\ &= g(v_0, n) - \sum_{i=1}^{n}\Big(\ln(l_{ii})\Big) - \frac{v_0 + n}{2}\ln(1 + \frac{(\phi(T_n) - \check{Z}_n)^T \alpha}{v_0}) \end{aligned} \tag{A.3b}$$

The derivative of Equation A.3 may then be written as:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln\big(p(\check{Z})\big) &= \frac{\partial}{\partial\theta}\Big((-\frac{1}{2}\ln(|\Psi_{nn}|)\Big)+\frac{\partial}{\partial\theta}\Big(-\frac{v_0+n}{2}\ln\big(1+\frac{(\phi(T_n)-\check{Z}_n)^T\Psi_{nn}^{-1}(\phi(T_n)-\check{Z}_n)}{v_0}\big)\Big)\\
&= \frac{\partial}{\partial\theta}\Big((-\frac{1}{2}\ln(|K|)\Big)+\frac{\partial}{\partial\theta}\Big(-\frac{v_0+n}{2}\ln\big(1+\frac{Y^TK^{-1}Y}{v_0}\big)\Big)\\
&= \frac{v_0+n}{v_0+Y^TK^{-1}Y}\times\frac{1}{2}Y^TK^{-1}\frac{\partial K}{\partial\theta}K^{-1}Y-\frac{1}{2}\mathrm{tr}(K^{-1}\frac{\partial K}{\partial\theta})\\
&= \frac{v_0+n}{v_0+Y^T\alpha}\times\frac{1}{2}\mathrm{tr}(\alpha\alpha^T\frac{\partial K}{\partial\theta})-\frac{1}{2}\mathrm{tr}(K^{-1}\frac{\partial K}{\partial\theta})\\
&= \frac{1}{2}\mathrm{tr}\Big(\frac{v_0+n}{v_0+Y^T\alpha}\times\alpha\alpha^T\frac{\partial K}{\partial\theta}-K^{-1}\frac{\partial K}{\partial\theta}\Big)\\
&= \frac{1}{2}\mathrm{tr}\Big(\Big(\frac{v_0+n}{v_0+Y^T\alpha}\times\alpha\alpha^T-(LL^T)^{-1}\Big)\frac{\partial K}{\partial\theta}\Big)\\
&= \frac{1}{2}\mathrm{tr}\Big(\Big(\frac{v_0+n}{v_0+(\check{Z}_n-\phi(T_n)^T)\alpha}\times\alpha\alpha^T-(LL^T)^{-1}\Big)\frac{\partial\Psi_{nn}}{\partial\theta}\Big)
\end{aligned} \tag{A.3c}$$

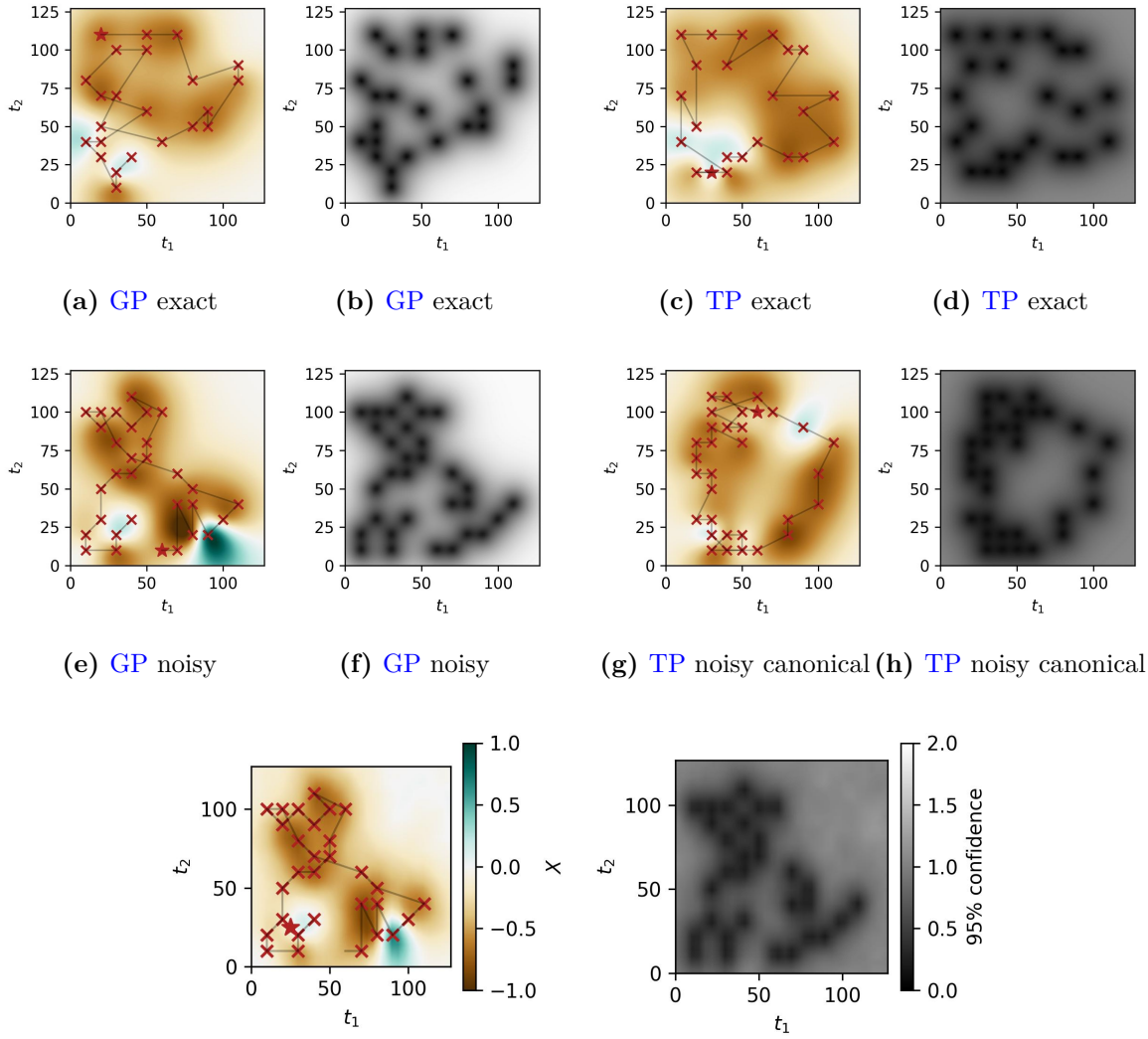
A.3 Mean Square Prediction Error

Equation A.4 shows how the MSPE (Equation 6.12) may be rearranged:

$$\begin{aligned}
&\mathrm{E}[(x-\tau)^2]\\
&=\mathrm{E}[x^2+2\tau x+\tau^2]\\
&=\mathrm{E}[x^2]-2\tau\mathrm{E}[x]+\tau^2\\
&=\mathrm{E}[x^2]+\mathrm{E}[x]^2-\mathrm{E}[x]^2-2\tau\mathrm{E}[x]+\tau^2\\
&=\mathrm{Var}[x]+(\mathrm{E}[x]-\tau)^2
\end{aligned} \tag{A.4}$$

A.4 Informative Path Planning Example Runs

The following two pages show runs of each IPPA given the under-confident and over-confident prior respectively. The ground truth of this region is shown in Figure 6.1e.

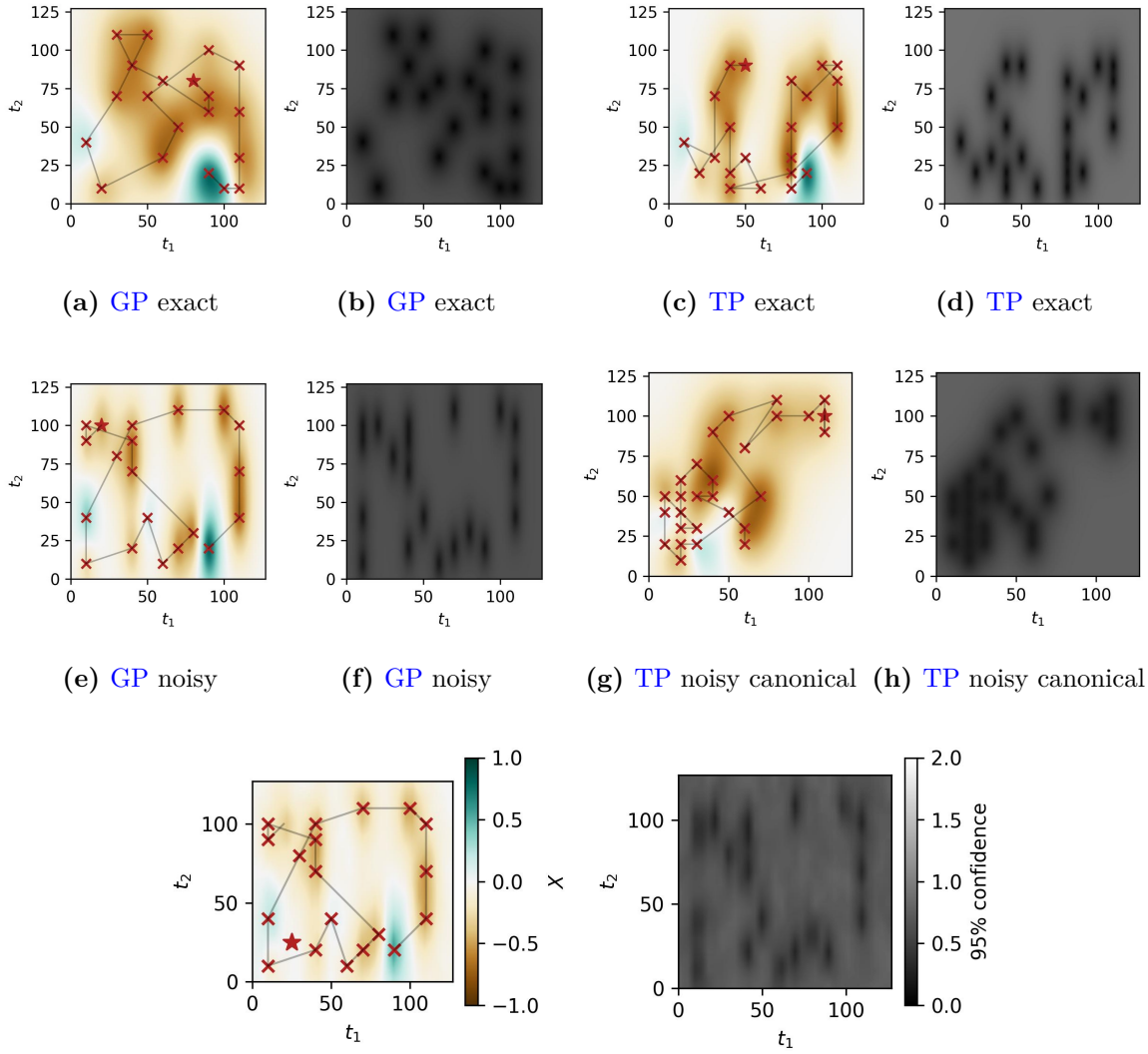


(a) GP exact **(b)** GP exact **(c)** TP exact **(d)** TP exact

(e) GP noisy **(f)** GP noisy **(g)** TP noisy canonical **(h)** TP noisy canonical

(i) TP noisy independent **(j)** TP noisy independent

Figure A.1 – IPPAs of Section 6.4.2 given the under-confident prior. The ground truth is shown in Figure 6.1e.



(a) GP exact **(b)** GP exact **(c)** TP exact **(d)** TP exact

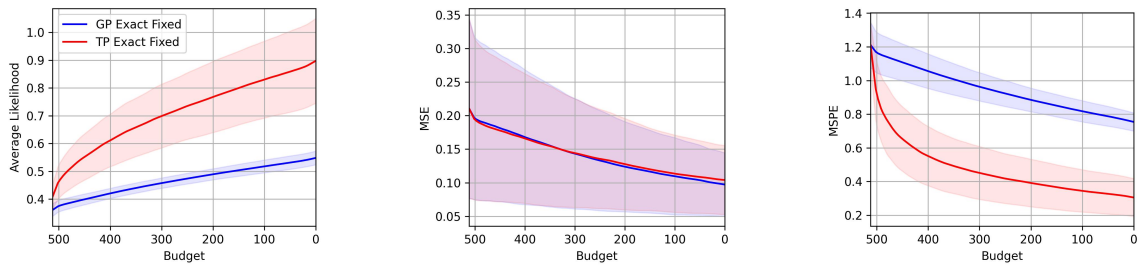
(e) GP noisy **(f)** GP noisy **(g)** TP noisy canonical **(h)** TP noisy canonical

(i) TP noisy independent **(j)** TP noisy independent

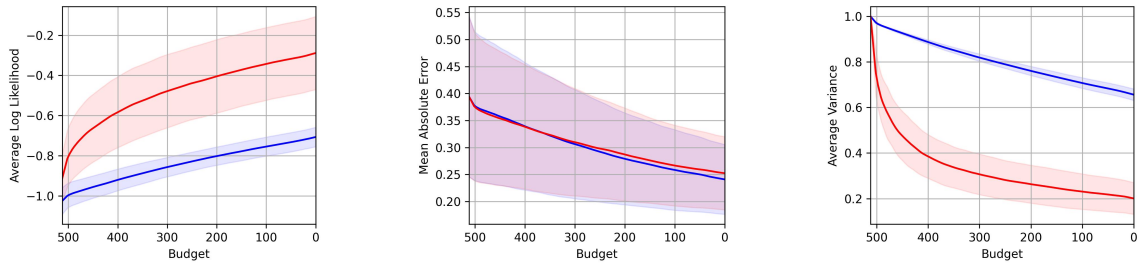
Figure A.2 – IPPAs of Section 6.4.2 given the over-confident prior. The ground truth is shown in Figure 6.1e.

A.5 Results Using Fixed Parameters

The results of Section 6.5 obtained using fixed kernel parameters are presented on the following two pages.

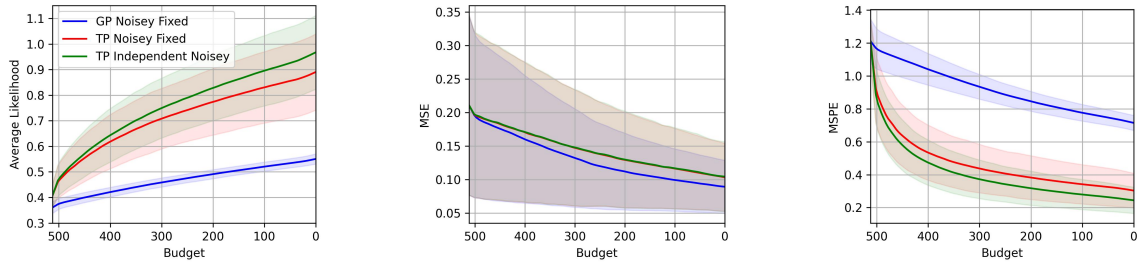


(a) Likelihood **(b)** MSE **(c)** MSPE

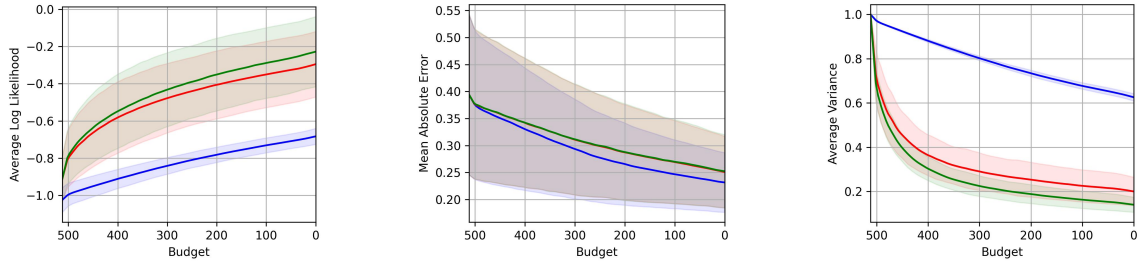


(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure A.3 – IPPA performance given **exact** observations and under-confident prior.

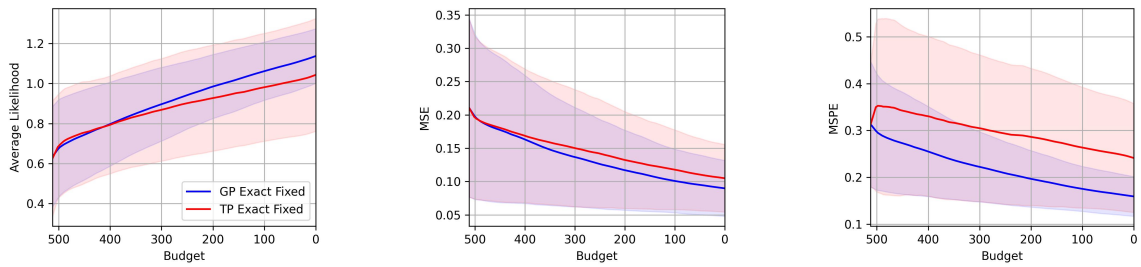


(a) Likelihood **(b)** MSE **(c)** MSPE

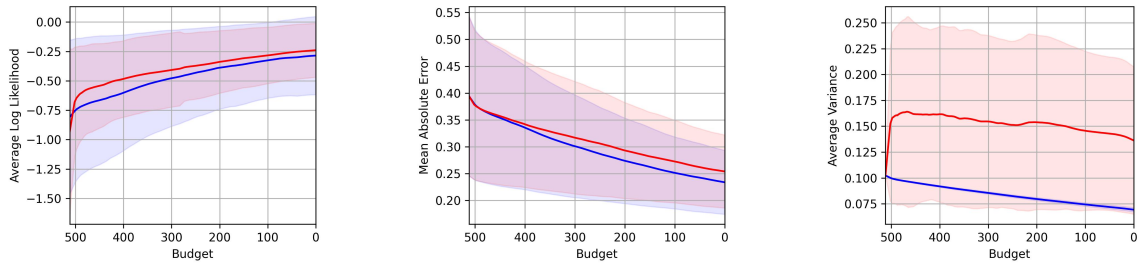


(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure A.4 – IPPA performance given **noisy** observations and under-confident prior.

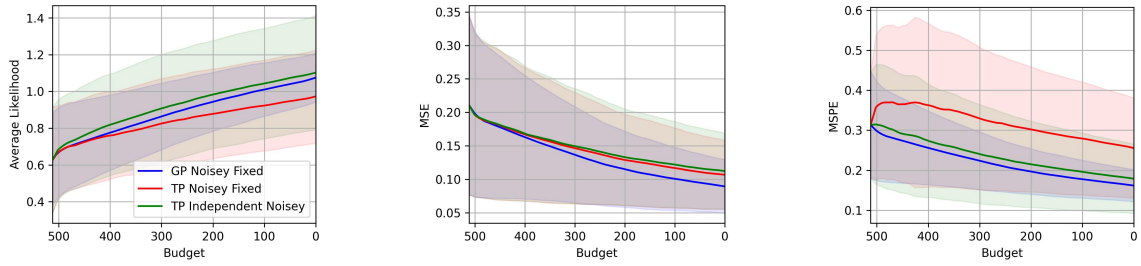


(a) Likelihood **(b)** MSE **(c)** MSPE

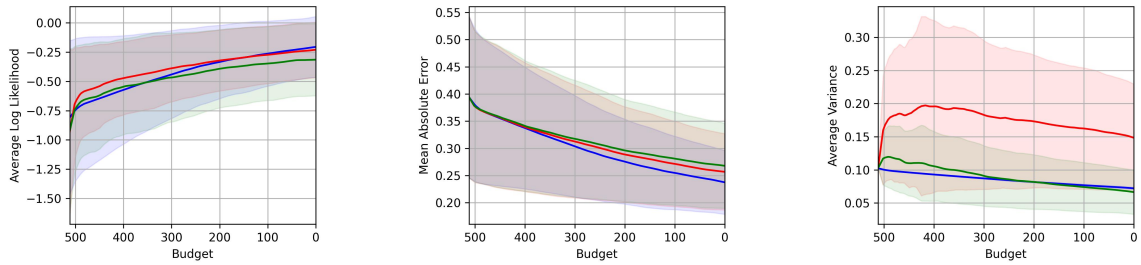


(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure A.5 – IPPA performance given **exact** observations and over-confident prior.



(a) Likelihood **(b)** MSE **(c)** MSPE



(d) Log Likelihood **(e)** Mean absolute error **(f)** Average Variance

Figure A.6 – IPPA performance given **noisy** observations and over-confident prior.

List of References

Ignacio Alvarez, Jarad Niemi, and Matt Simpson. Bayesian inference for a covariance matrix. In *Conference on Applied Statistics in Agriculture*, 2016.

J Ailton A Andrade. On the robustness to outliers of the student-t process. *Scandinavian Journal of Statistics*, 50(2):725–749, 2023.

Reinaldo B. Arellano Valle, J Jacier E. Contreras-Reyes, and Marc G. Genton. Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scandinavian Journal of Statistics*, 40(1):42–62, 2013.

Akash Arora. *Multi-Modal Active Perception for Robotic Information Gathering in Science Missions*. PhD thesis, University of Sydney, 2018.

Jonathan Binney, Andreas Krause, and Gaurav S. Sukhatme. Informative path planning for an autonomous underwater vehicle. In *IEEE International Conference on Robotics and Automation*, pages 4791–4796, 2010.

Weizhe Chen, Roni Khardon, and Lantao Liu. Adaptive robotic information gathering via non-stationary gaussian processes. *The International Journal of Robotics Research*, 43(4):405–436, 2024.

T. M. Cover and Joy A. Thomas. *Elements of information theory*. A Wiley-Interscience publication. Wiley-Interscience, Hoboken, N.J, 2nd edition, 2006. ISBN 9780471748816.

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3): 261–273, 2008. ISSN 1099-4300.

Morris H DeGroot. *Optimal Statistical Decisions*, volume 82 of *Wiley classics library*. Wiley-Interscience, Hoboken, N.J, wcl edition. edition, 2005. ISBN 9-780-47168-0291.

Peng Ding. On the conditional distribution of the multivariate t distribution. *The American Statistician*, 70(3):293–295, 2016.

David Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

Morris L Eaton. The wishart distribution. In *Multivariate Statistics*, volume 53, pages 302–334. Institute of Mathematical Statistics, 2007.

Garry Einicke. *Smoothing, filtering and prediction: Estimating the past, present and future*. Books on Demand, 2012.

Andrew Feutrill and Matthew Roughan. A review of shannon and differential entropy rate estimation. *Entropy*, 23(8), 2021.

Daniel Fink. A compendium of conjugate priors. *Technical Report*, 1997.

P. Michael Furlong. *Foraging, Prospecting, and Falsification - Improving Three Aspects of Autonomous Science*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, May 2018.

Xiang Gao and Tao Zhang. Unsupervised learning to detect loops using deep neural networks for visual slam system. *Autonomous robots*, 41:1–18, 2017.

Alberto Candela Garza. *Bayesian Models for Science-Driven Robotic Exploration*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, September 2021.

Andrew. Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki. Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 3ed edition, 2013. ISBN 9781439898208.

Marc G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312, Mar 2002. ISSN 1532-4435.

Leila Golshani. The rate of entropy for gaussian processes. *Journal of Statistical Research of Iran JSRI*, 12(1):71–82, 2015.

Eduardo Gutiérrez-Peña and Pietro Muliere. Conjugate priors represent strong pre-experimental assumptions. *Scandinavian journal of statistics*, 31(2):235–246, 2004.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

Nicholas Harrison, Nathan Wallace, and Salah Sukkarieh. Automated testing of spatially-dependent environmental hypotheses through active transfer learning. In *IEEE International Conference on Robotics - arXiv preprint*, 2024.

Grant Hillier and Raymond Kan. Properties of the inverse of a noncentral wishart matrix. *Econometric Theory*, 38(6):1092–1116, 2022.

David Hindley. Mathematical details for mean squared error of prediction. In *Claims Reserving in General Insurance*, International Series on Actuarial Science, page 465–470. Cambridge University Press, 2017.

Gregory Hitz, Enric Galceran, Marie-Ève Garneau, François Pomerleau, and Roland Siegwart. Adaptive continuous-space informative path planning for online environmental monitoring. *Journal of Field Robotics*, 34(8):1427–1449, 2017.

Roger A Horn and Fuzhen Zhang. Basic properties of the schur complement. pages 17–46, 2005.

Yulong Huang, Yonggang Zhang, Ning Li, Zhemin Wu, and Jonathon A. Chambers. A novel robust student's t-based kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 53(3):1545–1554, 2017.

Shields Jackson. *Planning Informative Benthic AUV Surveys*. PhD thesis, University of Sydney, 2023.

Kalvik Jakkala and Srinivas Akella. Multi-robot informative path planning from regression with sparse gaussian processes. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12382–12388, 2024.

Chen Quin Lam. *Sequential adaptive designs in computer experiments for response surface model fit*. PhD thesis, The Ohio State University, 2008.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Wenhao Luo and Katia Sycara. Adaptive sampling and online learning in multi-robot sensor coverage with mixture of gaussian processes. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6359–6364, 2018.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 05 1992. ISSN 0899-7667.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. ISBN 9780521642989.

Roman Marchant and Fabio Ramos. Bayesian optimisation for intelligent environmental monitoring. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2242–2249, 2012.

S.J. Matthews. Nsw statewide radiometric merge. Geological Survey of New South Wales, Department of Regional NSWs, 2023.

Merriam-Webster. tantalizing. 2004.

Joseph Victor. Michalowicz, Jonathan Michael. Nichols, and Frank. Bucholtz. *Handbook of differential entropy*. Taylor & Francis, Boca Raton, 1st edition, 2014. ISBN 0429072244.

Rajat Mishra, Mandar Chitre, and Sanjay Swarup. Online informative path planning using sparse gaussian processes. In *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pages 1–5, 2018.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *Technical Report*, 2007.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. ISBN 978-0-262-01802-9.

Javier Muñoz, Blanca López, Fernando Quevedo, Santiago Garrido, Concepción A. Monje, and Luis E. Moreno. Gaussian processes and fast marching square based informative path planning. *Engineering Applications of Artificial Intelligence*, 121:106054, 2023. ISSN 0952-1976.

Dimitris Papadimitriou and Somayeh Sojoudi. Control and uncertainty propagation in the presence of outliers by utilizing student-t process regression. In *2022 American Control Conference (ACC)*, pages 2748–2754. IEEE, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Marija Popović, Teresa Vidal-Calleja, Gregory Hitz, Jen Jen Chung, Inkyu Sa, Roland Siegwart, and Juan Nieto. An informative path planning framework for uav-based terrain monitoring. *Autonomous Robots*, 44:889–911, 2020.

Marija Popović, Gregory Hitz, Juan Nieto, Inkyu Sa, Roland Siegwart, and Enric Galceran. Online informative path planning for active classification using uavs. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5753–5758, 2017.

Howard Raiffa and Robert. Schlaifer. *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University, Boston, 1961. ISBN 978-0-471-38349-9.

Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

Isabel M. Rayas Fernández, Christopher E. Denniston, David A. Caron, and Gaurav S. Sukhatme. Informative path planning to estimate quantiles for environmental analysis. *IEEE Robotics and Automation Letters*, 7(4): 10280–10287, 2022.

Gernot Roetzer. *Efficient and Scalable Inference for Generalized Student-t Process Models*. PhD thesis, Trinity College Dublin, 2020.

Michael Roth, Tohid Ardeshiri, Emre Özkan, and Fredrik Gustafsson. Robust bayesian filtering and smoothing using student's t distribution. *ArXiv*, abs/1703.02428, 2017.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 877–885, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3: 54–70, 2023. ISSN 2667-2413.

Hamid Taheri and Zhao Chun Xia. Slam; definition and evolution. *Engineering Applications of Artificial Intelligence*, 97, 2021.

Qingtao Tang, Li Niu, Yisen Wang, Tao Dai, Wangpeng An, Jianfei Cai, and Shu-Tao Xia. Student-t process regression with student-t likelihood. In *International Joint Conference on Artificial Intelligence*, pages 2822–2828, 2017.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT Press, Cambridge, Mass., 2005. ISBN 0262201623 9780262201629.

B. D. Tracey and D. H. Wolpert. *Upgrading from gaussian processes to student's-t processes*. 2018.

Vinita Vasudevan and M. Ramakrishna. A hierarchical singular value decomposition algorithm for low rank matrices. *ArXiv*, abs/1710.02812, 2017.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0:

Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Alberto Viseras. *Distributed Multi-Robot Exploration under Complex Constraints*. PhD thesis, Pablo de Olavide University, 2018.

Nathan Daniel Wallace. *Energy-aware Planning and Control of Off-road Wheeled Mobile Robots*. PhD thesis, University of Sydney, 2020.

Zhanfeng Wang, Jian Qing Shi, and Youngjo Lee. Extended t-process regression models. *Journal of Statistical Planning and Inference*, 189:38–60, 2017. ISSN 0378-3758.

Yongyong Wei and Rong Zheng. Informative path planning for mobile sensing with reinforcement learning. *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pages 864–873, 2020.

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

Shipeng Yu, Volker Tresp, and Kai Yu. Robust multi-task learning with t-processes. In *international conference conference on Machine learning*, pages 1103–1110, 2007.

Yu Zhang and Dit–Yan Yeung. Multi-task learning using generalized t process. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 964–971, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.