NTNU
Norwegian University of
Science and Technology

Estimating the Value of Information Using Bayesian Optimization with Gaussian Process Surrogate Models

An Application to Failure Rates at Offshore Wind Farms

Hans Olav Vogt Myklebust

Master of Science in Physics and Mathematics
Submission date: June 2018
Supervisor: Jo Eidsvik, IMF

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This master thesis makes up the course TMA4900 for the Industrial mathematics study on NTNU. The focus in this thesis is the estimation of value of information in a complex decision situation with uncertain variables involved. All work related to this thesis has been performed during spring 2018.

I would like to express my gratitude to my supervisor, Jo Eidsvik at NTNU, for constructive discussions and feedback during the process of writing this thesis. I would also like to thank Iver Bakken Sperstad at Sintef Energy for his help and guidance with the simulation tool used in this thesis.

Trondheim, June 2018
Hans Olav Vogt Myklebust

Abstract

Finding operation and maintenance (O&M) strategies that increase the profitability of an offshore wind farm is essential in order to be competitive with other sources of renewable energy. An O&M strategy is characterized by a set of decision variables, such as vessel fleet and the number of available technicians. The presence of uncertain variables that influence the profitability of an O&M strategy, such as varying weather conditions, makes the search for good strategies difficult. Information about the uncertain variables will help to find good strategies, but such knowledge often comes with a price. It is therefore of great interest to find out whether buying information is a worthwhile investment.

By utilizing a simulation tool, the profitability of different O&M strategies can be investigated. The set of possible strategies are, however, large so that only a small subset may be explored. Bayesian optimization with a Gaussian process surrogate model is therefore used to find favorable O&M strategies in a high dimensional input space. After obtaining these strategies, the value of information for an interrelated uncertain variable was estimated. The search for favorable O&M strategies and estimation of the value of information for the failure rate for a wind turbine component category is demonstrated in a relevant case from the offshore wind industry. The strategies identified from the Bayesian optimization appear reasonable and the estimated value of information suggests that information gathering could be worthwhile, depending on the price of the information.

Sammendrag

For at offshore vindenergi skal være et konkurransedyktig alternativ til andre kilder til fornybar energi, er det viktig å finne drift- og vedlikeholdstrategier som øker lønnsomheten. En strategi kjennetegnes av et sett av beslutningsvariabler som fartøysflåte eller antall tilgjengelige teknikere. Usikre variabler som påvirker lønnsomheten til en strategi, for eksempel varierende værforhold, gjør det vanskelig å søke etter gode strategier. Informasjon om de usikre variablene vil bidra til å finne gode strategier, men slik kunnskap kommer ofte med en pris. Vi er derfor interessert i finne ut om kjøp av informasjon kan være en god investering.

Ved å bruke et simuleringsverktøy kan lønnsomheten til forskjellige strategier undersøkes. Det er imidlertid mange mulig strategier, slik at bare en liten del av mulighetsrommet kan utforskes. Bayesiansk optimering med en Gaussisk prosess surrogatmodell brukes til å finne gunstige strategier i et høy-dimensjonalt variabelrom. Etter å ha funnet disse strategiene ble verdien av informasjon for en relatert usikker variabel estimert. Søket etter gunstige strategier og estimering av verdien av informasjonen for en feilrate for en komponentkategori i en vindturbin er demonstrert i et relevant eksempel fra offshore vindindustri. Strategiene funnet fra Bayesiansk optimering synes å være rimelige, og den estimerte verdien av informasjon tyder på at innsamling av informasjon kan være profitabelt, avhengig av prisen på informasjonen.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The costs related to operation and maintenance (O&M) tasks constitute a major part of the overall expenses of an offshore wind farm. According to Welte et al. (2018), the O&M cost of an offshore wind turbine lies between 12-32% of the total income it provides. The electricity generated by offshore wind turbines is estimated to be 2.6 times more expensive than electricity generated by onshore wind turbines (Shafiee, 2015). Therefore, if we want offshore wind power to be competitive with other sources of renewable energy such as onshore wind, the costs related to O&M tasks must be reduced.

A strategy for O&M consists of several decisions, such as

- Location of the harbor, e.g distance to the wind farm
- Vessel fleet mix
- Number of technicians

In addition to the decision problems, we must also consider the vast amount of uncertainty involved. Some of the uncertainties are

- Wave height
- Wind speed
- Component failure rates.

From the mentioned examples of uncertainties and O&M strategy decisions, it is clear that they are interrelated. A good strategy for a location with rough weather might consist of robust vessels that can handle high waves and wind speeds, but travels at lower speeds. Locations with calm weather might favor travel speed above robustness. Typically a location will experience

both rough and calm weather, so a good O&M strategy must find a balance between the different weather conditions. Similarly, a wind farm where turbines are frequently breaking down requires a good amount of technicians and vessels, while a wind farm with few failures can perform O&M tasks with fewer technicians and vessels. Again, a good strategy must find a balance that works well on average.

Because of the mentioned interrelations, it is of interest to know as much as possible about the uncertain variables before choosing an O&M strategy. For instance, in the planning phase of a wind farm, it would be of interest to locate and orientate the wind farm so that the power production is maximized. A key uncertainty here is the wind direction. It might be possible to retrieve some information about historical wind directions so that the turbines are oriented the way that gives the highest expected revenue.

Another important uncertainty that we will focus on in this thesis is the component failure rates. Failures of different severity will occur from time to time, and they require both technicians and a vessel. The frequency of these failures is therefore of great interest when the number of technicians and vessels shall be decided. The manufacturer of a particular component might provide an estimate of the failure rate, but trusting this estimate might be a bit naive. The master thesis by Haraldsdottir and Sandstrom (2016) concludes that the lifetime of the main bearing depends heavily on the wind speed. The lifetime is shortest at the rated wind speed, which is the lowest wind speed at which the turbine reaches the highest possible power output. For higher and lower wind speeds the component lifetime is higher. Assuming that other components have failure rates that depend on the operating conditions, e.g. temperature or humidity therefore seems reasonable.

The search for favorable O&M strategies requires an understanding of the relationship between input, e.g O&M strategies, and output, e.g profit of wind farm. This understanding can be acquired by performing experiments. Physical experiments are intractable since it will require a lot of trial and error before a good, and probably sub-optimal O&M strategy is found. This motivates the use of simulation models, that mimic these physical experiments and provide the desired understanding. Several simulation tools for offshore wind farms have been developed, see for instance Hofmann (2011) for an overview. We will use NOWIcob (**N**orwegian **o**ffshore **wi**nd power life cycle **co**st and **b**enefit model) simulation tool in this thesis.

Much work has been done in the context of finding good O&M strategies for offshore wind farms using NOWIcob. For instance, in Sperstad et al. (2016) three decision problems with the aim to optimize the O&M strategies are investigated. In the master thesis by Gallala (2016), artificial neural networks are used as a surrogate model in order to find favorable O&M strategies. These contributions focus only on finding optimal O&M strategies. They do not look at the possibility of gathering information about uncertain variables that could improve the suggested optimal O&M strategy. The lack of such studies motivates this thesis.

1.2 Optimization and Surrogate Models

By performing several experiments, or simulations, with different inputs we can relate O&M strategies and the profit. However, the simulations in NOWIcob are time-consuming, so we can only explore a subset of possible input combinations. If we, for instance, want to find the strategy that maximizes the profit, without excessive evaluations, we can use Bayesian optimization. In this approach, a statistical model is used to approximate the relation between input and output to search for the global optima. The statistical model is fast to evaluate, which means that a search for the approximate global optima can be done efficiently. Bayesian optimization has a wide range of applications, for instance recommendation systems and speech recognition (Shahriari et al., 2016).

In this thesis, we will approximate the relationship between O&M strategies and the profit with a Gaussian process (GP). GPs are a popular tool for approximating unknown functions because of advantages such as the flexibility to input variables, ease of interpretation and the model provides uncertainty estimates. Other choices for surrogate model exists, for instance, artificial neural networks (Gallala, 2016). Artificial neural networks are more flexible to the input variables than GPs, but are more difficult to interpret and do not provide uncertainty estimates.

For efficient optimization of the approximation, we choose points for evaluation that are more promising based on the GP. We will discuss two methods for choosing new points here, probability of improvement and expected improvement. Optimizing the statistical approximation is done by sequentially evaluating new points, then fit a GP to the observations, use this GP to select new points for evaluation, and so on.

1.3 Value of Information

When an approximate relation is established for all alternatives and failure rates, we can compare strategies with and without the knowledge about the failure rate. We can then calculate an estimate of what we expect the profit to be when we know the failure rate and the profit averaged over different failure rates when the failure rate is unknown. The difference in these two estimates is the value of information (VOI), which is the main focus of this thesis. If we know the failure rate, we are likely to find a more optimal O&M strategy than without this knowledge. However, we will only gather this information if its price is less than our estimate of its value.

A thorough introduction to VOI and its application to the earth sciences can be found in Eidsvik et al. (2015). Other contexts where VOI analysis have been shown useful is the petroleum industry (Bratvold et al., 2009) and hydropower production (Ødegård et al., 2017). VOI analysis has not yet received much attention in the context of offshore wind farms. Seyr and Muskulus (2016) estimates the value of knowing the repair times for wind turbines, which can be used to improve the scheduling of maintenance. This was a high-level study only considering a small

subset of possible vessel fleet mix and number of technicians.

1.4 Contribution and Outline of Thesis

The main focus in this thesis is the estimation of the VOI for a component failure rate in offshore wind farms. Knowledge about the failure rate will help us finding more optimal O&M strategies which will make offshore wind power more competitive. Earlier work focuses on finding favorable O&M strategies, where it is assumed that the failure rate is fixed. It is reasonable to assume that the failure rate is unknown and that information about this quantity is available.

The rest of this thesis is outlined as follows: In Chapter 2 Bayesian optimization is discussed. GPs are introduced in Section 2.2, with focus on how we can approximate an unknown function. Methods for selection of new points for efficient optimization of the unknown function is provided in Section 2.3. In Chapter 3 decision theory is briefly introduced and the concept of VOI in a decision situation is discussed. Analytic expressions for VOI are developed in Section 3.2, and estimation of VOI is discussed in Section 3.3. In Section 4 NOWIcob is introduced and in Section 4.2 we formulate the O&M decision problem, and how the VOI can be estimated in this case. Results from this case study are presented in Chapter 5. In Chapter 6 the key findings in this thesis are presented and some suggestions to further work are provided.

Chapter 2

Bayesian Optimization

2.1 Introduction

Function optimization is of interest in many technical and scientific applications, and is a widely studied field in mathematics. We consider an objective function $f : \mathcal{X} \to \mathbb{R}$, that we want to maximize on a domain $\Omega \subset \mathcal{X}$. That is, we want to find

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \Omega}{\operatorname{argmax}} f(\boldsymbol{x}), \tag{2.1}$$

where boldface is used to emphasize that $\boldsymbol{x}$ is a vector. If the function and the gradient of the function are known, we can use one of many existing algorithms for maximizing f, for instance gradient descent or BFGS (Snyman, 2005). If the gradient is unknown, there are also existing algorithms, such as grid search or random search. In some cases, there exists no analytical expression for the function, and evaluations of f are typically time-consuming. Such functions are often referred to as "black-box functions", and optimization of these are more complicated. Bayesian optimization is a strategy for optimizing a black-box function in a sequential way.

The idea with Bayesian optimization is to use a surrogate model for f, that is a statistical approximation, and try to optimize this instead of f. There are many options for the surrogate model in the context of Bayesian optimization, for instance polynomial regression models (Queipo et al., 2005), artificial neural networks (Snoek et al., 2015) and Gaussian processes (Snoek et al., 2012). Gaussian processes (GPs) are perhaps the most popular choice for surrogate model and is what we will use in this thesis. The second part of Bayesian optimization is to use acquisition functions to determine where to evaluate the function next so that the number of evaluations is kept to a minimum. Since a function evaluation is expensive, we want a mechanism that helps us explore the areas where f is expected to be high. This exploration is done with respect to the fitted surrogate model.

GPs will be introduced in Section 2.2 and in Section 2.3, two acquisition functions will be derived and their weaknesses and strengths will be discussed. To get a better understanding of

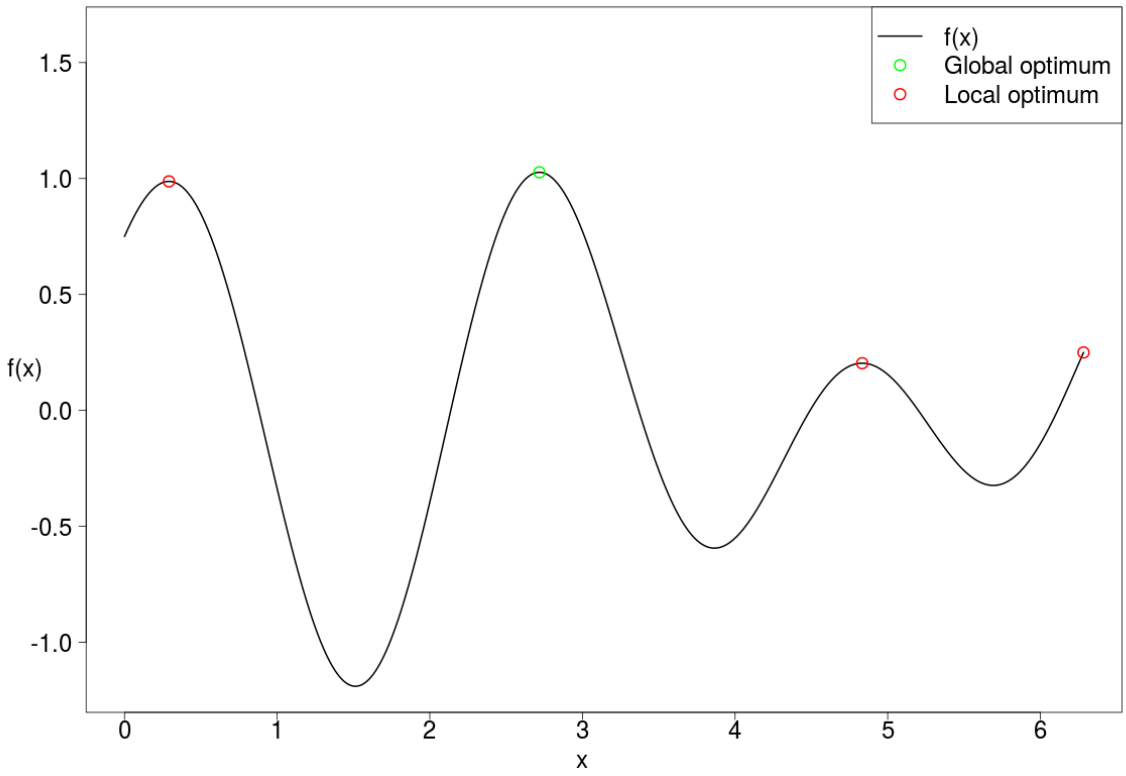


Figure 2.1: Plot of the function in (2.2). Global and local maximums are indicated.

the presented topics a running example will be used and is presented in the following.

2.1.1 Running Example

We will consider a function $f : \Omega \to \mathbb{R}$ given by

$$f(x) = \frac{1}{2}\cos(2x) + \frac{1}{2}\sin(3x) + \frac{1}{4}\cos(2.5x), \qquad 0 \leq x \leq 2\pi. \tag{2.2}$$

A plot of (2.2) is shown in Figure 2.1. We observe that the function has one global optimum and three local optimas, so it is not a straight forward function to optimize.

In Section 2.2 we will treat $f(x)$ as an unknown function and try to fit a GP model when we have some observations/evaluations from this function. In Section 2.3 we will see how new samples should be drawn in order to find the point that maximizes this assumed unknown function.

2.2 Gaussian Processes

In this section, we will introduce Gaussian processes, and see how they may be used to create a statistical approximation to an unknown function f given some training data/evaluations. First,

we will look at a basic GP with a simple mean and covariance function and see how we can use this model to generate a random function. Next, we will see how evaluations of the unknown function can be used to fit a GP. From there we move on to introducing hyperparameters in Section 2.2.1 that makes the model more adaptive to different types of functions, input data, covariance structures and noise. Finally, in Section 2.2.2, some more advanced aspects of GPs will be briefly discussed.

A GP is defined as a stochastic process, where every finite subset of elements has a multivariate normal distribution (Rasmussen, 2004). A GP is fully specified by its mean function, $\mu(\boldsymbol{x})$, and covariance function, $\Sigma(\boldsymbol{x}, \boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, and will be denoted here as

$$Y \sim \mathcal{GP}(\mu, \Sigma). \tag{2.3}$$

This can be seen as a generalization of a Gaussian distribution where the mean is a vector and the covariance is a matrix (Rasmussen, 2004). A GP can be thought of as a distribution over functions, while a Gaussian distribution is over vectors. GPs are therefore suitable to perform regression on unknown functions.

A key ingredient in the specification of a GP is the covariance structure, often given by the covariance function depending on the distance between points. The choice of covariance function encodes our prior knowledge about the unknown function, f, in particular the spatial dependency between variables. It will also, to some extent, determine the shape of the estimated function, as we will see later in an example in Section 2.2.2. A covariance function is a function of two p-dimensional variables, $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, that gives us the covariance between the function value at these two points in space. One widely used class of covariance function is the *stationary covariance functions*, which are a function of the Euclidean distance $d = \|\boldsymbol{x} - \boldsymbol{x}'\|$ between two points. Consider the following covariance function,

$$\mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = \Sigma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\}, \tag{2.4}$$

known as the squared exponential. This means that the covariance between $Y(\boldsymbol{x})$ and $Y(\boldsymbol{x}')$ decays as $\boldsymbol{x}$ and $\boldsymbol{x}'$ becomes farther apart in space. Note that $\Sigma(\boldsymbol{x}, \boldsymbol{x}) = 1$ and that $\Sigma(\boldsymbol{x}, \boldsymbol{x}') < 1$ for $\boldsymbol{x} \neq \boldsymbol{x}'$. The covariance function in (2.4) is a special case of the more advanced covariance functions we will look at in Section 2.2.1, where we allow $\Sigma(\boldsymbol{x}, \boldsymbol{x}') > 1$. The covariance function may take many different forms, but the function must produce a positive definite matrix to be a valid covariance matrix in a multivariate normal distribution.

To see how a GP can be used to perform regression, we will first see how they can be used to generate random data that follows a smooth functional relationship. Suppose we have

- A set of x-values, $(x_1, ..., x_n)$, where x_i is a scalar, not vector.
- A mean function, $\mu(x_i) = 0$ for $i = 1, ..., n$
- A covariance matrix Σ, where each entry $\Sigma_{ij} = \Sigma(x_i, x_j)$ is given by (2.4)

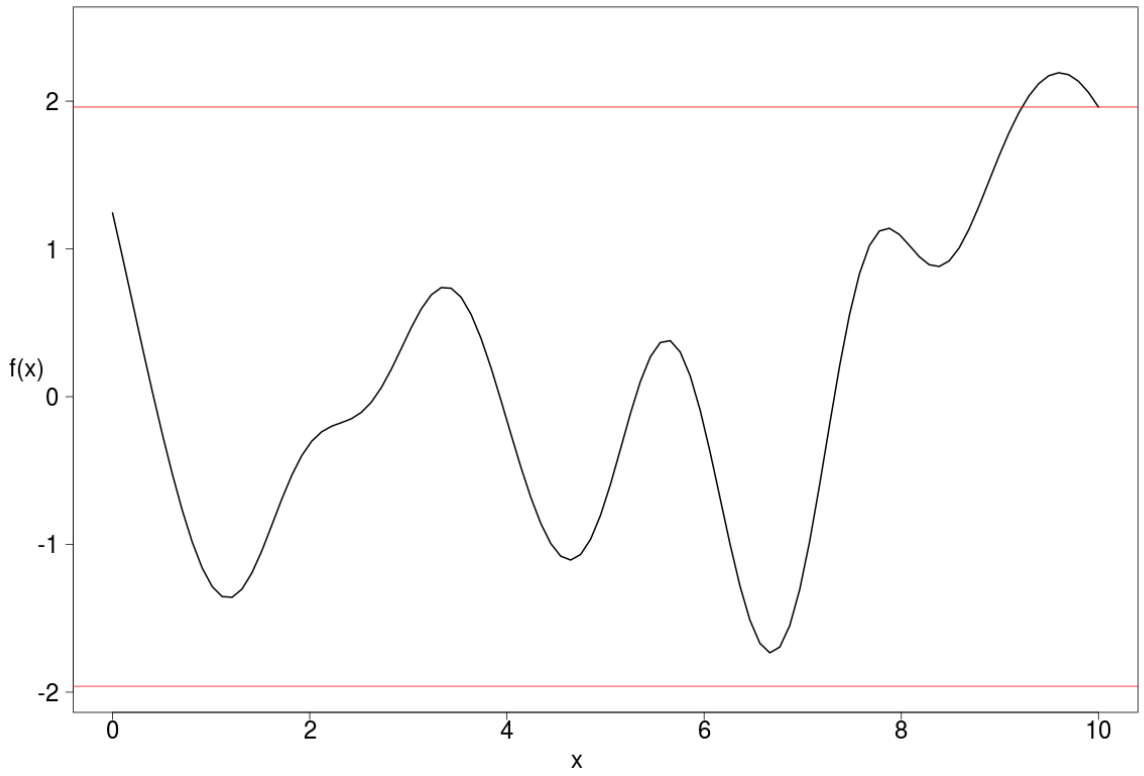


Figure 2.2: One realization of a GP with zero mean and covariance structure given in (2.4). The red lines indicate the 2.5% and 97.5% quantile of the $N(0,1)$ distribution.

Then we can draw an n-variate realization from a multivariate normal distribution, $Y \sim N_n(0, \Sigma_n)$, and plot the result for visualization. The realization can be seen in Figure 2.2. We observe that the function is smooth over the whole domain, and that Y lies between the horizontal red lines for most x-values. This is because the scale of the covariance is 1, so a 95% confidence interval for Y is then $[-1.96, 1.96]$. The smoothness is because of the choice of covariance function, making points close to each other highly correlated. Points far apart are essentially uncorrelated.

We are now able to generate samples from a random function, $Y(\boldsymbol{x}) \sim \mathcal{GP}(\mu, \Sigma)$. Assume next that we have evaluations of a function via observation pairs, $\mathcal{D}_n = \{X_n, Y_n)\}$. Here, X_n is a $n \times p$ matrix, where n is the number of observations and p is the number of variables. Y_n is a $n \times 1$ vector with function values. We are then interested in knowing the random function that could explain those values and predict future values. That is, we are interested in the conditional distribution $Y(\boldsymbol{x}) \mid \mathcal{D}_n$. If $Y(\boldsymbol{x}) \sim \mathcal{GP}(\mu, \Sigma)$ is the prior, then $Y(\boldsymbol{x}) \mid \mathcal{D}_n$ is the posterior. We recall from the definition of a GP that every finite subset of observations has a multivariate normal distribution specified by the mean and covariance function. Therefore, $Y(\boldsymbol{x})$ and $\mathcal{D}_n$ have a joint multivariate normal distribution. The posterior, $Y(\boldsymbol{x}) \mid \mathcal{D}_n$, is thus also multivariate normal. The conditional mean and covariance for a multivariate normal distribution are well known, but we will include it here for the sake of completeness. If we have a random vector $\boldsymbol{x}$,

and we partition it as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \tag{2.5}$$

with mean and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{2.6}$$

then the conditional distribution $\boldsymbol{x}_1 \mid \boldsymbol{x}_2$ is multivariate normal with mean and covariance,

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \\ \bar{\Sigma} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned} \tag{2.7}$$

We can use this to derive the posterior mean and covariance for $Y(\boldsymbol{x}) \mid \mathcal{D}_n$. Let Ω denote the set of feasible $\boldsymbol{x}$-locations, and X_n is the locations that are already evaluated with corresponding values Y_n. We can then state the conditional distribution as

$$Y(\Omega) \mid \mathcal{D}_n \sim N_m\left(\boldsymbol{\mu}(\Omega), \Sigma(\Omega)\right), \tag{2.8}$$

where

$$\begin{aligned} \boldsymbol{\mu}(\Omega) &= \Sigma(\Omega, X_n)\Sigma(X_n, X_n)^{-1}Y_n, \\ \Sigma(\Omega) &= \Sigma(\Omega, \Omega) - \Sigma(\Omega, X_n)\Sigma(X_n, X_n)^{-1}\Sigma(X_n, \Omega). \end{aligned} \tag{2.9}$$

Here, $\Sigma(\Omega, \Omega)$ is the $m \times m$ covariance matrix for the unobserved locations in Ω, where m is the number of feasible unobserved points, $\Sigma(X_n, X_n)$ is the $n \times n$ covariance matrix for the locations X_n that are observed, and $\Sigma(X_n, \Omega) = \Sigma(\Omega, X_n)^T$ is the covariance matrix for the unobserved and observed locations. We now have the posterior GP

$$Y(\Omega) \mid \mathcal{D}_n \sim \mathcal{GP}(\boldsymbol{\mu}(\Omega), \Sigma(\Omega)). \tag{2.10}$$

Consider again the running example defined in (2.2). We will use a GP prior and use $n = 6$ observations to find the posterior GP. We will use prior mean $\mu = 0$ and the covariance function in (2.4). By using the equations in (2.9) we find the posterior mean and covariance for $x \in \Omega$, which is the grid from $x = 0$ to $x = 2\pi$ with step size $\frac{2\pi}{100}$. The posterior mean with a 95% confidence interval and the true function is plotted in Figure 2.3. We see that the posterior mean interpolates the observed points of the true function, and to some extent captures the shape of $f(x)$. We expect $f(x)$ to be within the confidence interval in 95% of the domain, which it also is with a good margin. We observe that the uncertainty gets larger far from observations.

In Figure 2.4, we see the same as in Figure 2.3, but we have extended the set of feasible points a bit. The observations are the same as before. We see that the mean goes to zero, which

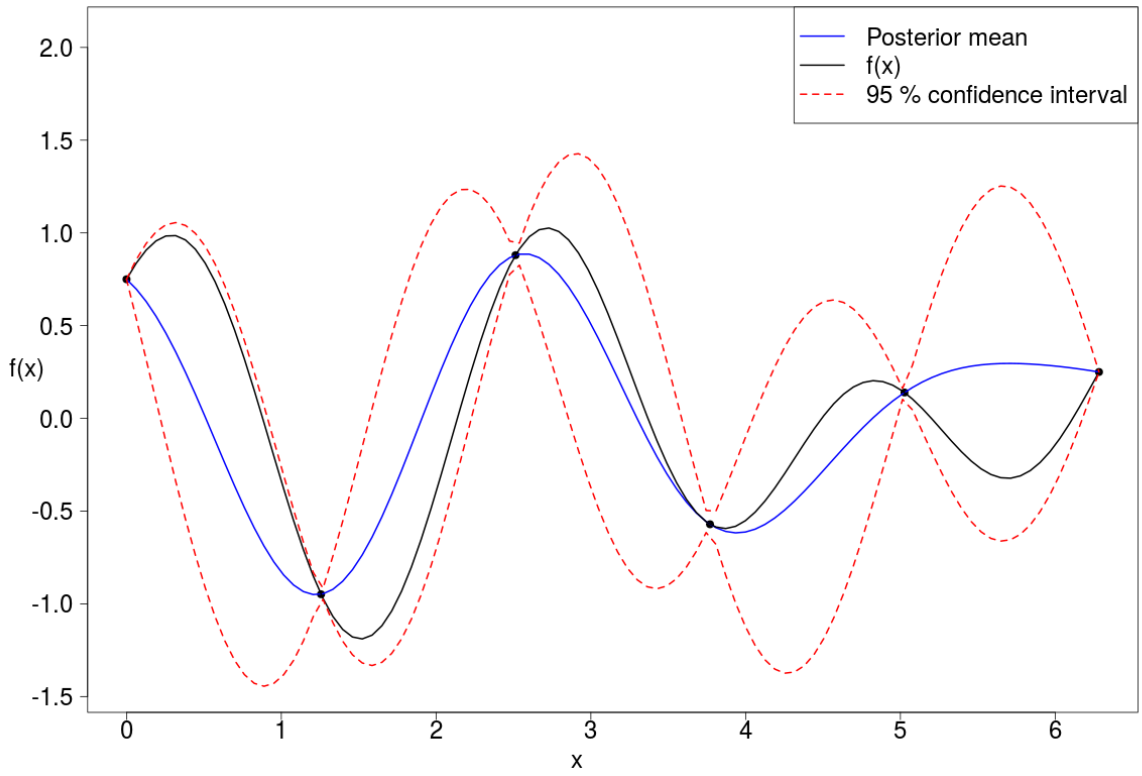


Figure 2.3: Posterior mean, 95% confidence interval and the true function, f. The observed values of f is plotted as black points.

is the prior value, when the distance to the nearest observation increases. This is because we are in an area where we know very little about the unknown function since we have no observations, and our best guess about it is the prior. Also, the upper and lower confidence bound converges toward a limit, namely the 2.5% and 97.5% quantiles. We should, therefore, be careful to trust predictions far outside the range of our observations.

2.2.1 Gaussian Process Hyperparameters

Thus far we have only considered somewhat simple GPs, and we have not estimated a single parameter yet. Introducing parameters may provide a better fit for our GP to the observed examples, but will also make the fitting process more complicated. We will in the following sections introduce three types of hyperparameters - scale, nugget and length scale. The scale parameter handles process variability, allowing the covariance to be larger than one. The nugget parameter is a way of handling noisy observations, as is often the case. The length scale parameter is an extension to the covariance function so that it takes variables having different spatial dependence into account.

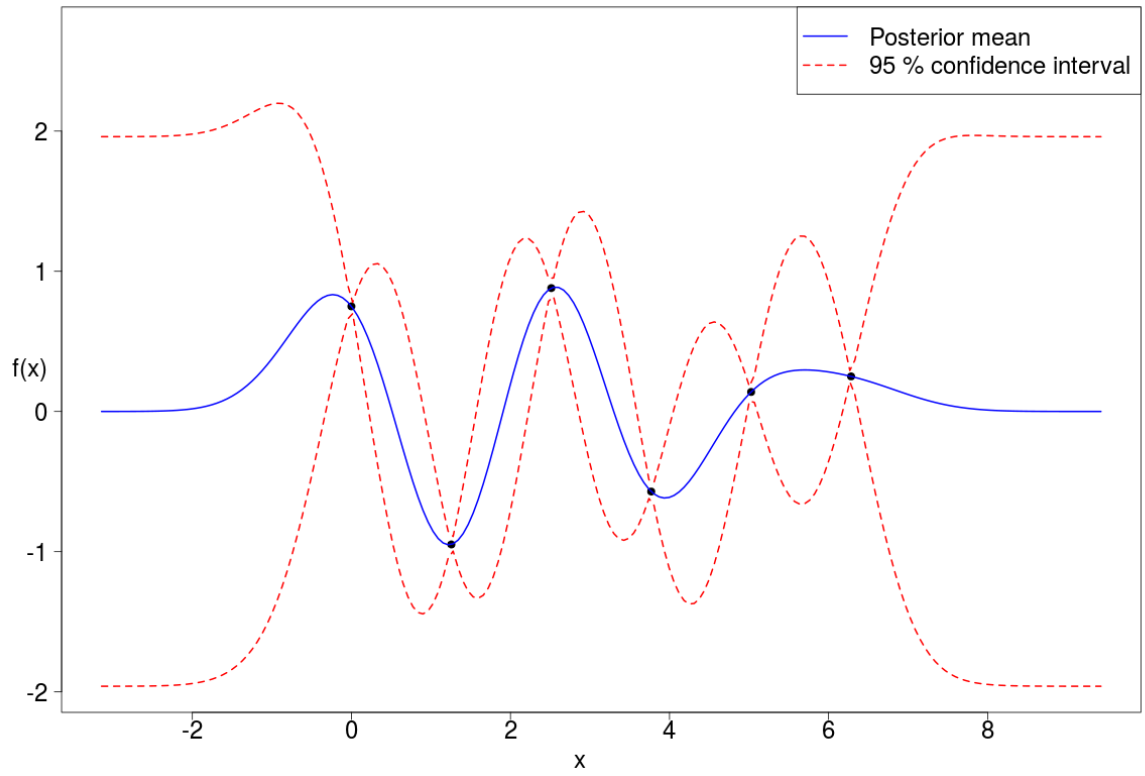


Figure 2.4: Posterior mean and 95% confidence interval for an extended grid. Both the posterior mean and the confidence interval flat out when the distance to the nearest observation increases.

Scale

The covariance function introduced in the previous section had one drawback - the covariance was bounded to be lower than or equal to 1. In practice, we need a way to scale the variability in the process. As a motivation, consider the following modification of the running example. Assume that the true function is scaled by a factor of ten, that is

$$f(x) = 10 \cdot \left(\frac{1}{2}\cos(2x) + \frac{1}{2}\sin(3x) + \frac{1}{4}\cos(2.5x) \right). \tag{2.11}$$

If we now obtain some evaluations from this function and use the formulas for the posterior mean and covariances that are given in (2.9) and plot the result, we get the plot in Figure 2.5. We see that the mean is still interpolating the observed values, but there is something wrong with the uncertainty. The true function lies mostly outside the 95% confidence interval, which it should not. The uncertainty is clearly underestimated. The reason for this is that the covariance function we use is not adequate for this function. The prior covariance matrix does only have entries that are smaller than, or equal to 1. From the equations in (2.9) we see that the posterior covariance is always smaller than the prior covariance. Therefore, since the prior covariance is bounded to be smaller than one, so is the posterior covariance.

The solution is therefore to introduce a way of scaling the covariance matrix. This can be

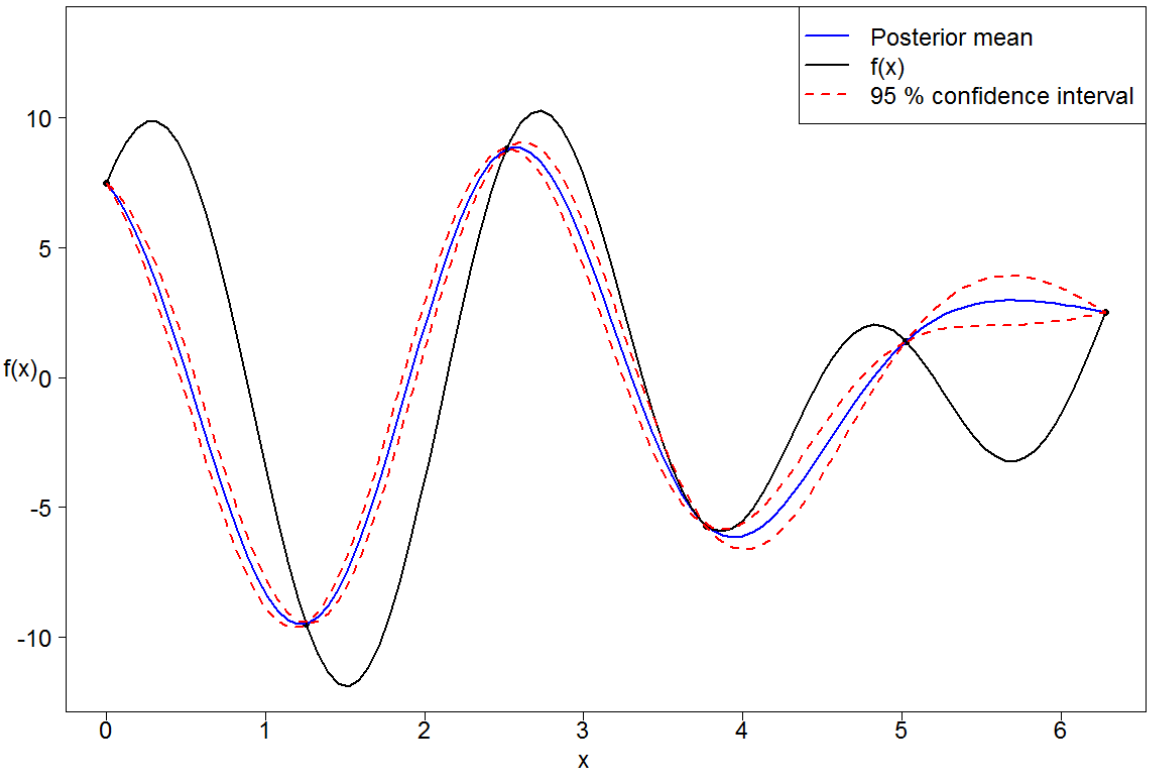


Figure 2.5: Posterior mean and 95% confidence intervals for the fitted GP to the scaled running example.

done by introducing a scale parameter τ^2 and define the covariance matrix $\Sigma_n = \tau^2 C_n$. Here, C_n is a correlation matrix with $C(x, x') < 1$ for $x \neq x'$, $C(x, x) = 1$ and positive definite. The GP can then be written as

$$Y \sim \mathcal{GP}(0, \tau^2 C_n). \tag{2.12}$$

Now we have an unknown parameter that we must estimate. This can be done in several ways, for instance maximum likelihood, Bayesian inference, e.g MCMC or manual tuning. In this thesis, we will focus on maximum likelihood estimation. See for instance Gramacy and Lee (2008) for an approach where parameters are estimated using Markov Chain Monte Carlo(MCMC). Manual tuning may give reasonable results, but the process of finding good estimates manually could be tedious.

We use that Y is normally distributed and set up the expression for the likelihood,

$$L = L(\tau^2) = \left(2\pi\tau^2\right)^{-\frac{n}{2}} |C_n|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau^2} Y_n^T C_n^{-1} Y_n\right\} \tag{2.13}$$

taking the log yields

$$l(\tau^2) = \log L(\tau^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tau^2) - \frac{1}{2}\log\left(|C_n|\right) - \frac{1}{2\tau^2} Y_n^T C_n^{-1} Y_n \tag{2.14}$$

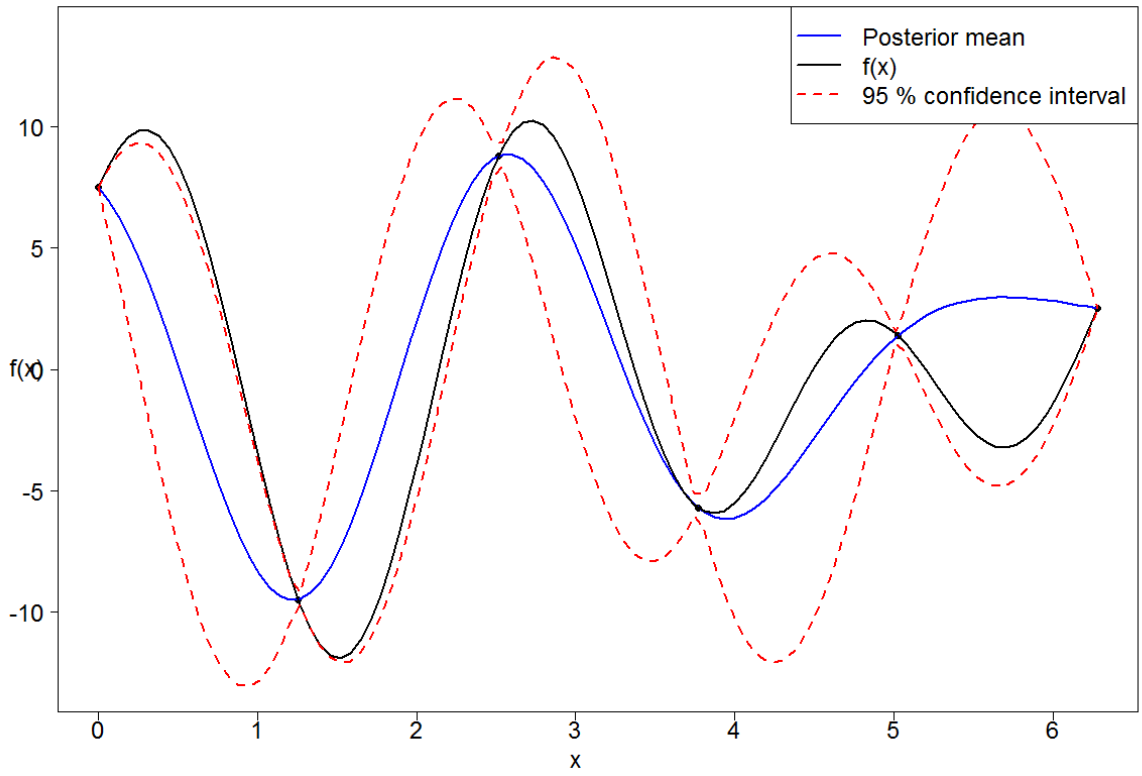


Figure 2.6: Posterior mean and 95% confidence intervals for the fitted GP to the scaled running example.

In order to maximize the likelihood, we differentiate with respect to τ^2, equate to zero and solve,

$$\begin{aligned} \frac{\partial l}{\partial \tau^2} = l' &= -\frac{n}{2\tau^2} + \frac{1}{2(\tau^2)^2} Y_n^T C_n^{-1} Y_n = 0 \\ \hat{\tau}^2 &= \frac{Y_n^T C_n^{-1} Y_n}{n}. \end{aligned} \tag{2.15}$$

We can now plug the estimated scale parameter into the posterior equations in (2.9).

$$\begin{aligned} \boldsymbol{\mu}(\Omega) &= C(\Omega, X_n) C(X_n, X_n)^{-1} Y_n \\ \Sigma(\Omega) &= \hat{\tau}^2 \left(C(\Omega, \Omega) - C(\Omega, X_n) C(X_n, X_n)^{-1} C(X_n, \Omega) \right). \end{aligned} \tag{2.16}$$

When comparing (2.16) and (2.9), we see that the posterior mean do not change, as $\hat{\tau}^2$ cancels out by C and C^{-1}. In the covariance however, $\hat{\tau}^2$ is just multiplied by the covariance in (2.9). An important note here is that by using $\hat{\tau}^2$, we are turning $Y(\Omega) \,|\, \mathcal{D}_n$ from a multivariate normal into a multivariate student-t distribution with n degrees of freedom Gramacy (2017). However, with a sufficiently large n, it will be approximately multivariate normal, so we will assume that for now.

If we return to our adjusted running example in (2.11), we can see how the use of a scale parameter changes the model. In Figure 2.5 we saw that the posterior covariance was underestimated and something was clearly not right. In Figure 2.6 we see that with the scale parameter,

things look much better. The estimate of the scale parameter is in this case $\widehat{\tau^2} = 65.96$. The shape of the confidence intervals looks much like what we can see in Figure 2.3, and the true function lies, as expected, mostly within this interval.

Noise

We can see, in both Figure 2.3 and Figure 2.6, that the estimated posterior mean interpolates the observed points. It is not always desirable that the fitted GP goes through the data points. One reason for that is that the observations are, in many cases, noisy. If we force our GP to interpolate all points, it could easily become non-smooth. For example, assume we have observations from an unknown function at two identical positions in the input space. If these observations have a different value, then it makes no sense to force the GP to interpolate both these points since the resulting GP would become non-smooth. Instead, we should allow the GP to go somewhere in between these two points so that our fitted GP remains smooth. In addition, there could be problems with inverting the covariance matrix since the matrix will no longer have full rank. The solution is to introduce a new parameter which is added to the diagonal in the covariance matrix. We write this as

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = C(\boldsymbol{x}_i, \boldsymbol{x}_j) + g\delta_{i,j}, \tag{2.17}$$

where $g > 0$ is the nugget parameter and $\delta_{i,j}$ is the Kronecker delta function, defined by

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.18}$$

Note that the nugget is only added when the indices of $\boldsymbol{x}$ are the same, not necessarily for equal $\boldsymbol{x}$. The temporally change of notation is to emphasize that. The GP can then be written as

$$Y \sim \mathcal{GP}(0, \tau^2 K_n), \tag{2.19}$$

where the entries in the covariance matrix are given by (2.17) multiplied by τ^2. In matrix form we have

$$\tau^2 K_n = \tau^2 (C_n + g\mathbb{I}_n), \tag{2.20}$$

where $\mathbb{I}_n$ is the $n \times n$ identity matrix. The next step is then to estimate g, which we also will do by maximum likelihood. Recall the expression for the MLE for τ^2 in (2.15). The MLE for τ^2 given g is then

$$\hat{\tau}^2 = \frac{Y_n^T K_n^{-1} Y_n}{n} = \frac{Y_n^T (C_n + g\mathbb{I}_n)^{-1} Y_n}{n}. \tag{2.21}$$

By inserting this expression for τ^2 in (2.14) we get the profile likelihood with only one unknown parameter, g,

$$\begin{aligned} l(g) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\tau}^2) - \frac{1}{2}\log(K_n) - \frac{1}{\hat{\tau}^2}Y_n^T K_n^{-1} Y_n \\ &= -\frac{n}{2}\left(\log(2\pi) + 1 - \log(n)\right) - \frac{n}{2}\log(Y_n^T K_n^{-1} Y_n) - \frac{1}{2}\log(|K_n|) \\ &= c - \frac{n}{2}\log(Y_n^T (C_n + g\mathbb{I}_n)^{-1} Y_n) - \frac{1}{2}\log(|C_n + g\mathbb{I}_n|). \end{aligned} \tag{2.22}$$

Maximizing $l(g)$ requires numerical methods since there are no closed form solution of $l'(g) = 0$. Fortunately, optimizing $l(g)$ can be done quite efficiently with gradient based methods since we can compute the gradient of $l(g)$. In order to obtain an expression for the gradient, the following formulas are useful (Petersen et al., 2008)

$$\begin{aligned} \frac{\partial K_n^{-1}}{\partial \phi} &= -K_n^{-1}\frac{\partial K_n}{\partial \phi}K_n^{-1} \\ \frac{\partial \log|K_n|}{\partial \phi} &= \text{tr}\left\{K_n^{-1}\frac{\partial K_n}{\partial \phi}\right\}. \end{aligned} \tag{2.23}$$

The differentiation is done with respect to a parameter ϕ that are involved in the entries of K_n. tr denotes the trace of a matrix, that is the sum of the diagonal entries. The derivative of a matrix K_n with respect to a scalar ϕ equals the matrix where each element is differentiated with respect to ϕ. Applying the formulas from (2.23) to the profile likelihood in (2.22) we get

$$\begin{aligned} l'(g) &= -\frac{n}{2}\frac{1}{Y_n^T K_n^{-1} Y_n}\frac{\partial}{\partial g}\left(Y_n^T K_n^{-1} Y_n\right) - \frac{1}{2}\frac{\partial \log|K_n|}{\partial g} \\ &= -\frac{n}{2}\frac{Y_n^T \frac{\partial K_n^{-1}}{\partial g} Y_n}{Y_n^T K_n^{-1} Y_n} - \frac{1}{2}\frac{\partial \log|K_n|}{\partial g} \\ &= \frac{n}{2}\frac{Y_n^T K_n^{-1}\frac{\partial K_n}{\partial g}K_n^{-1} Y_n}{Y_n^T K_n^{-1} Y_n} - \frac{1}{2}\text{tr}\left\{K_n^{-1}\frac{\partial K_n}{\partial g}\right\}. \end{aligned} \tag{2.24}$$

From (2.17) we see that g is only involved in the diagonal elements of K_n, the entries are simply $1 + g$. The off diagonal elements of k_n are constants in terms of g. Hence, $\frac{\partial K_n}{\partial g}$ is simply the $n \times n$ identity matrix. The final expression of $l'(g)$ is then

$$l'(g) = \frac{n}{2}\frac{Y_n^T \left(K_n^{-1}\right)^2 Y_n}{Y_n^T K_n^{-1} Y_n} - \frac{1}{2}\text{tr}\left\{K_n^{-1}\right\}. \tag{2.25}$$

See for instance Snyman (2005) for efficient gradient based algorithms for optimization. The MLE estimate of the nugget parameter, $\hat{g}$, can now be inserted in the expression for $\hat{\tau}^2$ which

yields the posterior equations

$$\begin{aligned}\boldsymbol{\mu}(\Omega) &= \hat{K}(\Omega, X_n)\hat{K}(X_n, X_n)^{-1}Y_n \\ \Sigma(\Omega) &= \hat{\tau}^2\left(\hat{K}(\Omega,\Omega) - \hat{K}(\Omega, X_n)\hat{K}(X_n, X_n)^{-1}\hat{K}(X_n, \Omega)\right),\end{aligned} \tag{2.26}$$

where the notation $\hat{K}$ is used to emphasize that the covariance matrix K is a function of an estimated quantity, $\hat{g}$.

To see how the nugget can be used in practice, we return to our running example. Again, we choose $n = 6$ points to evaluate the function, but now we make two observations at each point so we get in total 12 observations. In addition, we add some noise to the observations so that our observed value, y_i is the true function plus some normally distributed noise,

$$y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, 0.2^2) \quad \text{for } i = 1, .., 12. \tag{2.27}$$

By using the posterior equations in (2.26) we can compute the posterior mean and covariance, and plot the result. This can be seen in Figure 2.7. We observe a few differences from before. The posterior mean do no longer interpolate the observations, but go somewhere in between as we would expect. The confidence intervals looks a bit more irregular, and now is the uncertainty no longer zero at the points where we have observations. The true function is still mostly within the 95% confidence interval.

Length Scale

The third hyperparameter we will consider is a parameter related to the covariance function. The covariance function in (2.4) has one major drawback - it does not take into account that variables might have different spatial dependency. It is reasonable to learn this dependency through a parameter. A generalization of (2.4) that take this into account is the isotropic Gaussian, given by

$$C_\theta(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\theta}\right\}, \tag{2.28}$$

where θ is called the length scale parameter. Again we will use maximum likelihood to estimate the parameters. The procedure for estimating θ is no different from earlier, except that we have one more variable in the likelihood. The MLE estimate for τ^2 is now

$$\hat{\tau}^2 = \frac{Y_n^T(C_\theta + g\mathbb{I}_n)^{-1}Y_n}{n} = \frac{Y_n^T K_\theta^{-1} Y_n}{n}. \tag{2.29}$$

Inserting this expression into (2.14), we get an almost identical expression for the likelihood as in (2.22), but now the log likelihood is a function of two variables,

$$l(g, \theta) = c - \frac{n}{2}\log(Y_n^T K_\theta^{-1} Y_n) - \frac{1}{2}\log|K_\theta|. \tag{2.30}$$

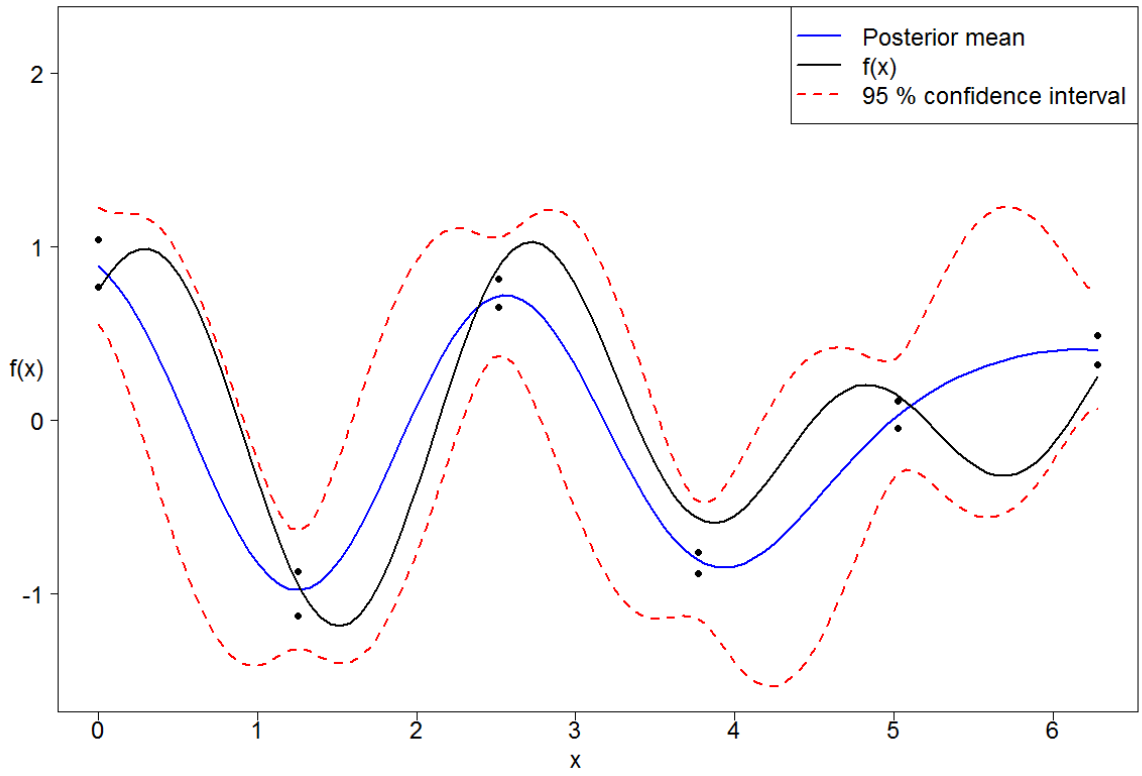


Figure 2.7: Posterior mean and 95% confidence intervals for a GP with noisy observations.

Also here, we need to use numerical methods to maximize the likelihood. The gradient of $l(g, \theta)$ is not hard to derive, and thus may gradient based method be used. The derivative with respect to g is given in (2.24), and the derivative with respect to θ is derived in the following. The diagonal entries of K_θ is still $1 + g$ since $C_\theta(x, x) = 1$. The derivative of the diagonal entries with respect to θ is therefore zero. The derivative of the off diagonal entries is given by

$$\begin{aligned} \frac{\partial K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left(\exp\left\{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\theta}\right\}\right) \\ &= K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\theta^2}, \qquad i \neq j. \end{aligned} \tag{2.31}$$

Therefore, we have that

$$\dot{K}_\theta = \frac{\partial K_\theta}{\partial \theta} = \frac{K_\theta \cdot D_n}{\theta^2}, \tag{2.32}$$

where D is a $n \times n$ matrix with Euclidean distances. The dot$(\cdot)$ in (2.32) denotes a component-wise product, not ordinary matrix multiplication. Since the diagonal entries of D_n is zero, the derivative of K_θ will also have zeros on the diagonal. For the off diagonal elements, we get the

same as in (2.31). The derivative of $l(g, \theta)$ with respect to θ is then

$$\begin{aligned}\frac{\partial l(g,\theta)}{\partial \theta} &= -\frac{n}{2}\frac{1}{Y_n^T K_\theta^{-1} Y_n}\frac{\partial}{\partial \theta}\left(Y_n^T K_\theta^{-1} Y_n\right) - \frac{1}{2}\frac{\partial}{\partial \theta}\log|K_\theta| \\ &= \frac{n}{2}\frac{Y_n^T K_\theta^{-1}\dot{K}_\theta K_\theta^{-1} Y_n}{Y_n^T K_\theta^{-1} Y_n} - \frac{1}{2}\text{tr}\left\{K_\theta^{-1}\dot{K}_\theta\right\}.\end{aligned} \tag{2.33}$$

Again, gradient based methods may be used to maximize the log likelihood in order to obtain the maximum likelihood estimates of the hyperparameters.

An additional improvement of the covariance structure, which is a further generalization of the isotropic Gaussian, is the separable (anisotropic) Gaussian covariance function,

$$\begin{aligned}C_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') &= \exp\left\{-\sum_{k=1}^{p}\frac{(x_k - x_k')^2}{\theta_k}\right\} \\ &= \exp\left\{-\frac{(x_1 - x_1')^2}{\theta_1}\right\} \cdot \ldots \cdot \exp\left\{-\frac{(x_p - x_p')^2}{\theta_p}\right\}.\end{aligned} \tag{2.34}$$

Thus, we allow the covariance decay at different rates for different input variables. The estimation of $\theta_1, ..., \theta_p$ is done in a similar way as before, but now the profile log likelihood has $p + 1$ variables. The expression for the log likelihood is almost identical as before, we just replace K_θ with $K_{\boldsymbol{\theta}}$. As before, the diagonal elements of $K_{\boldsymbol{\theta}}$ does not involve $\boldsymbol{\theta}$, so the derivative with respect to θ_k is still zero. The derivative of the off diagonal elements in $K_{\boldsymbol{\theta}}$ with respect to θ_k is given by

$$\frac{\partial K_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \theta_k} = K_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\frac{(x_{ik} - x_{jk})^2}{\theta_k^2}, \qquad i \neq j. \tag{2.35}$$

and in matrix form,

$$\dot{K}_{\theta_k} = \frac{\partial K_{\boldsymbol{\theta}}}{\partial \theta_k} = \frac{K_{\boldsymbol{\theta}} \cdot D_k}{\theta_k^2}, \tag{2.36}$$

where D_k is a $n \times n$ matrix with square distances between the $k - th$ elements of the input variables. The derivative of the profile log likelihood with respect to θ_k is then

$$\frac{\partial l(g,\boldsymbol{\theta})}{\partial \theta_k} = \frac{n}{2}\frac{Y_n^T K_{\boldsymbol{\theta}}^{-1}\dot{K}_{\theta_k} K_{\boldsymbol{\theta}}^{-1} Y_n}{Y_n^T K_{\boldsymbol{\theta}}^{-1} Y_n} - \frac{1}{2}\text{tr}\left\{K_{\boldsymbol{\theta}}^{-1}\dot{K}_{\theta_k}\right\}. \tag{2.37}$$

The hyperparameters can now be numerically optimized with a gradient based algorithm. The expressions for the posterior mean and covariance is almost identical to what have in (2.26), but now $\widehat{K}$ is now a function of $\widehat{g}$ and $\widehat{\boldsymbol{\theta}}$.

Note that the inverse of the covariance matrix, $K_{\boldsymbol{\theta}}$, must be computed in order to compute the gradient. This means that in each iteration of the optimization algorithm, a $n \times n$ matrix must be inverted. When optimizing the log likelihood with respect to many variables, many iterations are typically required in order to obtain a global maximum. If n is large, this will be very time-consuming.

2.2.2 More Advanced Gaussian Processes

Covariance Structure

The covariance structures introduced in the previous section is only a small selection of what one may use when fitting a GP. The choice of covariance structure depends on what kind of function or physical process the model is meant for. The function in our running example is very smooth and infinitely differentiable, so the isotropic Gaussian is an adequate choice in that case. If one wants to model something less smooth, for instance a stock price, another covariance function should be used to capture the noisy behavior. One class of covariance functions that allows less smooth processes is the *Matern class* of covariance functions given by

$$k(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}d}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}d}{l} \right), \tag{2.38}$$

where K_ν is the modified Bessel function (Abramowitz and Stegun, 1964), d is the Euclidean distance, l is the length scale parameter discussed in section 2.2.1 and ν is a positive, real valued parameter. It can be shown that if $\nu \to \infty$, the covariance function simplifies to the squared exponential defined in (2.4) (Rasmussen, 2004).

In Figure 2.8 are $k(d)$ plotted for three different values of ν. We see that $k(d)$ are decreasing monotonically towards zero for all values of ν, but in different ways. For $\nu = 0.5$, points that are close to each other will be less correlated, which will give a less smooth GP compared to GPs with $\nu = 2$ and $\nu \to \infty$.

In Figure 2.9, three samples from a GP with zero mean and Matern covariance are plotted. The differences in smoothness are evident. For $\nu = 0.5$, the process is very noisy and non-smooth and resembles a somewhat volatile stock price. For $\nu \to \infty$ the process is very smooth, just as the GP sample in Figure 2.2.

Non-Stationary Gaussian Processes

Until now, we have only considered the case where we fit one GP to the entire input space. The underlying assumption is that we have the same mean and covariance structure over the whole domain, i.e stationary. This assumption might be too strong in some cases, as some modeling problems require more flexibility than what a stationary GP provides. Gramacy and Lee (2008) argue for this and discusses a method for fitting non-stationary GPs by treed partitioning. This is done by partitioning the input space into distinct subsets and fit a stationary GP in each subset. The treed partitioning is done by a reversible jump MCMC and parameter estimation is done by Metropolis-Hastings draws. Gramacy et al. (2007) provides practical examples of treed GPs and a thorough review of the `R` package `tgp`.

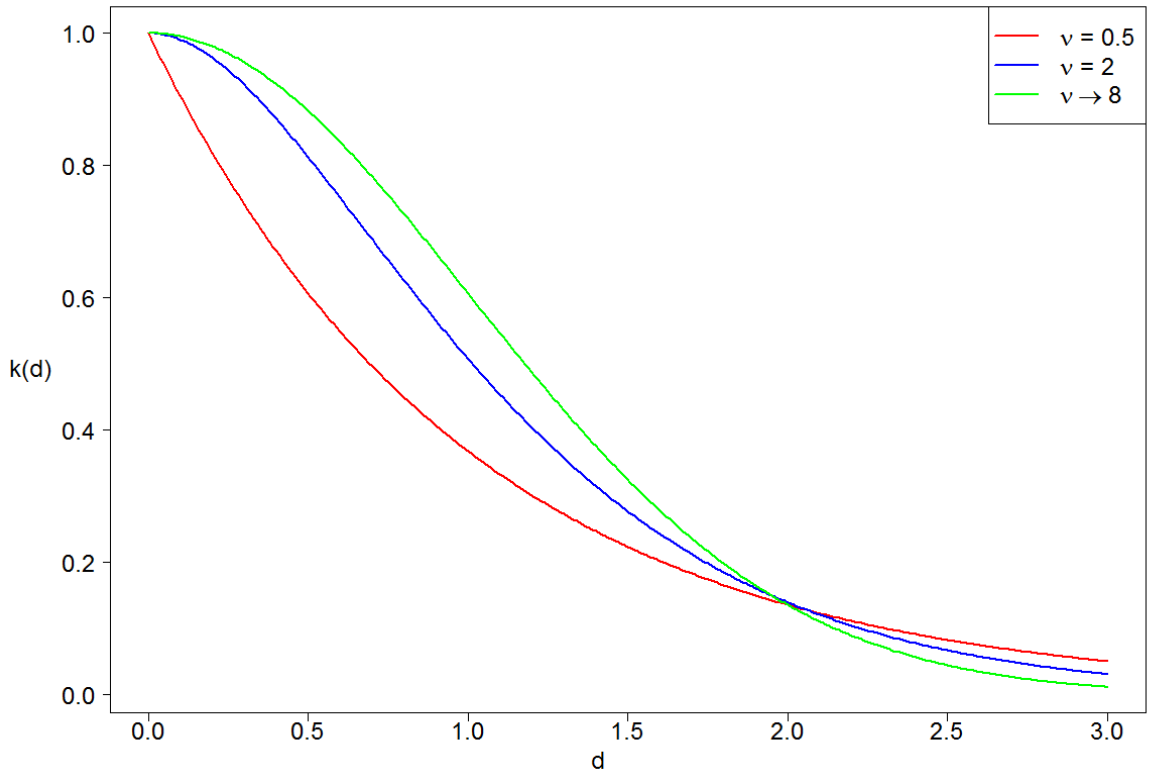


Figure 2.8: Matern covariance function as a function of the distance for three different values of the parameter ν and $l = 1$.

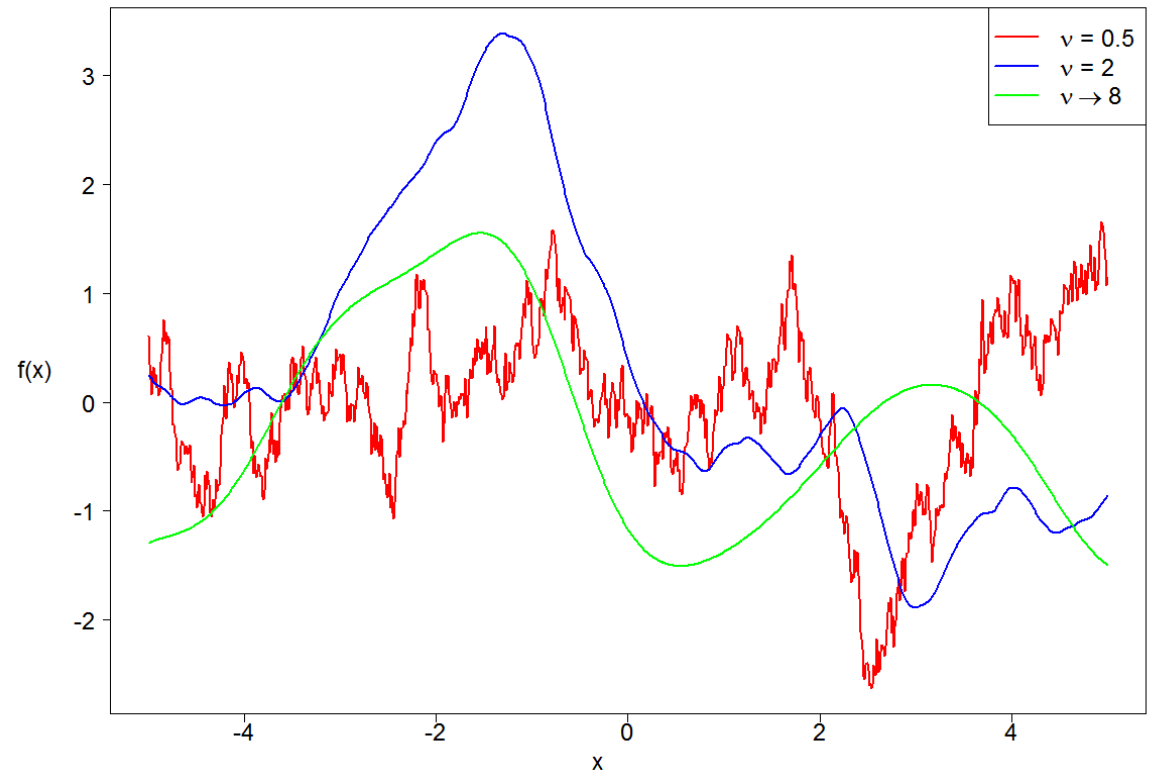


Figure 2.9: Three samples from a GP with Matern covariance structure with $l = 1$.

Categorical variables

We have, thus far, only considered variables that are real-valued. There are two good reasons for that - it is the easiest to handle and it is the most common type of variable. There is one major problem with categorical variables when it comes to the covariance functions that we have discussed here: how do we define a distance between two such variables in space? One obvious solution is to use a covariance function that does not depend on the distance between two points in space. Still, an appropriate choice is hard and depends heavily on the data and underlying model.

Roustant et al. (2018) address covariance functions for categorical data that yields a positive definite covariance matrix based on a nested Bayesian linear model. Gramacy et al. (2010) solve the problem with categorical data by partitioning the data set by these variables and exclude them in the GP regression. Platt et al. (2002) uses a binary covariance function in the context of using a GP to generate music playlists. We will not go further in depth on this topic, and let the interested reader consult with the proposed literature.

Handling Large Data Sets

We briefly mentioned in the introduction that one weakness of GPs is the problems when the number of data points and the number of variables is large. These problems arise in the estimation of the hyperparameters. Assuming that we have p variables and are including a scale parameter, nugget and use the separable Gaussian covariance matrix, we will have $p+2$ parameters to estimate. The profile log likelihood in (2.30) that need to be maximized with numerical methods will now have $p+1$ variables. In each iteration of the optimization algorithm, a $n \times n$ matrix must be inverted, where n is the number of observations. The time complexity of inverting a matrix is $\mathcal{O}(n^3)$, so a large n gives high computation time.

Different solutions to the scaling problem of GPs are proposed in the literature. Liu et al. (2014) discuss how a GP surrogate model with many variables may be assisted by evolutionary algorithms to reduce the dimension of the input space to achieve lower computation times. Smola and Bartlett (2001) suggest a sparse greedy technique to approximate the posterior mean and covariance, with a time complexity of $\mathcal{O}(np^2)$, where n is the sample size and p is the number of variables. Vecchia approximations is an approximation of a GP that leads to a sparse Cholesky factor of the inverse covariance matrix. A more thorough discussion of Vecchia approximations can be found in Katzfuss and Guinness (2017). Bayesian committee machine is a method where the observed data is divided into m distinct subsets and a model is fitted in each subset. When making predictions, one combines the posterior mean and covariance from all the m models to make an approximation of the posterior mean and covariance. See for instance Xiuliang et al. (2009) for an application of Bayesian committee machine to GPs.

2.3 Acquisition Functions

Bayesian optimization is a sequential optimization procedure where we in each step must decide where to evaluate the unknown function next. We can either choose one point in each iteration or a batch of B points. We will focus on the case with only one point, and then briefly discuss how B points can be selected jointly.

An acquisition function, $\alpha(\boldsymbol{x})$, is a cheap way of finding where we should evaluate our unknown function f next. When we are in the context of maximization, we choose new points for evaluation by solving

$$\boldsymbol{x}^{\text{new}} = \underset{\boldsymbol{x}\in\Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x}). \tag{2.39}$$

When designing an acquisition function one must keep in mind that

1. f is typically an expensive function to evaluate so we want to minimize the number of evaluations.

2. One can easily be stuck in local optimums, so a mechanism for exploring globally is desirable.

The acquisition function we choose should, therefore, find a balance between exploring and exploiting the unknown objective function. The acquisition functions we will consider here may be interpreted as evaluating the expected utility associated with evaluating f at a new point $\boldsymbol{x}$. We will derive and discuss probability of improvement (PI) and expected improvement (EI). Both will be illustrated by applying them to our running example.

2.3.1 Probability of Improvement

Let y^* denote the optimal value thus far, that is

$$y^* = \max Y_n, \tag{2.40}$$

where Y_n is the vector of evaluated points from the unknown function f. Probability of improvement (PI) evaluates f at the point which is most likely to improve on y^*. Let $Y = Y(\boldsymbol{x})$ be a random variable that denotes the value of f at $\boldsymbol{x}$. Recall that Y is a normally distributed variable, since it is a finite subset of the GP, with mean μ and variance σ^2. Next, define the utility function associated with evaluating f on a point $\boldsymbol{x}$ by

$$u(Y) = \begin{cases} 0 & \text{if } y \leq y^* \\ 1 & \text{if } y > y^*. \end{cases} \tag{2.41}$$

If the a new evaluating turns out to be better than our current maximum, we receive a unit reward. Otherwise, no reward are received. Note that the reward is independent of the size of

y. The acquisition function is the expected value of this utility function with respect to Y,

$$\alpha_{\mathrm{PI}}(\boldsymbol{x}) = \mathrm{PI} = \mathrm{E}\left[u(Y)\right] = \int_{-\infty}^{y^*} 0 \cdot p(y)dy + \int_{y^*}^{\infty} 1 \cdot p(y)dy. \tag{2.42}$$

The notation $a_{\mathrm{PI}}(\boldsymbol{x})$ is used to emphasize that PI is a function of the input variable $\boldsymbol{x}$. In the following we will only use the notation PI. The first term in (2.42) is zero, and the second term is an integral over the probability density function $p(y)$,

$$\mathrm{PI} = \int_{y^*}^{\infty} p(y)dy = P(Y > y^*) = 1 - \Phi\left(\frac{y^* - \mu}{\sigma}\right) = 1 - \Phi\left(z\right), \tag{2.43}$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution and $z = \frac{y^* - \mu}{\sigma}$. We see from the expression that a low z yields a high PI. Thus, points where μ is high and σ is low will give a high probability of improvement.

Assume that we want to maximize the function $f(x)$ in our running example and the next point for evaluation shall be chosen based on PI. We have $n = 6$ initial samples, and have computed the posterior mean and covariance using the posterior equations in (2.9). Thus, for each $x_i \in \Omega$ we have a posterior mean, μ_i and posterior standard deviation $\sigma_i = \sqrt{\Sigma(x_i, x_i)}$. By inserting these values into the expression in (2.43), we obtain the PI for all input values, x_i.

In Figure 2.10 the PI for our running example is plotted. We observe that there are three areas with high probability of improvement, which corresponds well with the interpretation and Figure 2.3. For x close to 0 and 6, we see in Figure 2.3 that the mean is high, but so is the uncertainty. The probability of improvement should therefore be high. The other peak in PI is for $x \approx 3$, and this is where the probability of improvement is highest. We can see in Figure 2.3 that the mean is high here, and that the uncertainty is at least smaller than for x close to 0 or 6.

Even though the resulting probability of improvement in our running example seems reasonable, there is something odd about the utility function. The utility of improving on the current maximum does not depend on the actual improvement. This might cause that optimization algorithm gets stuck in a local maximum. A more reasonable utility function takes the improvement into account, and that is what the expected improvement acquisition function does.

2.3.2 Expected Improvement

Again, let y^* denote the optimal value thus far. Expected improvement (EI) considers the following utility function,

$$u(Y) = \max\left\{0, Y - y^*\right\}, \tag{2.44}$$

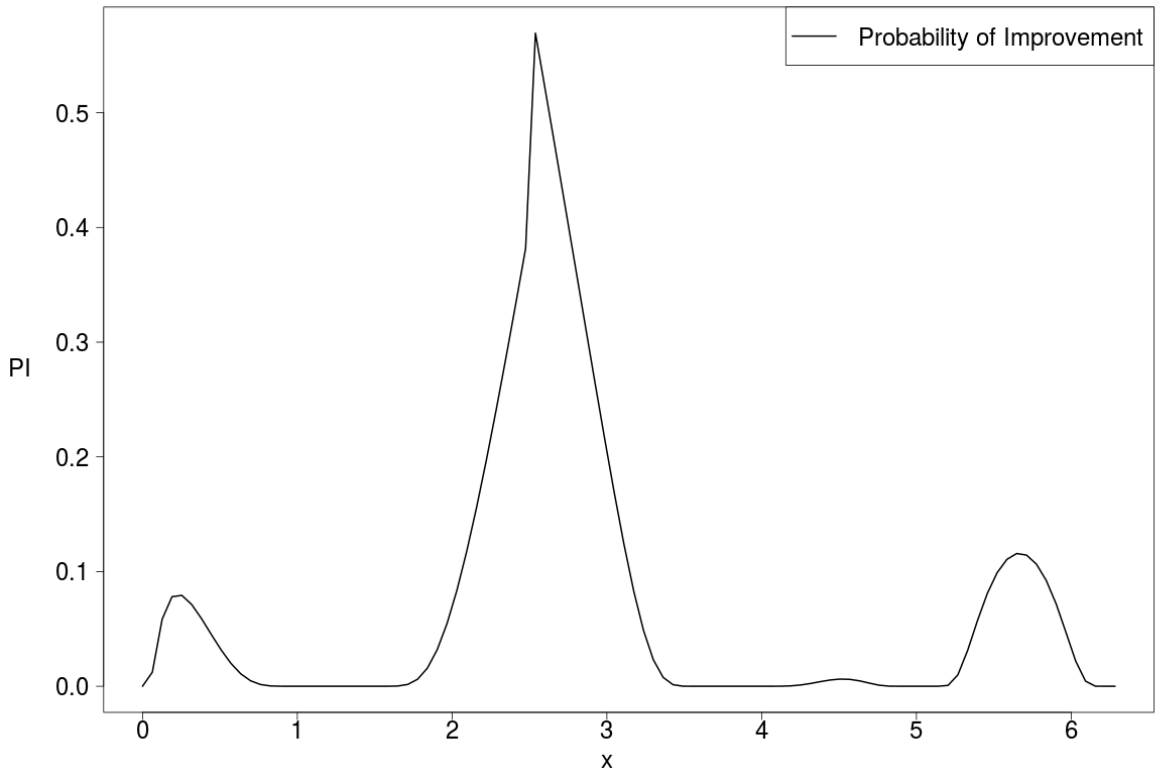


Figure 2.10: Probability of improvement for running example.

meaning that an improvement in the function value gives a reward of this improvement. The EI is defined as the expected value of the utility function,

$$\begin{aligned} \alpha_{\mathrm{EI}}(\boldsymbol{x}) = \mathrm{EI} = \mathrm{E}[u(Y)] &= \int_{-\infty}^{\infty} \max\left\{0, y - y^*\right\} p(y)dy \\ &= \int_{-\infty}^{y^*} 0 \cdot p(y)dy + \int_{y^*}^{\infty} (y - y^*)p(y)dy. \end{aligned} \tag{2.45}$$

The first term is obviously zero. By splitting the second term, EI can be written as

$$\mathrm{EI} = \int_{y^*}^{\infty} y \cdot p(y)dy - y^* \int_{y^*}^{\infty} p(y)dy. \tag{2.46}$$

The second term is a constant, y^*, multiplied by PI given by (2.43). This inserted in (2.46) yields

$$\mathrm{EI} = \int_{y^*}^{\infty} y \cdot p(y)dy - y^* \left(1 - \Phi\left(\frac{y^* - \mu}{\sigma}\right)\right). \tag{2.47}$$

For computing the first term in (2.47), we will use the fact that Y is normally distributed with mean μ and variance σ^2 so that

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \tag{2.48}$$

We will also use a change of variable $Z = \frac{Y-\mu}{\sigma}$. Inserting this in the first term in (2.47) yields

$$\begin{aligned}
\int_{y^*}^{\infty} y \cdot p(y)dy &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{y^*}^{\infty} y \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} dy \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{z}^{\infty} (\sigma z + \mu) \exp\left\{-\frac{z^2}{2}\right\} \sigma dz \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{z}^{\infty} z \exp\left\{-\frac{z^2}{2}\right\} dz + \frac{\mu}{\sqrt{2\pi}} \int_{z}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz,
\end{aligned} \tag{2.49}$$

where $z = \frac{y^*-\mu}{\sigma}$. The integral in the first term can easily be computed by using the substitution $u = \frac{z^2}{2}$. We recognize the second term as the PI for a standard normal distributed variable, multiplied by μ. Hence,

$$\int_{y^*}^{\infty} y \cdot p(y)dy = \sigma \cdot \phi(z) + \mu \cdot (1 - \Phi(z)). \tag{2.50}$$

Combining the results from (2.47) and (2.50) yields

$$\begin{aligned}
\text{EI} &= \sigma \cdot \phi(z) + \mu \cdot (1 - \Phi(z)) - y^* \left(1 - \Phi\left(z\right)\right) \\
&= \sigma \cdot \phi(z) + (\mu - y^*) \cdot (1 - \Phi(z)).
\end{aligned} \tag{2.51}$$

The expression for expected improvement in (2.51) has two components. The first term involves the predictive standard deviation, σ. Higher values of σ gives a higher expected improvement, which encourages exploration. The second term has a factor $(\mu - y^*)$, which increases with μ and then yield a higher EI. Thus, both areas with high mean values and areas with high uncertainty are subjects for further exploration.

Returning to our running example, we can also compute the EI in a similar way we did for PI. For each $x_i \in \Omega$ we can insert μ_i and σ_i into (2.51) and plot EI against the input values. This can be seen in Figure 2.11. We observe that our interpretation of EI is clearly illustrated here. From Figure 2.3 we see that the posterior mean is highest around $x = 3$, where the uncertainty is also low. For points nearby, the mean is a bit lower, but the uncertainty is higher which gives a larger EI. There are also a significant EI for x close to 0 and 6, just as for PI.

When choosing the next point for evaluating, the point with the highest value of the acqui-

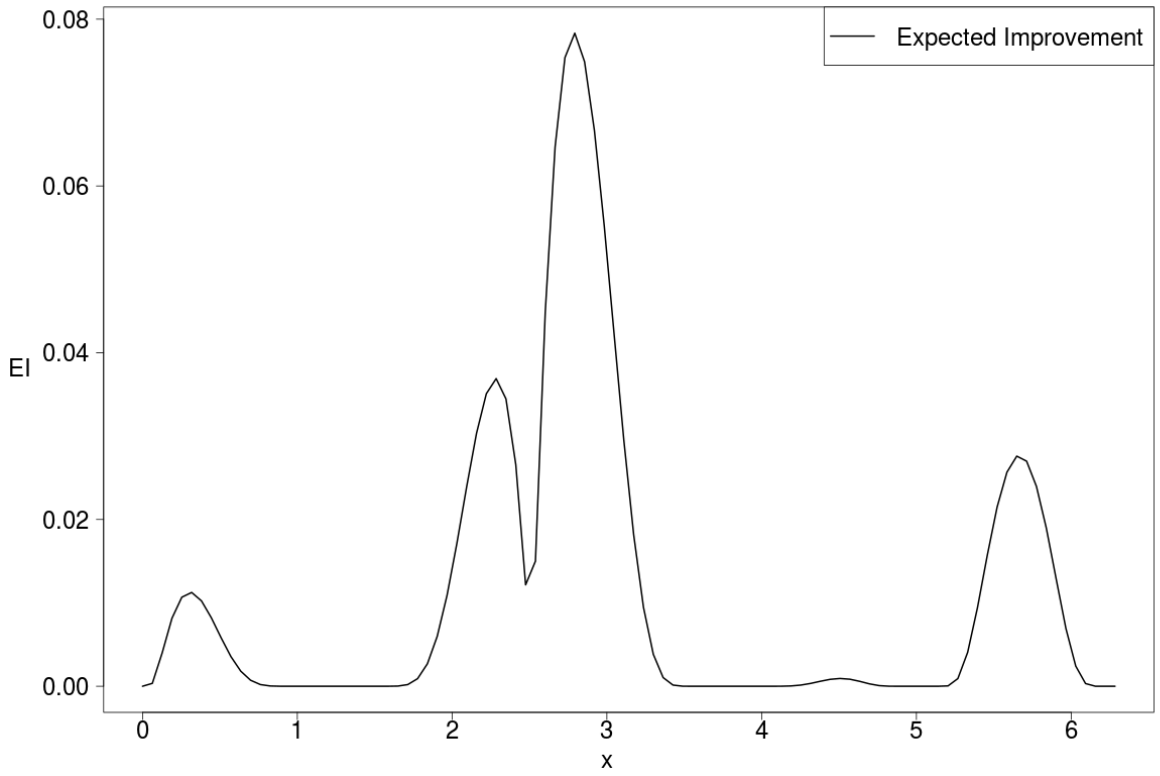


Figure 2.11: Expected improvement for running example.

sition function, $a(\boldsymbol{x})$, is chosen. If the function evaluation is done in batches of size B, what points should then be chosen? The B points with the largest value of the acquisition function are no good strategy for obvious reasons. The areas where $a(\boldsymbol{x})$ is high, the points nearby are also high. In our running example, if one should run a batch of 3 function evaluations, and choose the 3 largest values, all values would be around $x = 3$ when using EI. This gives us unwanted exploitation. An additional mechanism for avoiding exploitation is therefore needed when function evaluations are done in batches.

One simple and intuitive solution is to force the distance between two potential points for evaluation to be larger than some predefined value, that is

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq \delta. \tag{2.52}$$

The choice of δ depends on the problem under consideration. Then, if B new points shall be

chosen, we can do selection in a sequential way,

$$\begin{aligned}
\boldsymbol{x}^{n+1} &= \underset{\boldsymbol{x}\in\Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x}) \\
\boldsymbol{x}^{n+2} &= \underset{\boldsymbol{x}\in\Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x}) \;\text{ where }\; \|\boldsymbol{x}-\boldsymbol{x}^{n+i}\| \geq \delta \;\text{ for } i=1 \\
&\vdots \\
\boldsymbol{x}^{n+B} &= \underset{\boldsymbol{x}\in\Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x}) \;\text{ where }\; \|\boldsymbol{x}-\boldsymbol{x}^{n+i}\| \geq \delta \;\text{ for } i=1,...,B.
\end{aligned} \tag{2.53}$$

With this scheme, new points will always be in a given distance away from the other points in space, which will give a broader exploration of the input space. This strategy will be used in the case study in Chapter 4.

The proposed scheme in (2.53) for selection of several new points is not necessarily optimal. The choice of points to evaluate can be seen as a joint optimization problem of B variables, given by

$$\begin{aligned}
\left\{\boldsymbol{x}^{n+1},...,\boldsymbol{x}^{n+B}\right\} &= \operatorname{argmax}\, \alpha(\boldsymbol{x}^{n+1},...,\boldsymbol{x}^{n+B}) \\
&\text{where} \|\boldsymbol{x}^i-\boldsymbol{x}^j\| \geq \delta \;\; \forall \boldsymbol{x}^i \neq \boldsymbol{x}^j, \;\; i,j=1,...,B.
\end{aligned} \tag{2.54}$$

We have not considered acquisition functions with several input variables, and will not go further in depth here. An analytically expression for the function may not be available, and numerical optimization of the function might be time consuming. We let the interested reader consult with existing literature on the subject, for instance Schonlau et al. (1998) where the extension of acquisition functions to several variables was defined. Ginsbourger et al. (2008) provides a more in depth study of multi point selection with the expected improvement acquisition function.

2.4 Algorithm for Bayesian Optimization

We have now discussed the two major parts in Bayesian optimization - a surrogate model and acquisition functions. We will now combine these to construct an algorithm for maximizing a function $f(\boldsymbol{x})$. This is summarized in Algorithm 1.

The algorithm starts by specifying the GP that will be used as a surrogate model. This includes specification of mean and covariance structure. The initial sample consist of n_0 points at some specified or random locations in the input space. After initialization continues the search for the global optimizer until the most promising point suggested by the acquisition function falls below a lower limit ϵ. For each iteration is the set of observations, $\mathcal{D}_n$, augmented by new observations. The algorithm returns $\mathcal{D}_n$ so that is may be further analyzed.

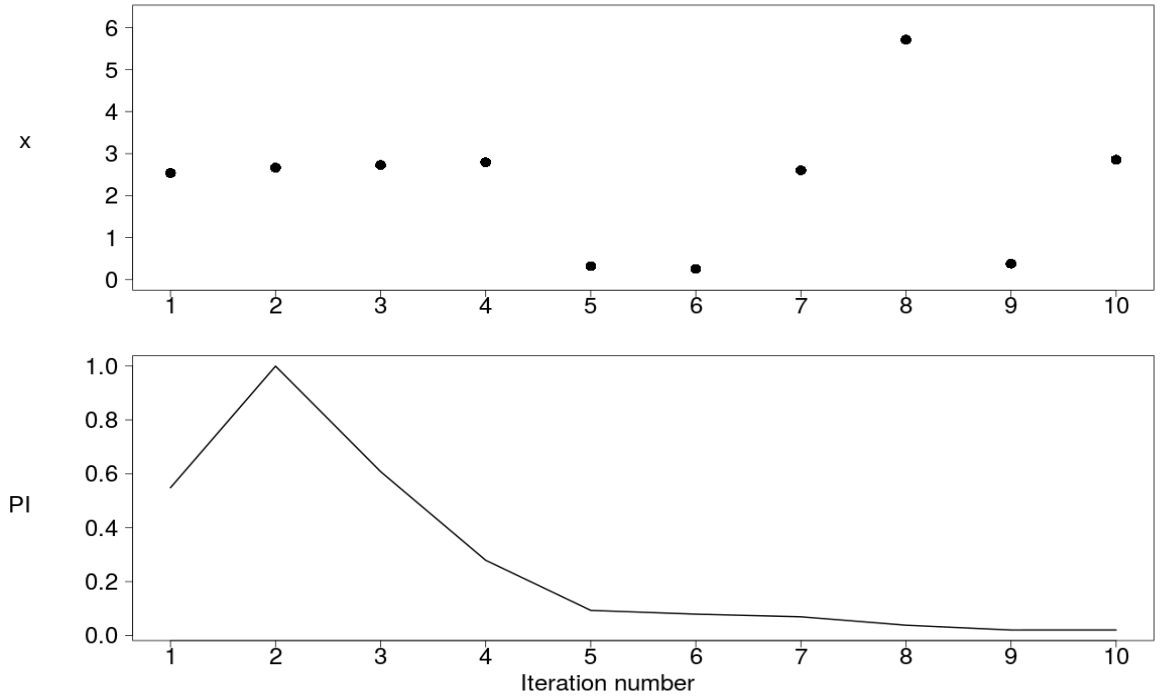
We return to our running example to demonstrate how Bayesian optimization works in practice, in particular how new points are selected based on EI and PI. The selected points in each iteration for EI and PI are visualized in Figure 2.12. Both EI and PI suggests points around

Algorithm 1 Bayesian optimization with GP prior

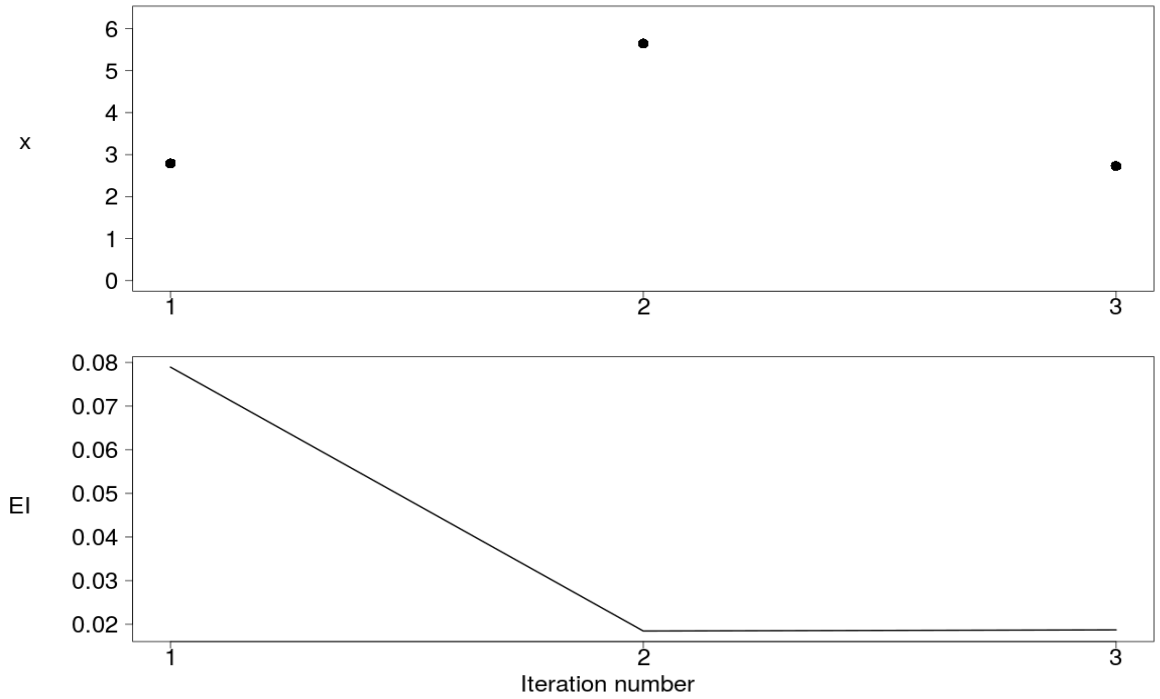
1. Specify prior GP
2. Obtain initial sample, $\mathcal{D}_n = \{\boldsymbol{x}_i, f(\boldsymbol{x}_i)\}_{i=1}^{n_0}$
3. Specify minimum improvement, ϵ
4. **repeat**
5. Fit the surrogate model, $\mathcal{GP} \mid \mathcal{D}_n$
6. Find $\boldsymbol{x}^{\text{new}} = \underset{\boldsymbol{x}\in\Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x})$
7. evaluate $y^{\text{new}} = f(\boldsymbol{x}^{\text{new}})$
8. Augment $\mathcal{D}_n = \{\mathcal{D}_n, (\boldsymbol{x}^{\text{new}}, y^{\text{new}})\}$
9. **until** $\alpha(\boldsymbol{x}^{\text{new}}) < \epsilon$
10. **return** $\mathcal{D}_n$

2.7 in the first iteration, which is where the optimal solution lies. After this, the two acquisition functions suggests different points for evaluation. PI exploits the area where the optimal solution lies, while EI explores a different area. In the third iteration, EI suggests again a point around 2.7, and the algorithm terminates after this because all points have an EI less than $\epsilon_{\text{EI}} = 0.01$. PI suggest to evaluate 10 points in total before all points have a PI less than $\epsilon_{\text{PI}} = 0.01$, and the algorithm terminates.

We note that when we use PI, we explore the input space more than when we use EI. This is because the minimum PI that is needed for evaluating a point, ϵ_{PI}, is less restrictive than the minimum EI, ϵ_{EI}. The values for ϵ_{EI} and ϵ_{PI} are set to arbitrary values here to illustrate how the acquisition functions work. Both acquisition functions suggested points that eventually lead to a solution close to, or equal the optimal solution.



(a)



(b)

Figure 2.12: PI and EI in each iteration of Bayesian optimization of the function in (2.2)

Chapter 3

Value of Information

3.1 Introduction

Information has been and will remain a highly relevant topic in our society. More information than ever is available due to the digital evolution we have seen in the last years. Many mechanical components are now equipped with sensors, which provide useful information about the state of the component. Such information may, for instance, be used to make decisions about the scheduling of maintenance (Jardine et al., 2006). This raises the questions: What is this information worth? Under what circumstances should this information be gathered at all?

We will address these questions in the following, where we will see how value of information (VOI) in a decision situation can be estimated. First, we will introduce decision theory and the associated notation that we will use. We will also use a running example that is given in Section 3.1.2. In Section 3.2 will the concept of VOI be introduced and expressions for prior- and posterior value that are needed to compute VOI will be derived. Methods for estimating VOI are discussed in Section 3.3.

3.1.1 Decision Theory

Decisions of varying importance are made by everyone, every day. Some decisions are trivial and do not require much thought, while other decisions might be more complex and have a meaningful impact on the society. Often will decisions involve uncertainty which makes the process of making the best decision more complicated. Decision analysis is about studying a decision situation and provide guidance to make better decisions.

A decision situation involves two types of variables - decisions and uncertainties. A decision is denoted by $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of alternatives available for the decision maker. A decision situation may involve several decisions, so for generality we will use vector notation $\boldsymbol{a} = (a_1, ..., a_d)$ where $a_i \in \mathcal{A}_i$, $i = 1, ..., d$. Decisions are variables the decision maker may control, but the outcome of a decision remains uncertain if uncertain variables influencing the

outcome are present.

Uncertainties, or random variables, are what the decision maker can not control, and are denoted by $\boldsymbol{\lambda} \in \mathcal{L}$. We will use vector notation to emphasize that there are several uncertain variables involved. Each element in $\boldsymbol{\lambda}$ may be continuous, discrete or categorical.

Once the uncertainties and the possible decisions are identified, the decision maker can consider the different scenarios. A scenario is a set consisting of a possible combination of a decision and outcome of the random variable. Each scenario is associated with a value that represents how happy we are with this particular outcome. We will denote this value by the value function, $v(\boldsymbol{\lambda}, \boldsymbol{a})$. Note that since $\boldsymbol{\lambda}$ is a random variable, $v(\boldsymbol{\lambda}, \boldsymbol{a})$ is also a random variable. We typically want to choose the alternative that maximizes the value function.

A decision situation with a simple value function is introduced next, and is an example that we will use throughout this chapter to illustrate the presented topics.

3.1.2 Running Example

Consider a decision situation with the following simple value function,

$$v(\lambda, a) = \begin{cases} 10 - 3\lambda & \text{if } a = a_1 \\ 8 - \lambda & \text{if } a = a_2, \end{cases} \tag{3.1}$$

where λ is assumed to come from a discretized truncated normal distribution defined on a regular grid between 0 and 3 with step size 0.1. The probability mass function is given by

$$P(\Lambda = \lambda) = P(\lambda - 0.05 \leq \bar{X} \leq \lambda + 0.05), \quad \lambda = 0, 0.1, ..., 3, \tag{3.2}$$

where $\bar{X}$ is distributed according to a truncated normal distribution. The set of actions consists of a_1 and a_2. The value function is visualized in Figure 3.1. We see that for $\lambda < 1$, $a = a_1$ is the alternative that yields the highest value function, while $a = a_2$ is the best alternative for $\lambda > 1$. In the point of intersection, $\lambda = 1$, we are indifferent between choosing a_1 or a_2. If we knew what value λ will take, the decision would be easy. However, if such information where available, it would most likely come with a price. VOI analysis provides a tool for finding a fair price for this information.

3.2 Value of Information

When facing a decision situation, the decision maker is also faced with an auxiliary decision pertained to the underlying decision - information gathering. It might be possible to acquire information about the uncertain variables that are involved in the decision situation so that a more informed decision can be made. For the simple decision situation defined in (3.1), it would be very useful to know the value of λ before a decision is made. We are then certain to

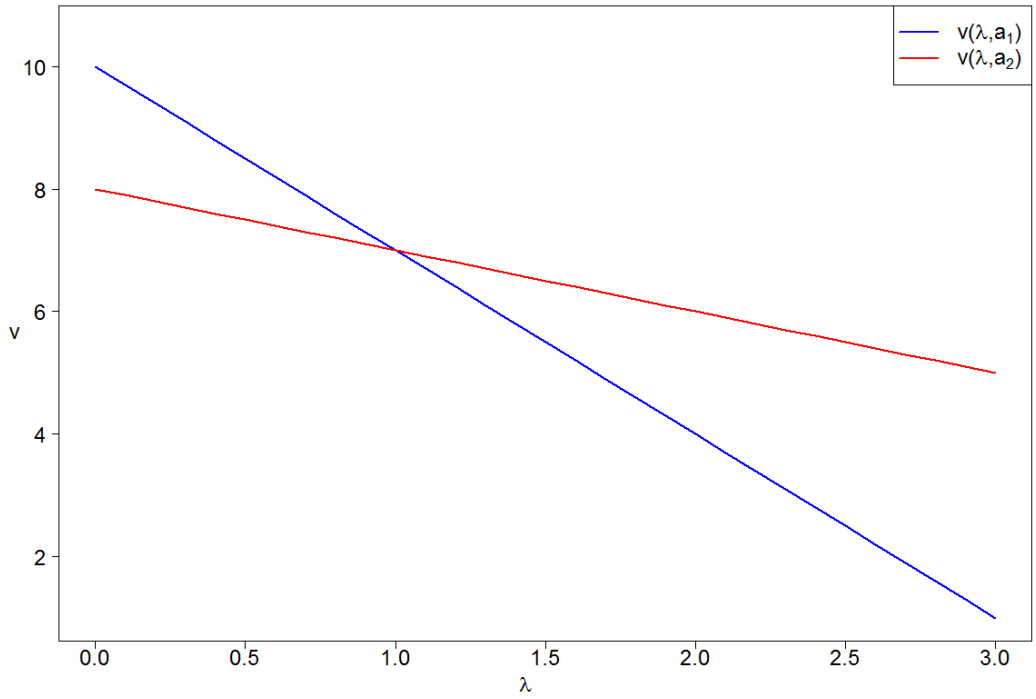


Figure 3.1: Example of a value function with two possible alternatives, a_1 and a_2.

make a better decision that yields a scenario with a higher value attached. The information about λ will come with a price, and it might be so high that the information is not worth gathering.

VOI analysis is useful in many applications. Bratvold et al. (2009) addresses how VOI analysis have been used in the petroleum industry, and highlight decision situations where the use of this analysis was successful. An application of VOI in hydropower production is discussed in Ødegård et al. (2017). The decision under consideration is the scheduling of power production where the inflow of water to the reservoir is uncertain. Measurements of the snow in nearby areas provide information about the inflow, and the value of this information is estimated. Seyr and Muskulus (2016) estimates the value of information for repair times for offshore wind turbines. This can be used to improve the O&M strategy for an offshore wind farm.

Eidsvik et al. (2015) define the value of information (VOI) for an uncertain variable $\boldsymbol{\lambda}$ and decision $\boldsymbol{a}$ as the price at which the decision maker is indifferent between purchasing the information and not. By indifferent, we mean that the decision maker would be indifferent between making the decision with or without the information. The price of indifference is thus the maximum amount that the decision maker is willing to pay for the information. In order to find the VOI, we must quantify the value of making a decision without the information and the value without the information. We will refer to these values as prior value(PV) and posterior value(PoV). We can then express the value of information as

$$\text{VOI} = \text{PoV} - \text{PV}\,. \tag{3.3}$$

Analytic expressions for PV and PoV will be derived in the following. For simplicity, we will derive the expressions for PV and PoV for a risk-neutral decision maker. See for instance Kahneman and Tversky (2013) for decision theory under risk.

3.2.1 Prior Value

The PV is the value associated with making the optimal decision without acquiring any additional information. We will assume that the uncertain variables $\boldsymbol{\lambda}$ has a known prior distribution, $p(\boldsymbol{\lambda})$, that captures the decision makers prior beliefs about $\boldsymbol{\lambda}$. The value that follows from an action $\boldsymbol{a}$ and random variable $\boldsymbol{\lambda}$ is denoted by $v(\boldsymbol{\lambda}, \boldsymbol{a})$. By using the assumption that the decision maker is risk neutral, the optimal decision is the one that maximizes the expected value of $v(\boldsymbol{\lambda}, \boldsymbol{a})$, given by

$$\mathrm{PV} = \max_{\boldsymbol{a}\in\mathcal{A}} \{E(v(\boldsymbol{\lambda}, \boldsymbol{a})\} = \max_{\boldsymbol{a}\in\mathcal{A}} \left\{ \int_{\boldsymbol{\lambda}\in\mathcal{L}} v(\boldsymbol{\lambda}, \boldsymbol{a}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \right\}. \tag{3.4}$$

We return to the running example to illustrate how the prior value can be computed in practice. The expected value of λ is known,

$$\mathrm{E}(\Lambda) = \sum_{\lambda=0}^{3} \lambda \cdot P(\Lambda = \lambda) = 1.5, \tag{3.5}$$

and hence the expected value,

$$\begin{aligned} \mathrm{E}(v(\lambda, a_1)) &= 10 - 3 \cdot \mathrm{E}(\lambda) = 10 - 3 \cdot \frac{3}{2} = 5.5 \\ \mathrm{E}(v(\lambda, a_2)) &= 8 - \mathrm{E}(\lambda) = 8 - \frac{3}{2} = 6.5. \end{aligned} \tag{3.6}$$

The prior value is then

$$\mathrm{PV} = \max_{a} \{\mathrm{E}\left(v(\lambda, a)\right)\} = \max\{5.5, 6.5\} = 6.5. \tag{3.7}$$

Thus, the alternative that gives the highest expected value is a_2, with a value of 6.5. This is the optimal decision when no additional information about λ is available.

3.2.2 Posterior Value

We are now in the setting where we have some additional information that is in some way related to the uncertain quantity $\boldsymbol{\lambda}$. We distinguish between perfect and imperfect information, where the former means that we observe the quantity $\boldsymbol{\lambda}$ itself without any form of noise or measurement error. With imperfect information, we observe a quantity $\boldsymbol{y}$ that are for instance $\boldsymbol{\lambda}$

plus some noise. The probabilistic dependence between $\boldsymbol{\lambda}$ and $\boldsymbol{y}$ is represented through $p(\boldsymbol{\lambda}|\boldsymbol{y})$. The posterior value for both types of information will be derived in the following.

Perfect Information

Perfect information means that we observe the uncertain quantity directly, without any measurement noise or such. Assuming that $\boldsymbol{\lambda}$ has a prior distribution $p(\boldsymbol{\lambda})$ which is known, we start from the expression for PV in (3.4) and condition on $\boldsymbol{\lambda}$ to get

$$\mathrm{PV} \mid \boldsymbol{\lambda} = \max_{\boldsymbol{a}\in\mathcal{A}} \{E(v(\boldsymbol{\lambda}, \boldsymbol{a}) \mid \boldsymbol{\lambda}\} = \max_{\boldsymbol{a}\in\mathcal{A}} \{v(\boldsymbol{\lambda}, \boldsymbol{a})\}. \tag{3.8}$$

The expectation may be removed since we are conditioning on all the random variables present, so that $v(\boldsymbol{\lambda}, \boldsymbol{a})$ is no longer stochastic. Since we do not know what $\boldsymbol{\lambda}$ are before we actually observes it, we must consider the expected value with respect to $\boldsymbol{\lambda}$ to obtain the posterior value,

$$\mathrm{PoV}(\boldsymbol{\lambda}) = \mathrm{E}\left(\max_{\boldsymbol{a}\in\mathcal{A}} \{v(\boldsymbol{\lambda}, \boldsymbol{a})\}\right) = \int_{\boldsymbol{\lambda}} \max_{\boldsymbol{a}\in\mathcal{A}} \{v(\boldsymbol{\lambda}, \boldsymbol{a})\} p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}. \tag{3.9}$$

Inserting (3.4) and (3.9) into (3.3) yields the VOI for perfect information. Note that the posterior value is a function of the information we observe.

Returning to the running example, we can compute the posterior value, using (3.9),

$$\begin{aligned} \mathrm{PoV}()\lambda &= \mathrm{E}\left(\max_{\boldsymbol{a}\in\mathcal{A}} \{v(\boldsymbol{\lambda}, \boldsymbol{a})\}\right) \\ &= \sum_{\lambda=0}^{1} (10 - 3\lambda) P(\Lambda = \lambda) + \sum_{\lambda=1.1}^{3} (8 - \lambda) P(\Lambda = \lambda) \\ &\approx 6.5808. \end{aligned} \tag{3.10}$$

Thus, the value of making a decision after acquiring additional information is 6.5808. The VOI in this case is then

$$\mathrm{VOI} = 6.5808 - 6.5 = 0.0808, \tag{3.11}$$

which means that we are willing to pay 0.0808 monetary units for the perfect information about λ.

Imperfect Information

Suppose that we are not able to observe $\boldsymbol{\lambda}$ directly, but through another random variable, $\boldsymbol{y}$. Again, we start with the expression for the prior value. Now, we condition on the observed variable, $\boldsymbol{y}$, with corresponding density $p(\boldsymbol{y})$,

$$p(\boldsymbol{y}) = \sum_{\boldsymbol{\lambda}} p(\boldsymbol{y} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda}). \tag{3.12}$$

The value conditioned on $\boldsymbol{y}$ is given by

$$\max_{\boldsymbol{a}\in\mathcal{A}}\{E(v(\boldsymbol{\lambda},\boldsymbol{a})\mid\boldsymbol{y}\}=\max_{\boldsymbol{a}\in\mathcal{A}}\left\{\int v(\boldsymbol{\lambda},\boldsymbol{a})p(\boldsymbol{\lambda}\mid\boldsymbol{y})d\boldsymbol{\lambda}\right\}. \tag{3.13}$$

Since we do not know $\boldsymbol{y}$ before we actually have observed it, we must take expectation with respect to $\boldsymbol{y}$, as before, yielding

$$\begin{aligned}\mathrm{PoV}(\boldsymbol{y})&=\mathrm{E}\left(\max_{\boldsymbol{a}\in\mathcal{A}}\{E(v(\boldsymbol{\lambda},\boldsymbol{a})\mid\boldsymbol{y}\}\right)=\int_{\boldsymbol{y}}\max_{\boldsymbol{a}\in\mathcal{A}}\{E(v(\boldsymbol{\lambda},\boldsymbol{a})\mid\boldsymbol{y}\}\,p(\boldsymbol{y})d\boldsymbol{y}\\&=\int_{\boldsymbol{y}}\left(\max_{\boldsymbol{a}\in\mathcal{A}}\left\{\int_{\boldsymbol{\lambda}}v(\boldsymbol{\lambda},\boldsymbol{a})p(\boldsymbol{\lambda}\mid\boldsymbol{y})d\boldsymbol{\lambda}\right\}\right)p(\boldsymbol{y})d\boldsymbol{y}.\end{aligned} \tag{3.14}$$

Inserting (3.4) and (3.14) into (3.3) yields the VOI for imperfect information.

Again we return to the running example to illustrate how we can compute the posterior value of imperfect information. We will now assume that we observe y that have the following relation to λ,

$$y=\lambda+\Delta, \tag{3.15}$$

where Δ is a random variable defined on the grid $\Delta\in[-0.1,0,0.1]$. We will assume that λ and Δ are independent. The conditional distribution $P(Y=y\mid\Lambda=\lambda)$ is given by

$$\begin{aligned}P(Y=y\mid\lambda)&=\begin{cases}p & \text{for } y=\lambda-0.1\\ q & \text{for } y=\lambda\\ p & \text{for } y=\lambda+0.1\end{cases}\\ P(Y=y\mid\lambda=0)&=\begin{cases}p+q & \text{for } y=\lambda\\ p & \text{for } y=\lambda+0.1\end{cases}\\ P(Y=y\mid\lambda=3)&=\begin{cases}p & \text{for } y=\lambda-0.1\\ p+q & \text{for } y=\lambda\end{cases}\end{aligned} \tag{3.16}$$

p and q are set so that $2p+q=1$. The first step in computing the posterior value is to find the conditional expectation $E(v(\lambda,a)\mid y)$. To to that, we must compute the conditional expectation $\mathrm{E}(\lambda\mid y)$, which is found by,

$$\mathrm{E}(\lambda\mid y)=\sum_{\lambda}\lambda\cdot P(\Lambda=\lambda\mid Y=y)=\sum_{\lambda}\lambda\cdot\frac{P(Y=y\mid\Lambda=\lambda)P(\Lambda=\lambda)}{P(Y=y)}, \tag{3.17}$$

where the marginal distribution of y is found by

$$P(Y=y)=\sum_{\lambda}P(Y=y\mid\Lambda=\lambda)P(\Lambda=\lambda). \tag{3.18}$$

Table 3.1: PoV and VOI for some values of p and q.

p	q	PoV	VOI
$\frac{1}{20}$	$\frac{9}{10}$	6.5804	0.0804
$\frac{1}{6}$	$\frac{2}{3}$	6.5795	0.0795
$\frac{1}{3}$	$\frac{1}{3}$	6.5782	0.0782

Let $\mu(y)$ denote the conditional expectation of λ given y. The posterior value is then given by

$$\begin{aligned} \mathrm{PoV} &= \mathrm{E}\left(\max_{\boldsymbol{a}\in\mathcal{A}}\{E(v(\boldsymbol{\lambda},\boldsymbol{a}) \mid \boldsymbol{y}\}\right) \\ &= \sum_{y} \max\{10-3\mu(y), 8-\mu(y)\}\, P(Y=y). \end{aligned} \tag{3.19}$$

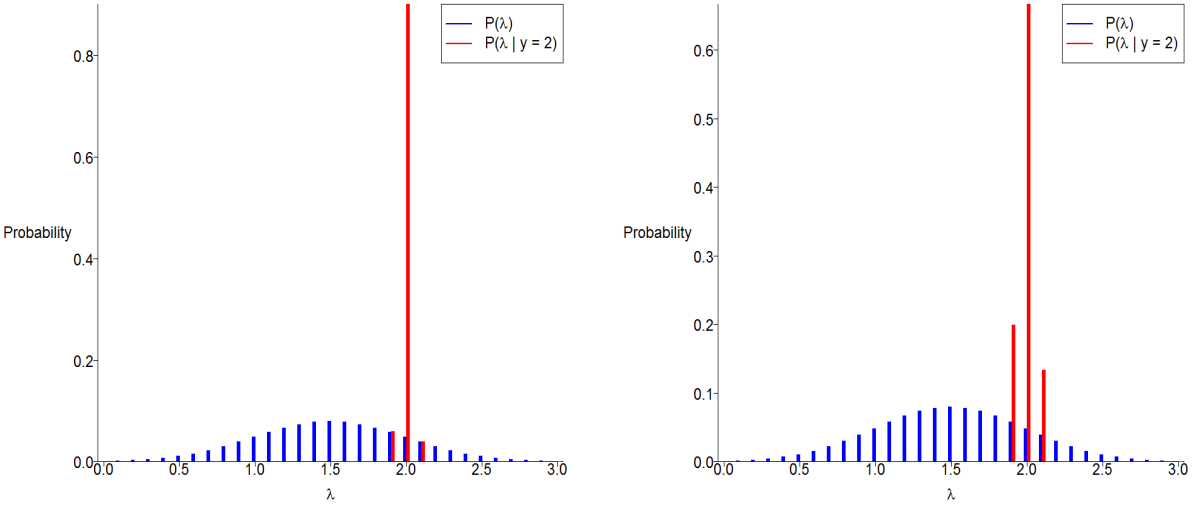
The posterior value for some values of p and q are computed and listed in Table 3.1. We note that when p increases and q decreases, the PoV also decreases. This is because a larger p gives higher uncertainty in y which again makes the information less accurate. The value of imperfect information is smaller than the value of perfect information, which is as we would expect. When $q \to 1$, the PoV for imperfect information will approach the PoV for perfect information.

Histograms showing the distribution of λ and $\lambda \mid y = 2$ is displayed in Figure 3.2. We see that when q is close to 1, in Figure 3.2a, it is very likely that $\lambda = 2$ when we observe $y = 2$. In Figure 3.2b we see that when q decreases, the probability of $\lambda = 2$ when we observe $y = 2$ decreases as well. This means that the uncertainty in the observed quantity y increases and hence the value of knowing y will decrease.

When developing the expressions for the PoV above, we assumed that we had information about the uncertain quantity $\boldsymbol{\lambda}$ in its entirety. In some cases, we are only able to observe parts of $\boldsymbol{\lambda}$, that is we observe a subset $\boldsymbol{\lambda}_{\mathcal{K}} \subset \boldsymbol{\lambda}$. This is referred to as partial information. Analytic expressions and algorithms for estimating these quantities can be found in Eidsvik et al. (2015).

3.3 Estimation of Value of Information

Analytic expressions for prior- and posterior values are in most cases unavailable. We will restrict to the case when $\boldsymbol{\lambda}$ is discrete, which is what we will be considering in our case study in Chapter 4. Algorithms for estimating prior- and posterior value for continuous $\boldsymbol{\lambda}$ can be found in Eidsvik et al. (2015). We also assume that the set of alternatives, $\mathcal{A}$, is a finite set. The



(a) $p = \frac{1}{20}$ and $q = \frac{9}{10}$ **(b)** $p = \frac{1}{6}$ and $q = \frac{2}{3}$

Figure 3.2: Histograms showing $p(\lambda)$ and $p(\lambda \mid y = 2)$ for two set of values for p and q.

value function will be unknown in our case study, so we introduce the notation $\widehat{v}$ for the value function to emphasize that it is estimated. The estimation of prior- and posterior value will be treated separately, in the following.

3.3.1 Prior Value Estimation

When estimating the prior value, defined in (3.4), we first consider an estimate of the expected function. By the definition of expected value for a discrete random variable we have,

$$\mathrm{E}(\widehat{v(\boldsymbol{\lambda}, \boldsymbol{a})}) = \sum_{\boldsymbol{\lambda} \in \mathcal{L}} \widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a}) p(\boldsymbol{\lambda}). \tag{3.20}$$

The prior value estimator can then be expressed as

$$\widehat{\mathrm{PV}} = \max_{\boldsymbol{a} \in \mathcal{A}} \left\{ \sum_{\boldsymbol{\lambda} \in \mathcal{L}} \widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a}) p(\boldsymbol{\lambda}) \right\}. \tag{3.21}$$

Note that if $\mathcal{A}$ is unbounded, the maximization must be done with some numerical optimization method.

3.3.2 Posterior Value Estimation

The estimate of the posterior value will depend on the type of information we have. We will, therefore, provide formulas for both perfect and imperfect information.

Substituting the integral with a sum in (3.9) yields

$$\widehat{\mathrm{PoV}} = \sum_{\boldsymbol{\lambda} \in \mathcal{L}} \max_{\boldsymbol{a} \in \mathcal{A}} \{\widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a})\}\, p(\boldsymbol{\lambda}). \tag{3.22}$$

for perfect information. For imperfect information we have

$$\widehat{\mathrm{PoV}} = \sum_{\boldsymbol{y}} \max_{\boldsymbol{a} \in \mathcal{A}} \{\mathrm{E}\left(\widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a}) \mid \boldsymbol{y}\right)\}\, p(\boldsymbol{y}). \tag{3.23}$$

The VOI estimator is then simply

$$\widehat{\mathrm{VOI}} = \widehat{\mathrm{PoV}} - \widehat{\mathrm{PV}}. \tag{3.24}$$

When we have a finite set of alternatives, and the value function, the computation of $\widehat{\mathrm{VOI}}$ is straightforward. However, when the value function is unknown and must be estimated, the computation of $\widehat{\mathrm{VOI}}$ is more complicated and time-consuming. In our case study in Chapter 4, the value function will be unknown so we will use Bayesian optimization with a GP surrogate model to maximize the value function and to compute $\widehat{\mathrm{VOI}}$.

When using Bayesian optimization, we are treating the value function as stochastic which means that there will be uncertainty in the estimate of VOI. We will, therefore, in the following briefly discuss how to access the accuracy and the uncertainty of the VOI estimate.

3.3.3 Accuracy of the VOI Estimate

When estimating VOI it is of interest to know how accurate this estimate is. Typically, we would consider the variance of the estimator, but the analytic expression for the variance of VOI is not available. Instead, we can look at how the VOI estimate evolves when we are optimizing the value function. Recall that in Bayesian optimization we are sequentially choosing points to evaluate based on the fitted surrogate model. Thus, in each iteration, we have an estimate of the value function that can be used to compute $\widehat{\mathrm{VOI}}$. The evolution of $\widehat{\mathrm{VOI}}$ should converge, and stabilize around the final estimate of VOI when the number of evaluations grows. We believe that this is a useful diagnostic in our situation, as it monitors the convergence of the VOI estimate over evaluations.

In addition to a point estimate of the VOI, it is of interest to quantify the uncertainty of the estimate. Since we have a GP representation of the value function, one can generate samples of the value function to explore the variability of the VOI estimate. We will next discuss a method to access this variability using the samples from the GP. This is not directly connected to the conditional mean uncertainty, however, which is really required in the $\widehat{\mathrm{VOI}}$ calculation.

Recall that when we use a GP to fit the value function, every finite subset of observations is normally distributed. Let $\widehat{v}$ the value function obtained through Bayesian optimization, where μ and Σ are the posterior mean and covariance in the fitted GP, respectively. The realizations v^b

is drawn by

$$v^b = \mu + L \cdot N_n(\mathbf{0}, \mathbb{I}_n), \qquad b = 1, ..., B, \tag{3.25}$$

where L is the Cholesky decomposition of Σ, so that $LL^T = \Sigma$. Note that the generated value, v^b, will be a sample and not the conditional mean. For each sample, we can go through the PV and PoV calculation. An estimate of the VOI based on $\widehat{\mathrm{VOI}}^b$ for $b = 1, ..., B$ is then

$$\overline{\mathrm{VOI}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathrm{VOI}}^b, \tag{3.26}$$

and an estimate of the standard deviation is

$$\mathrm{SD}(\widehat{\mathrm{VOI}}) = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\mathrm{VOI}}^b - \overline{\mathrm{VOI}} \right)^2 \right)^{\frac{1}{2}}. \tag{3.27}$$

In general, we have that $\widehat{\mathrm{VOI}} \neq \overline{\mathrm{VOI}}$ because the VOI is a non-linear function, and the expected value of the functional is different from the functional of the mean. The definition of PoV state that PoV should be computed using the conditional expectation, $\mathrm{E}(\widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a}) | \boldsymbol{y})$, and not sampled values. The estimate of the standard deviation could still give us an indication of the magnitude of the uncertainty.

Chapter 4

Case Study - NOWIcob Simulation Model

In this chapter we will use the theory from Chapter 2 and 3 to show how one can estimate the VOI in a complex decision situation. The decision problem under consideration is the selection of optimal O&M strategy in order to maximize the profits from an offshore wind farm. The optimal strategy depends on several uncertain variables so it is of interest to estimate the value knowing a subset of these uncertainties. We focus on the failure rate for a medium repair failure here. We will use the simulation tool NOWIcob to evaluate different vessel fleet combinations so that we can estimate VOI. An introduction to NOWIcob and the decision problem is given in Section 4.1, and the estimation of VOI in such a decision situation is described in Section 4.2. The final algorithm and some case specific details for the estimation of VOI are discussed in Section 4.3.

4.1 Introduction

NOWIcob (**N**orwegian **o**ffshore **wi**nd power life cycle **co**st and **b**enefit model) is a simulation tool developed by SINTEF, designed to mimic the daily operation of a offshore wind farm. The primary use of NOWIcob is for simulation and optimization of different aspects of an offshore wind farm (Hofmann et al., 2017a). The simulation is used to estimate performance parameters such as profits and operation costs by simulating maintenance activities and related logistics of a wind farm. It can also be used as a decision support tool for decision problems such as choosing the vessel fleet or the location of the maintenance bases. We will focus on the decision support application of NOWIcob.

The model has several input variables, and we can divide all variables into two distinctions - decision variables and uncontrollable/uncertain variables. Decision variables are what the user of the simulation tool may control, for instance the vessel fleet mix or the location of maintenance bases. The choice of these variables makes up the strategy for operation and maintenance of the wind farm. We will restrict us to three decision variables here - the number of Crew Transfer Vessels (CTV), the number of Surface Effect Ships(SES) and the number

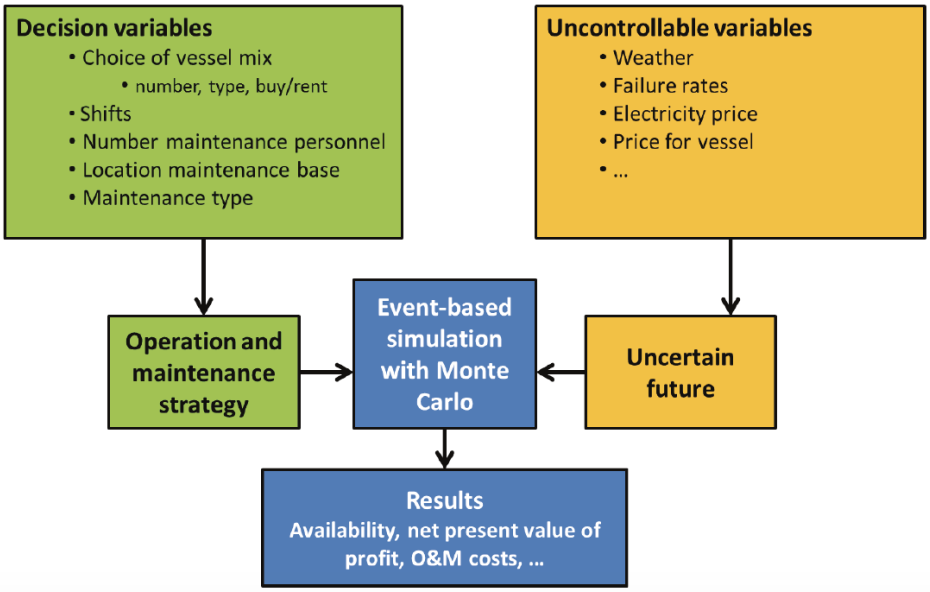


Figure 4.1: Decision variables and uncertain/uncontrollable variables in NOWIcob. The figure is taken from Hofmann et al. (2017a).

of maintenance personnel. Vessels are defined by properties such as travel speed, passenger capacity and what weather it may operate in, e.g wind speed and wave height. Both CTV and SES can transfer up to 12 people from the harbor to the wind farm. CTV has a top speed of 20 knots and can access the wind farm with wave heights up to 1.5 meters. SES has a top speed of 35 knots and tolerates waves up to 2 meters high. A more detailed description of CTV and SES can be found in the appendix and Hofmann et al. (2017b), where also other vessels are described in detail.

Uncertain variables include weather, e.g wave heights and wind speed, and much more. These variables are the basis for the uncertain future in the model. Uncontrollable variables may be the price for a particular boat or the annual salary of an employee. These variables are fixed and are not subject for change in any part of the simulation. Figure 4.1 depicts the distinction between the different types of input variables in NOWIcob in a simple flowchart. The next step after the input variables are specified and processed is an event-based simulation with Monte Carlo (MC) where the operation of the wind farm is mimicked over time.

Since there are several uncertain variables involved in the simulation, the output will vary between simulations with the same input. By running several MC iterations, the uncertainty will be captured. Each MC iteration provides several performance parameters, for instance, the availability and the profit of the wind farm. Here we are most interested in the profit, which we will denote by $v(\boldsymbol{\lambda}, \boldsymbol{a})$. That is, the profit, or value, depends on a set of uncertain variables, $\boldsymbol{\lambda}$, and decision variables $\boldsymbol{a}$. The profit is measured in GBP in NOWIcob.

NOWIcob distinguishes between two types of maintenance tasks - preventive and corrective. The former is maintenance tasks that are scheduled to predetermined time steps, for instance

annual services that occur once a year for each turbine. The latter is maintenance tasks related to failures that occur randomly. These failures are divided into categories based on their severity. Failures within each category are assumed to follow a homogeneous Poisson process with an average yearly rate. This rate might differ between years, and can for instance have a bathtub shape so that the rate is higher the first few and last few years. The assumption of a homogeneous Poisson process means that the time t until next failure is exponentially distributed,

$$t \mid \lambda \sim \exp(\lambda), \tag{4.1}$$

where λ is the failure rate. We condition on λ in (4.1) to emphasize that the failure rate itself might be unknown, which is a reasonable assumption for several reasons. Firstly, obtaining an estimate of the failure rate for one particular category requires a comprehensive data collection. This collection of data is done over several years since some components might have an expected lifetime of up to 20 years (Faulstich et al., 2011). Such a long period of data collection raises another problem - the components in the wind turbines will be further developed and probably improved over the years. Thus, the estimates based on the data collection may not be accurate for future, improved components.

Comprehensive data collections have been done, but the resulting data sets are confidential for the public. Faulstich et al. (2011) analyze the results from a large monitoring survey for onshore wind turbines conducted by Fraunhofer IWES. This survey collected 64 000 maintenance and repair reports from 1500 wind turbines. The result from the analysis of this concludes that the annual failure rate for minor repairs is 1.8 and 0.6 for major repairs. Here, minor repairs are defined as failures with downtimes less than a day, while a major repair is defined as a failure with downtimes that exceeds one day.

Carroll et al. (2016) has conducted an analysis of a data set based on ~ 350 offshore wind turbines, where all turbines are between 3 and 10 years old from between 5 and 10 wind farms in Europe. Exact numbers of turbines, the age of the turbines and number of wind farms are not provided in the paper for confidentiality reasons. Specifications of the types of wind turbines are not provided either. Here, a minor repair is defined as a repair due to a failure that costs less than 1000 euros. A major repair has a cost between 1000 and 10000 euros, and repair costs above this are defined as a major replacement. These costs are based on material costs only. The paper concludes that the annual failure rate for minor repairs is 6.81, 1.17 for major repair and 0.29 for major replacement.

Dinwoodie et al. (2015) presents a reference case for simulation of O&M tasks for offshore wind farms. Five different failure categories are used here - manual reset, minor repair, medium repair, major repair and major replacement. The proposed values for the annual failure rates are 7.5 for manual reset, 3.0 for minor repair, 0.275 for medium repair, 0.04 for major repair and 0.08 for major replacement.

The large variation of estimates of failure rates makes it difficult to draw any conclusion of

what the true values are. The estimates from the proposed literature use different categorizations of failure rates and are based on different data that are collected on different premises, e.g onshore vs. offshore. This motivates the idea of treating the error rate λ as a random variable with some distribution $p(\lambda)$. We will consider four different distributions of λ, mainly inspired by Dinwoodie et al. (2015), which we will address in Section 4.3.2.

4.2 Estimation of Value of Information in O&M Strategy Decision Problem

We will use NOWIcob as a decision tool for the selection of vessel fleet and number of maintenance personnel for an offshore wind farm. In this context, we are interested in estimating the value of knowing the failure rate for a failure in the medium repair category. In this section, we will set up the problem, using the notation introduced in Chapter 3.

4.2.1 Variables

We will denote the uncertain variables by $\boldsymbol{\lambda}$ as before. All uncertain variables such as weather, electricity prices, etc. are included in $\boldsymbol{\lambda}$. We not will dive deep into the impact of every uncertain variable, except the failure rate for a failure in the medium repair failure category. We will denote this variable by $\lambda \in \boldsymbol{\lambda}$ in the following. A discretized state space is used for this failure rate. In reality, will λ be continuous, but we must consider a finite set of feasible values to limit the number of evaluations in NOWIcob. We will assume that $\lambda \in \Omega_\lambda = \{0.1, 0.2, ..., 1.5\}$, $|\Omega_\lambda| = M = 15$ and that λ has a known probability mass function $p(\lambda)$.

The decision variables we will consider here are

- $a_p \in \mathcal{A}_p = \{10, 11, ..., 50\}$, $|\mathcal{A}_p| = 41$
- $a_{CTV} \in \mathcal{A}_{CTV} = \{0, 1, ..., 4\}$, $|\mathcal{A}_{CTV}| = 5$
- $a_{SES} \in \mathcal{A}_{SES} = \{0, 1, ..., 4\}$, $|\mathcal{A}_{SES}| = 5$

Here, a_p denotes the number of personnel available on each shift, a_{CTV} denotes the number of CTV and a_{SES} denotes the number of SES. We will denote the set of all possible combinations of a_p, a_{CTV} and a_{SES} by $\mathcal{A}$.

4.2.2 Value Function

The quantity of interest is the profit of the offshore wind farm. We will denote the value function by v, and assume that it is a function of the variables discussed in Section 4.2.1,

$$v = v(\boldsymbol{a}, \lambda), \qquad \boldsymbol{a} = (a_p, a_{CTV}, a_{SES}). \tag{4.2}$$

v will depend on many other variables, such as the location of the wind farm, but we will treat these variables as constant in this case study. A major part of estimating the value of information is the optimization of the value function. An evaluation of the value function is time consuming, and we have an input space of size $|\mathcal{A}_p| \cdot |\mathcal{A}_{CTV}| \cdot |\mathcal{A}_{SES}| \cdot |\Omega_\lambda| = 15375$. One can evaluate all points, which will yield the optimal solution, but that will take a lot of time because each evaluation could take several minutes. A smarter approach is to use Bayesian optimization that will be more effective because fewer points need to be evaluated. This is also a more scalable approach if one wants to include even more decision variables.

We will use Bayesian optimization with a GP as a surrogate model to maximize v. To do so, we must make some assumption about v and the variables. Firstly, we must specify the covariance structure. We will assume that v is a smooth function so that a separable Gaussian covariance structure is applicable. Then we also assume that the variables have a different spatial dependency, which seems like a reasonable assumption. Two points where a_{CTV} differ by one should have a different covariance than two points where a_p differ by one. The distance in space is the same, but while a difference of one person does not matter much, will a difference of one vessel have a large impact on the profit.

Since there are many uncertain variables involved in the simulation tool, we must have a way to include uncertainty in our model. The uncertainty is obtained by running several MC iterations in NOWIcob. Because of randomly occurring failures will the profit vary between simulations. By including a nugget in the GP, as discussed in Section 2.2.1, we include this uncertainty in the estimate of the value function.

The optimization of the value function is done sequentially, as described in Algorithm 1. First, an initial batch of points is obtained and used to fit a GP. Next, we use an acquisition function to suggests the next points for evaluation, and then we add these points to the set of observations. These two steps are repeated until we reach a stopping criterion.

4.3 Algorithm for Estimating Value of Information

In the following, we will go through the algorithm for estimating the VOI in our case study. First, in Section 4.3.1, will the case-specific acquisition function will be formulated. In Section 4.3.3 will the algorithm used to estimate VOI be presented.

4.3.1 Acquisition Function

The choice of acquisition function must be done based on the problem under consideration, in particular on the data that are involved. We are not only interested in maximizing the function, but we want to find the maximum value for each value of λ. This makes standard acquisition functions inadequate since areas where λ is large will not be explored at all because the profit will be lower in these areas. We must, therefore, add a mechanism that forces the exploration

of promising points for all rates.

Let $\mathcal{M}_k$ be the optimal value thus far for each rate λ_k, that is

$$\mathcal{M}_k = \max_{\boldsymbol{a}\in\mathcal{A}}\{v(\boldsymbol{a}, \lambda_k)\}, \quad k = 1, ..., M. \tag{4.3}$$

Note that $\mathcal{M}_k$ corresponds to y^* when we derived the acquisition functions in Section 2.3. The EI can then be expressed as

$$\mathrm{EI}\left(\boldsymbol{a}, \lambda_k\right) = \mathrm{E}\left(\max\{v(\boldsymbol{a}, \lambda_k) - \mathcal{M}_k, 0\}\right). \tag{4.4}$$

The numerical value for EI can be computed by using (2.51). The EI in each point will now be set relative to the current maximum for each rate, and not the overall maximum. The algorithm for finding the expected improvement for all feasible points is summarized in Algorithm 2.

Algorithm 2 Expected improvement for case study

1. Compute $\mathcal{M}_k$ for $k = 1, 2, ..., M$
2. **for** $k = 1, 2, ..., M$
3. **for all** $\boldsymbol{a} \in \mathcal{A}$
4. Compute $\mathrm{EI}(\boldsymbol{a}, \lambda_k) = \mathrm{E}\left(\max\{v(\boldsymbol{a}, \lambda_k) - \mathcal{M}_k, 0\}\right)$
5. **end for**
6. **end for**
7. **return** EI

Since we must explore areas for all values of the rate, it is reasonable to force the selection of one point for each value of rates. Thus, each time we select new points for evaluation, M points are chosen - one for each value of λ. In addition to force the selection of points for all values of failure rates, we also want a mechanism that ensures the selected points to have some distance between each other. For instance, we do not want to evaluate both $(a_p, a_{CTV}, a_{SES}, \lambda) = (22, 1, 1, 0.3)$ and $(a_p, a_{CTV}, a_{SES}, \lambda) = (22, 1, 1, 0.4)$ since these points are likely to have very similar value. We will therefore use the scheme in (2.53) to select new points for evaluations, forcing the new points to be a certain distance from each other.

The final algorithm for selecting new points for evaluation is summarized in Algorithm 3. In line 5 is a new point proposed based on the expected improvement. If this point has no nearby points that are already accepted candidates for new evaluation and has an expected improvement larger than ϵ, the point is added to the set of candidates. Otherwise, the point is removed as a candidate for evaluation. We will define distance, $\|\boldsymbol{x}_{\text{cand}} - \boldsymbol{x}\|$, as

$$\|\boldsymbol{x}_{\text{cand}} - \boldsymbol{x}\| = |a^p_{\text{cand}} - a^p| + |a^{CTV}_{\text{cand}} - a^{CTV}| + |a^{SES}_{\text{cand}} - a^{SES}| + 10 \cdot |\lambda_{\text{cand}} - \lambda|, \tag{4.5}$$

where $\boldsymbol{x} = (\boldsymbol{a}, \lambda)$ is used to ease the notation. We will use $\delta = 15$ as a minimum distance between new points for evaluation. The algorithm proceeds until $M = 15$ new points are selected, or until there are no more candidates with an EI higher than ϵ.

Algorithm 3 New point selection for case study

1. Compute EI for all feasible points
2. Initialize $X_{\text{cand}} = \{\}$
3. **for** $k = 1, 2, ..., M$
4. **repeat**
5. Propose $\boldsymbol{x}_{\text{cand}} = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, \text{EI}(\boldsymbol{a}, \lambda_k)$
6. **if** $\|\boldsymbol{x}_{\text{cand}} - \boldsymbol{x}\| > \delta \quad \forall \boldsymbol{x} \in X_{\text{cand}}$ and $EI(\boldsymbol{x}_{\text{cand}}) > \epsilon$
7. $X_{\text{cand}} = \{X_{\text{cand}}, \boldsymbol{x}_{\text{cand}}\}$
8. **else**
9. Remove $\text{EI}(\boldsymbol{x}_{\text{cand}})$ from EI
10. **until** $\boldsymbol{x}_{\text{cand}} \in X_{\text{cand}}$ or no more candidates available
11. **end for**
12. **return** X_{cand}

4.3.2 Distribution of Failure Rate

We argued in Section 4.1 that it is reasonable to treat the failure rate as stochastic, but the distribution remains unknown. We will therefore consider four different distributions, which are reasonable guesses on what might actually be the truth. These distributions are

1. $p_1(\lambda) = \text{Gamma}(k = 0.903, \beta = 0.316)$
2. $p_2(\lambda) = \text{Gamma}(k = 0.226, \beta = 1.263)$
3. $p_3(\lambda) = N(0.285, 0.3^2)$
4. $p_4(\lambda) = N(0.285, 0.6^2)$

where k is the shape parameter and β is the scale parameter in the gamma distribution. The distributions is truncated on the interval $[0.1, 3]$ and discretized into the feasible values for λ. The distributions is plotted in Figure 4.2. The mean in all four distributions is 0.285, which is the failure rate for a medium type repair suggested by Dinwoodie et al. (2015). The standard deviation in p_1 and p_3 is 0.3, and 0.6 for p_2 and p_4.

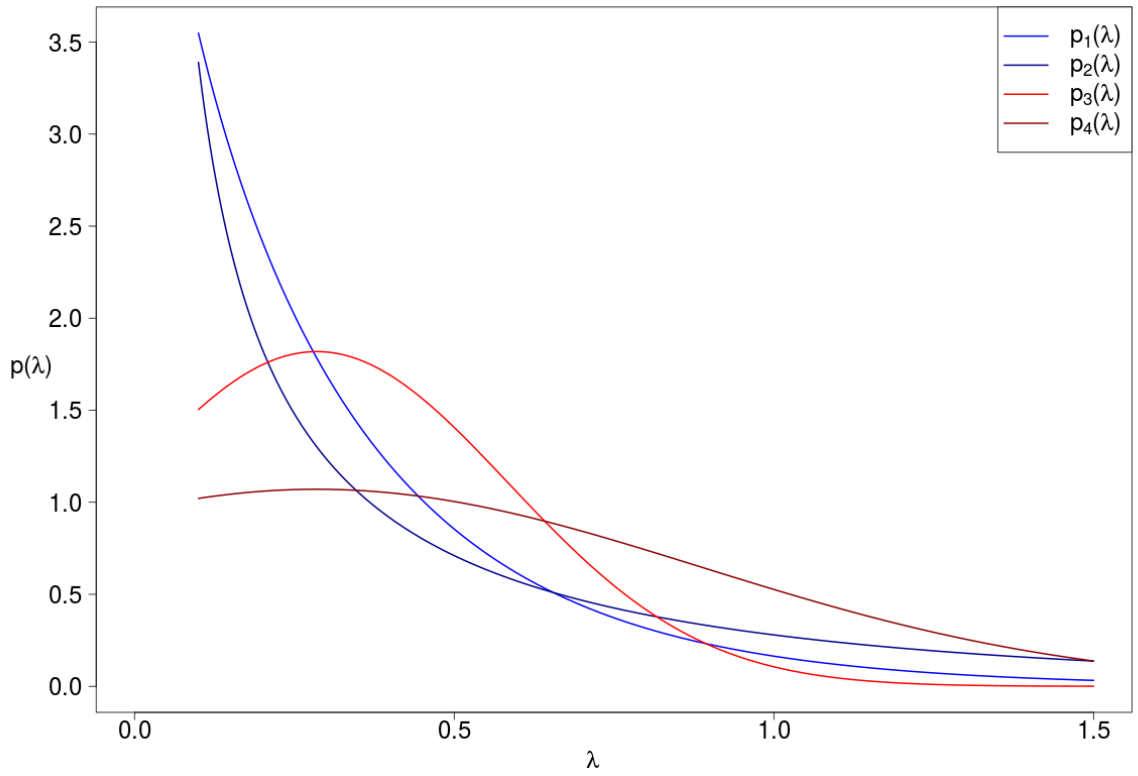


Figure 4.2: Plot of the three suggested distributions of the failure rate, λ.

4.3.3 Algorithm

The resulting algorithm for estimating the VOI is summarized in Algorithm 4. The initial points $\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_0}$ are selected by using a Latin Hypercube sampling procedure. See for instance Santner et al. (2013) for a thorough review of Latin Hypercube sampling and other space filling sampling procedures. To ease the notation in the algorithm will $\boldsymbol{x} = (\boldsymbol{a}, \lambda)$ denote all the variables used to fit the GP.

The first step in each iteration is to fit a GP to the observations in $\mathcal{D}_n$. The necessary theory for this was discussed in Section 2.2. The library `laGP` (Gramacy and Sun, 2017) in R provides an efficient way of fitting a GP, including the estimation of the hyperparameters discussed in Section 2.2.1. Next, new points for evaluations are chosen, using Algorithm 3. The new points are then evaluated in NOWIcob, which is our "black-box function". NOWIcob is written in MATLAB, which is called from R. Finally are all new observations added to the set of observations, $\mathcal{D}_n$.

The search of a the optimal configuration for each failure rate continues until there are no more candidates, X_{cand}, that have an expected improvement above the lower limit, ϵ. Then may the $\widehat{\text{VOI}}$ be computed using the suggested distribution for λ discussed in 4.3.2.

We will consider both perfect and imperfect information about the failure rate. For the computation of $\widehat{\text{VOI}}$ with perfect information, we can plug the estimated value function and distribution for λ into (3.22). For imperfect information, we will assume the same relation

Algorithm 4 Estimation of VOI for case study

1. Define
   - Feasible values for decision variables, a_p, a_{CTV}, a_{SES}
   - M feasible values for uncertain variable, λ
2. Specify
   - Acquisition function $\alpha(\boldsymbol{x})$
   - Distribution of failure rate, $p(\lambda)$
3. Obtain initial sample
   - Set initial sample size, n_0
   - Sample initial points $\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_0}$
   - Evaluate $y_i = f(\boldsymbol{x}_i)$ for $i = 1, ..., n_0$
   - $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_0}$
4. **repeat**
5. Fit the surrogate model, $\mathcal{GP} \mid \mathcal{D}_n$
6. Find $X_{\text{cand}} = \underset{\boldsymbol{x} \in \Omega}{\operatorname{argmax}}\, \alpha(\boldsymbol{x})$
7. Evaluate $y_i = f(\boldsymbol{x}_i) \quad \forall \boldsymbol{x}_i \in X_{\text{cand}}$
8. Augment $\mathcal{D}_n$ with new observations
9. **until** $X_{\text{cand}} = \emptyset$
10. Compute $\widehat{\text{VOI}} = \widehat{\text{PoV}} - \widehat{\text{PV}}$
11. **return** $\widehat{\text{VOI}}$

between what we observe, y, and the failure rate as what we used in the example in Section 3.2.2, that is

$$y = \lambda + \Delta, \qquad \Delta \in \{-0.1, 0, 0.1\}. \tag{4.6}$$

The conditional distribution $P(Y = y \mid \Lambda = \lambda)$ is given in (3.16). The posterior value for imperfect information can then be estimated by

$$\begin{aligned} \widehat{\mathrm{PoV}} &= \mathrm{E}\left(\max_{\boldsymbol{a}\in\mathcal{A}}\left\{\mathrm{E}\left(\widehat{v}(\boldsymbol{a},\lambda) \mid y\right)\right\}\right) \\ &= \sum_{y}\max_{\boldsymbol{a}\in\mathcal{A}}\left\{\sum_{\lambda}\widehat{v}(\boldsymbol{a},\lambda)\frac{P(Y=y \mid \Lambda=\lambda)P(\Lambda=\lambda)}{P(Y=y)}\right\}P(Y=y). \end{aligned} \tag{4.7}$$

$P(\Lambda = \lambda)$ is given by the proposed distributions in Section 4.3.2 and $P(Y = y)$ is easily computed by using (3.18).

Chapter 5

Experimental Results

In this chapter, we will present the results from the estimation of the value of knowing the failure rate for a medium repair type of failure. We will first present some results from the optimization of the value function in Section 5.1. Next, we will present the estimated VOI for different distributions of the failure rate for both perfect- and imperfect information in Section 5.2. In Section 5.2.3 we will look at the accuracy in the VOI estimate and briefly discuss the uncertainty in the estimate.

5.1 Optimization of the Value Function

The verification of the proposed optimal values is not straightforward since the value function is unknown. We will get a better understanding of the optimization process by looking at the EI at different stages. Since the value function has four input variables, it can not be visualized in its entirety. We will, therefore, consider subsets of the input space, where the vessel alternatives, a_{CTV} and a_{SES} are fixed.

Figure 5.1 shows contour plots of the relative EI for the points with $(a_{CTV}, a_{SES}) = (0, 1)$ plotted at different stages in the optimization process. The relative EI is the ratio between the EI and the current best observation thus far. In the upper left corner, we see that there are many points that are candidates for being the maximum. Most of these points are for low values of λ. Points with high a_p at this stage are assumed to be promising points, but in the following iterations are these points discarded. We see that the most promising points moves toward lower values of a_p, as we can see in Figure 5.1b, 5.1c and 5.1d.

Contour plots of the relative EI for the points with $(a_{CTV}, a_{SES}) = (0, 2)$ are shown in Figure 5.2. After the initial batch the situation is much alike what we saw for $(a_{CTV}, a_{SES}) = (0, 1)$ as can be seen in Figure 5.2a. In the following iterations are many points discarded as candidates for being the maximum, and the area of high EI gathers around $a_p = 20$ and $\lambda > 0.4$.

We see in Figure 5.1 and 5.2 that we discard many more points than we evaluate since just 15 points are evaluated in each iteration. This is just what we wanted - minimize the number of

evaluations in NOWIcob, and still find good solutions.

When the relative EI is lower than 0.001 for all feasible points, we choose to terminate the algorithm and we have obtained the estimate of v. In total was 50 iterations and 499 evaluations in NOWIcob required. This is a large reduction from the number of feasible points, 15375. Figure 5.3a and 5.3b depict the estimated posterior mean and uncertainty, respectively, for $(a_{CTV}, a_{SES} = (0, 2)$. The estimated value function is smooth and the shape is what we would expect - monotonically decreasing when λ increases. For $a_p > 24$, the profit decays monotonically which is what we expect since all values of personnel larger than 24 are only an additional expense.

The posterior standard deviation is shown in Figure 5.3b. The uncertainty should be lower in the areas that are well explored, and higher elsewhere. By comparing with Figure 5.1 we see that the areas that had a significant relative EI are the same that have low uncertainty. The uncertainty is high in the same areas as the points that were discarded as candidates for being the maximum. Thus, promising areas are well explored and less promising areas are not.

In Table 5.1 the optimal configurations are shown, as suggested by the fitted GP, for all feasible values of the failure rate, λ. The value function, $\widehat{v}$, decays as the failure rate increases just as expected. There are two different vessel configurations, $(a_{CTV}, a_{SES}) = (0, 1)$ and $(a_{CTV}, a_{SES}) = (0, 2)$. For low values of λ is one vessel enough, while for higher rates there are more failures so an additional vessel is required to avoid too much down time. We see that the optimal configuration are in the areas that had a high relative EI in Figure 5.1 and 5.2.

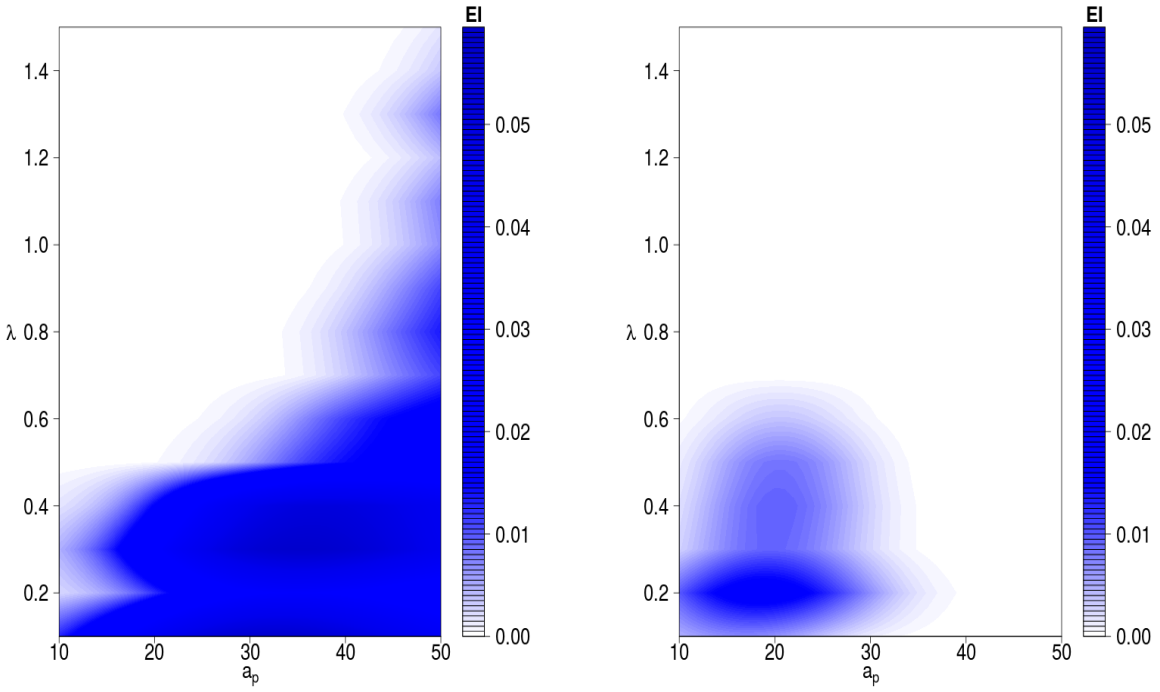
We used a separable Gaussian covariance function, defined in (2.34), when estimating the value function with a GP. Then we assumed that the variables had a different spatial dependency. In the resulting GP, the length scale parameters, $\boldsymbol{\theta} = (\theta_p, \theta_{CTV}, \theta_{SES}, \theta_\lambda)$, was estimated to be

- $\theta_p = 2915$
- $\theta_{CTV} = 11.9$
- $\theta_{SES} = 3.4$
- $\theta_\lambda = 37$

This means that our assumption that the variables have different spatial dependency was correct. Two points that differ by one in a_p are more correlated than two points that differ by one vessel or rate. A small change in the number of personnel should not affect the overall profit much. Changing the number of vessels might lead to large losses either because there are too few vessels available to transport personnel to the wind farm, or because there are too many vessels so that it becomes an unnecessary additional expense.

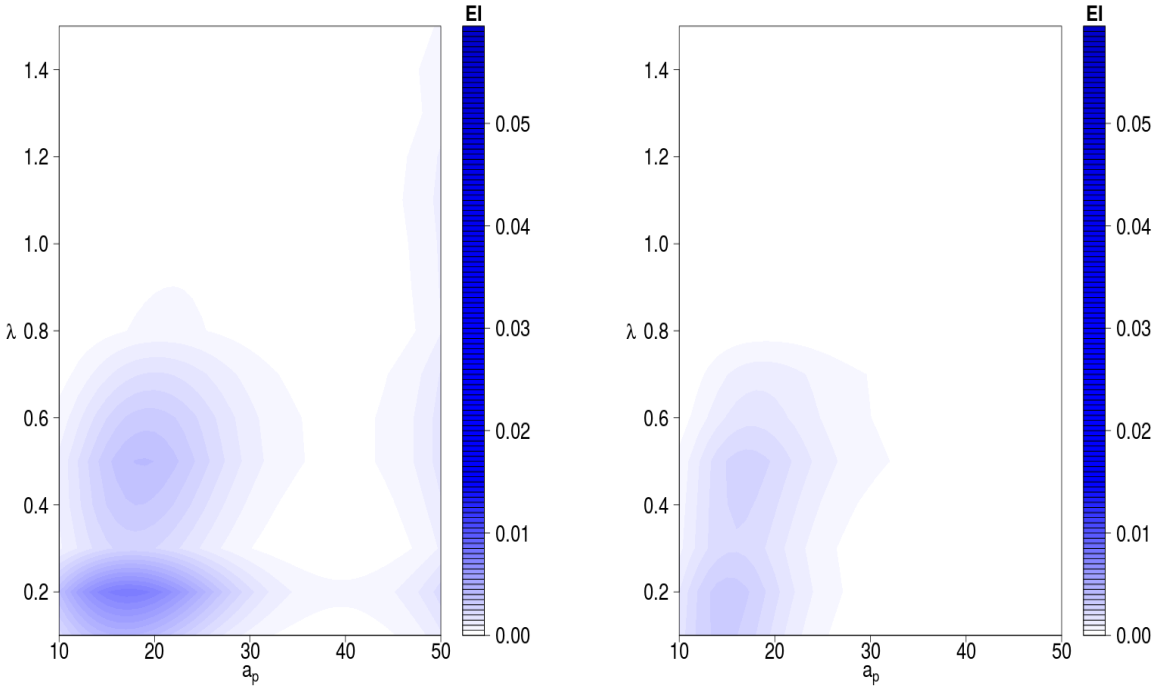
5.1.1 Check of the Normality Assumption

When we are estimating the value function fitted with a GP, we are assuming that each point is Gaussian distributed with some mean and variance. We briefly mentioned in Section 2.2.1 that



(a) Relative expected improvement after initial batch of samples.

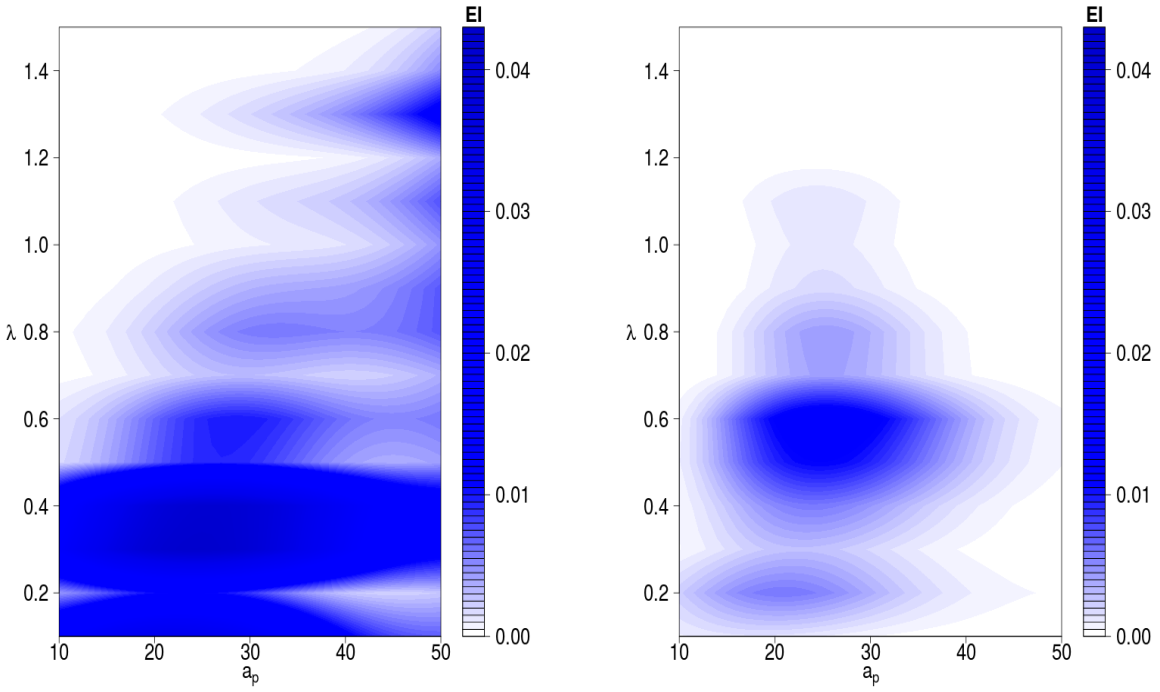
(b) Relative expected improvement after 5 iterations of adaptive sampling.



(c) Relative expected improvement after 10 iterations of adaptive sampling.

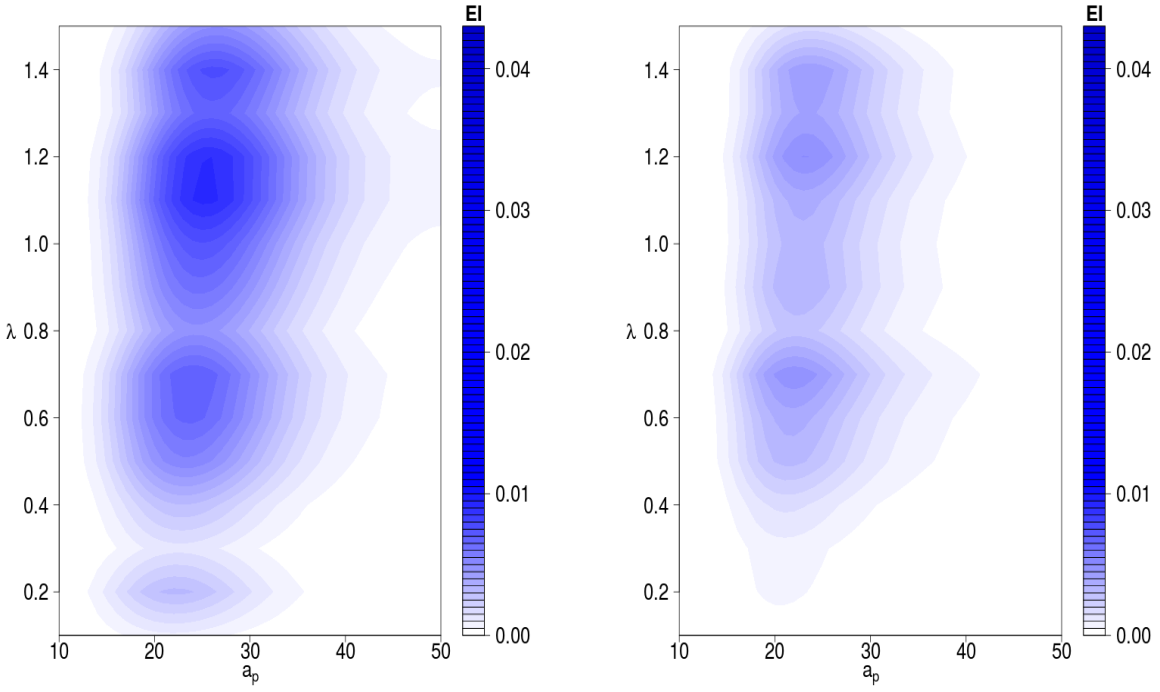
(d) Relative expected improvement after 20 iterations of adaptive sampling.

Figure 5.1: Relative expected improvement at different stages of the optimization process for $(a_{CTV}, a_{SES}) = (0, 1)$.



(a) Relative expected improvement after initial batch of samples.

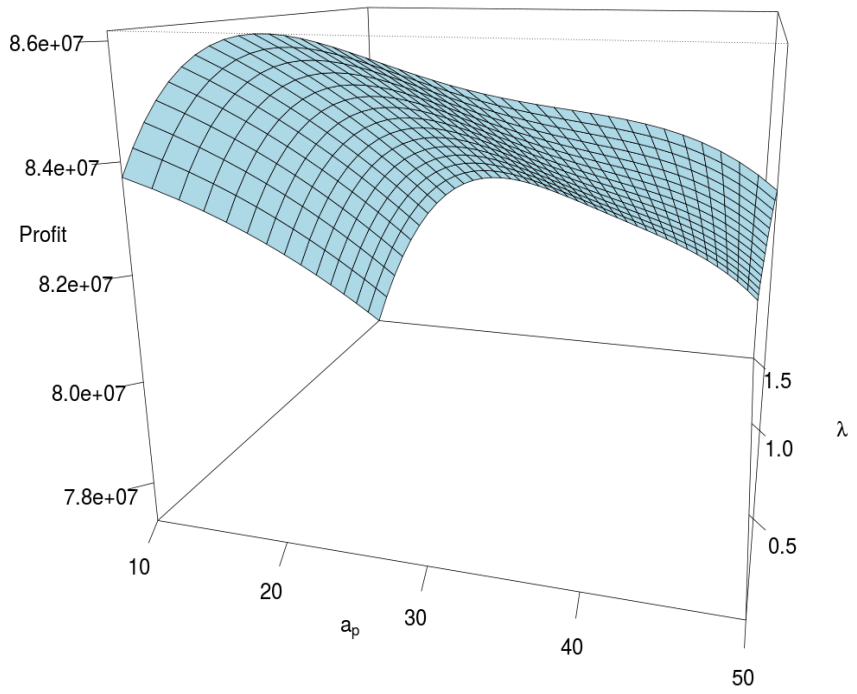
(b) Relative expected improvement after 5 iterations of adaptive sampling.



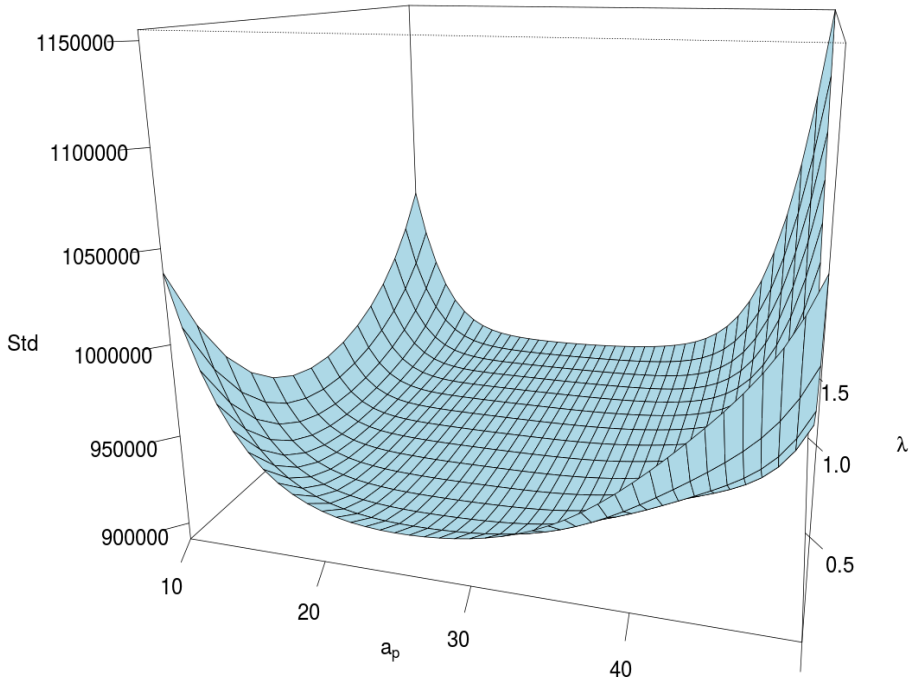
(c) Relative expected improvement after 10 iterations of adaptive sampling.

(d) Relative expected improvement after 20 iterations of adaptive sampling.

Figure 5.2: Relative expected improvement at different stages of the optimization process for $(a_{CTV}, a_{SES}) = (0, 2)$.



(a) Estimated value function.



(b) Estimated uncertainty.

Figure 5.3: Estimated value function $\widehat{v}(\lambda, a_p, a_{CTV} = 0, a_{SES} = 2)$ and corresponding uncertainty.

Table 5.1: The estimated optimal configuration for each feasible value of the failure rate, λ. $\widehat{v}$ is the estimated posterior mean and $\sigma(\widehat{v})$ is the estimated posterior standard deviation in the GP.

a_{CTV}	a_{SES}	a_p	λ	$\widehat{v}$	$\sigma(\widehat{v})$
0	1	12	0.1	87026026	896420
0	1	12	0.2	86726854	893357
0	1	12	0.3	86345384	892446
0	1	12	0.4	85881964	892720
0	1	12	0.5	85337141	893575
0	2	22	0.6	84728132	891470
0	2	22	0.7	84411937	891025
0	2	22	0.8	84087095	890875
0	2	22	0.9	83754600	890833
0	2	22	1.0	83415403	890804
0	2	23	1.1	83076993	890734
0	2	23	1.2	82734091	890733
0	2	23	1.3	82387465	890909
0	2	24	1.4	82039205	891500
0	2	24	1.5	81694414	893044

when we are estimating the scale parameter, τ^2, each point is turned from a normal to a student-t distribution with n degrees of freedom. We have assumed that we have a sufficiently large n so that we may assume that the points are still normally distributed. Since we are making these assumptions, we should check whether they are valid or not. There are several ways of checking for normality. We will do it by looking at how new observations match the proposed model.

Denote a new observation by Y, with cumulative distribution function (cdf) $F_Y(y)$. In our case Y is normally distributed. Define $Z = F_Y(Y)$ and note that Z will take values in the interval $[0, 1]$. The cdf of Z is given by

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(F_Y(Y) \leq z) \\ &= P(Y \leq F_Y^{-1}(z)) = F_Y\left(F_Y^{-1}(z)\right) = z. \end{aligned} \tag{5.1}$$

We recognize $F_Z(z)$ as the cdf of a uniform distribution on $[0, 1]$. Thus, if we collect m new observations, $Y_1, ..., Y_m$, we expect $Z_1, ..., Z_m$ to be uniformly distributed on $[0, 1]$.

A histogram showing the distribution of $Z_1, ..., Z_m$ is shown in Figure 5.4. This is computed by evaluating m new points in NOWIcob, and computing $Z_m = F_{Y_m}(Y_m)$, where F_{Y_m} is given from the fitted GP. We see that there are some spikes for z close to zero and one, and a slightly overweight of observations where $z > 0.5$. These spikes indicate that there are some extreme observations that are either a lot smaller or a lot larger than what the estimated model predicts. This might be observations in areas where we have few observations and thus poor estimates of the mean and variance.

By closer inspection, we observe that the points with $z < 0.05$ are characterized by few

vessels and a high failure rates. In particular, points with $(a_{CTV}, a_{SES}) = (0, 0)$ had z close to zero in all cases as the profit where dramatically overestimated. This is because the GP never learned that zero vessels mean that no repairs will be done if a failure occurs which again is caused by the lack of observations. The lack of observations is because the GP learned that $(a_{CTV}, a_{SES}) = (0, 0)$ is an area of low profits. This is just as we want, in terms of efficiency in the optimization process, but it gives poor predictions in this part of the input space. The points where $z > 0.95$ are characterized by many vessels and personnel and a low failure rate, which is also caused by the lack of observations.

In addition to assuming normality, we also assumed that we would have the same covariance structure over the whole feasible domain. The extreme observations we saw in Figure 5.4 might indicate that this assumption was inadequate. For instance, points with a high failure rate, and few vessels and technicians might be subject to a higher variation in the output. Once a failure occurs, it could take a long time before it gets repaired and failures might pile up so that a lot of income might be lost. The time in which these failures occurs will vary between simulations, and with limited maintenance resources will this yield a high variation in the output. The points with $(a_{CTV}, a_{SES}) = (0, 0)$ illustrates this well. If a failure occurs during the first time step of the simulation, this wind turbine will not contribute to the overall profit. If this failure occurs at the very end of the simulation, the wind turbine will have produced a lot of energy and then contributed a lot to the overall profit. The difference in profit between simulations could therefore be large, which leads to a higher uncertainty in this part of the input space.

Because of these differences in the covariance structures, it could be of interest to partition the input space so that we have the same covariance structure in each partition. This was briefly discussed in Section 2.2.2. We note this as a possible improvement of our approach.

5.2 Estimation of Value of Information

The optimal O&M strategies without information about the failure rates for the distributions for λ under consideration are listed in Table 5.2. The PV and strategies are found by using (3.21). Note that $\widehat{v}$ in (3.21) corresponds to the posterior mean in the GP. We see that the distributions with the highest variance yield the lowest PV and the distributions with lowest variance yields the highest PV. The optimal configurations suggested by the distributions with low variance agrees well with Table 5.1 and Figure 4.3.2. For $\lambda \leq 0.5$ is the optimal configuration $(a_p, a_{CTV}, a_{SES}) = (12, 0, 1)$. Most of the probability mass for these two distributions are for $\lambda \leq 0.5$, so that even though the configurations are sub-optimal for $\lambda > 0.5$, they are not given much weight.

The optimal configurations suggested for the distributions with high variances have both two vessels, since a high λ is now more likely. Again we see that the suggested configurations agree well with Table 5.1 and Figure 4.3.2.

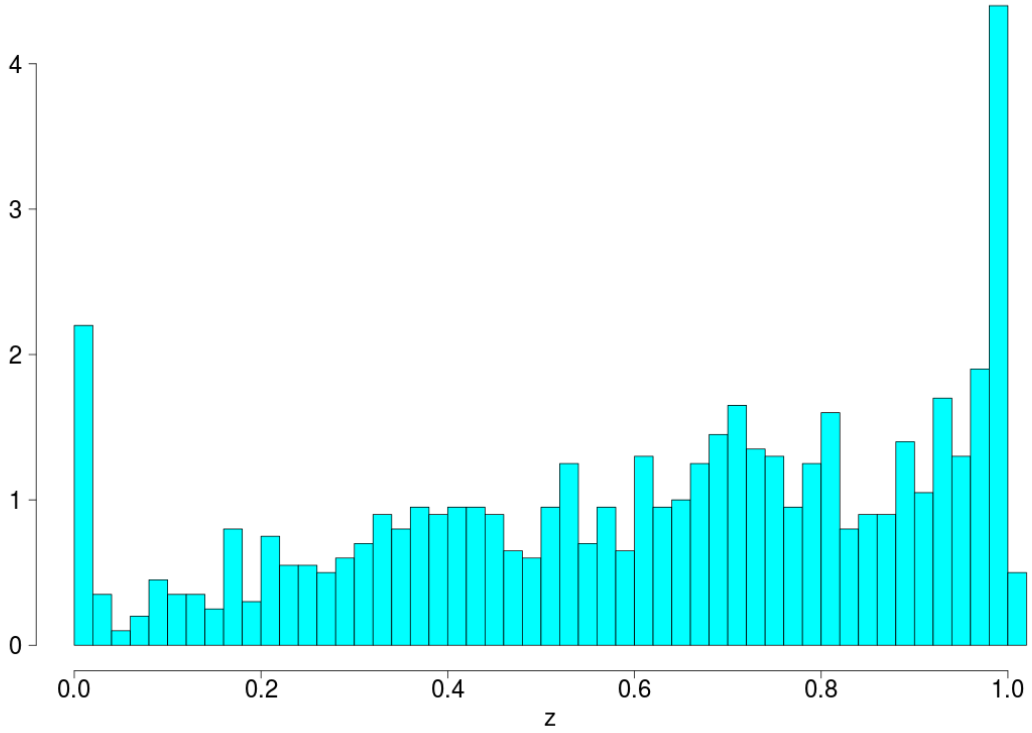


Figure 5.4: Distribution of $z_1, ..., z_{1000}$ when computing the cumulative probability for new observations.

Table 5.2: Summary of optimal O&M strategy without additional knowledge about the failure rate for the four distributions under consideration.

Distribution	a_p	a_{CTV}	a_{SES}	PV
Gamma$(k = 0.903, \beta = 0.316)$	12	0	1	85876527
Gamma$(k = 0.226, \beta = 1.263)$	21	0	2	85153821
Normal$(0.285, 0.3^2)$	12	0	1	85832979
Normal$(0.285, 0.6^2)$	22	0	2	84742937

5.2.1 Perfect Information

The PoV for perfect information is computed by using (3.22). For all λ, are $\max_{\boldsymbol{a} \in \mathcal{A}} \{\widehat{v}(\boldsymbol{\lambda}, \boldsymbol{a}\}$ listed in Table 5.1 and the prior values are listed in Table 5.2. The PoV and VOI for the four distributions under consideration are summarized in Table 5.3. As expected are the VOI largest for the distributions with highest variance, where the gamma distribution yields the overall highest VOI. The estimated VOI is, for all distributions of λ, a good amount of money. It is, therefore, reasonable to assume that information gathering is worthwhile.

Table 5.3: Estimated PoV and VOI for the four distributions of the failure rate under consideration.

Distribution	PoV	VOI
Gamma$(k = 0.903, \beta = 0.316)$	86082262	205735
Gamma$(k = 0.226, \beta = 1.263)$	85679872	526050
Normal$(0.285, 0.3^2)$	85933890	100910
Normal$(0.285, 0.6^2)$	85099062	356125

5.2.2 Imperfect Information

The PoV for imperfect information is computed by using (4.7) and the four proposed distributions for λ. Just as we did in the running example in Chapter 3, we will compute the VOI for three sets of values for p and q. The result in summarized in Table 5.4. For $p = \frac{1}{20}$ and $p = \frac{9}{10}$ are the differences in VOI for imperfect and perfect information small. The differences gets larger when q decreases and p increases, but they are still relatively small. The estimates for VOI are still so high that information gathering is likely to be worthwhile, depending on the price of the information.

Table 5.4: Estimated PoV and VOI for the four distributions of the failure rate under consideration.

p	q	Distribution	PoV	VOI
$\frac{1}{20}$	$\frac{9}{10}$	Gamma$(k = 0.903, \beta = 0.316)$	86081259	204732
		Gamma$(k = 0.226, \beta = 1.263)$	85678795	524973
		Normal$(0.285, 0.3^2)$	85932296	99317
		Normal$(0.285, 0.6^2)$	85097466	354528
$\frac{1}{6}$	$\frac{2}{3}$	Gamma$(k = 0.903, \beta = 0.316)$	86078994	202467
		Gamma$(k = 0.226, \beta = 1.263)$	85676361	522539
		Normal$(0.285, 0.3^2)$	85928972	95992
		Normal$(0.285, 0.6^2)$	85093740	350802
$\frac{1}{3}$	$\frac{1}{3}$	Gamma$(k = 0.903, \beta = 0.316)$	86076561	200034
		Gamma$(k = 0.226, \beta = 1.263)$	85672885	519063
		Normal$(0.285, 0.3^2)$	85925486	92506
		Normal$(0.285, 0.6^2)$	85088417	345480

5.2.3 Accuracy of the VOI Estimate

We will in the following investigate the accuracy and uncertainty of the VOI estimate for one distribution of λ, $p(\lambda) = \text{Gamma}(k = 0.226, \beta = 1.263)$. The evolution of $\widehat{\text{VOI}}$ is shown in Figure 5.5. We observe that the VOI estimate changes a lot early in the optimization process

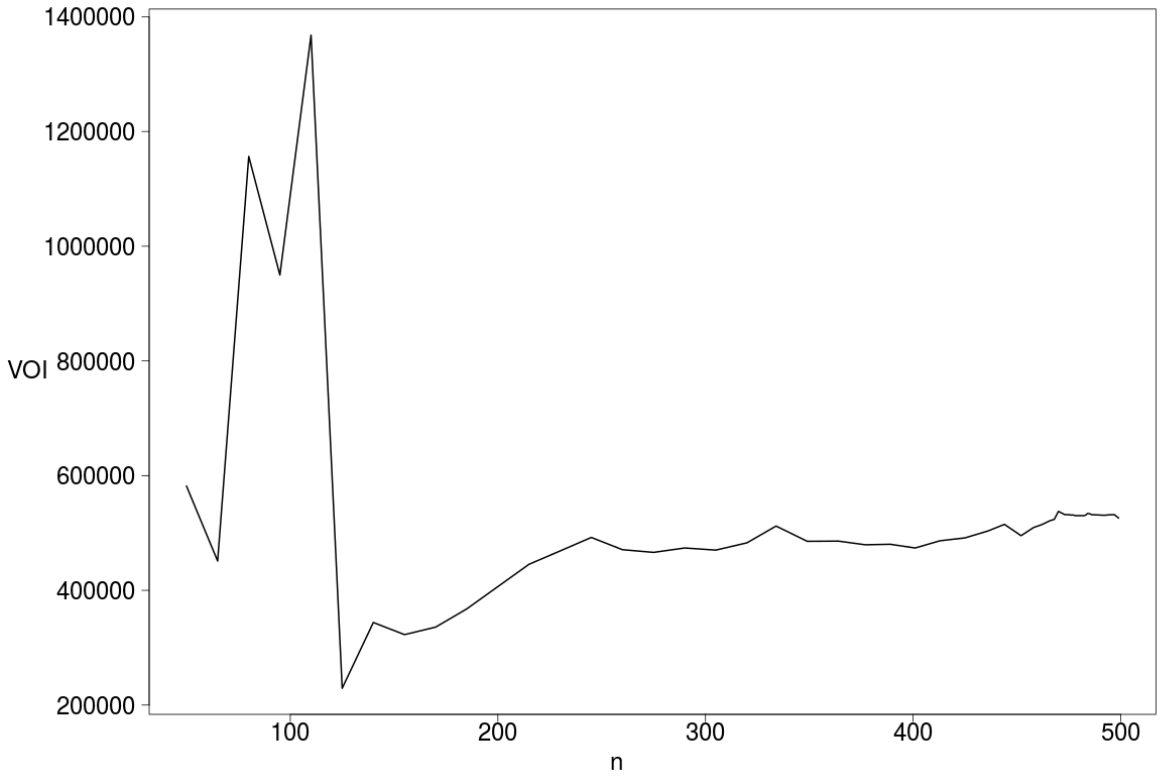


Figure 5.5: The estimated VOI as a function of the number of evaluated points, n.

and stabilizes as the number of evaluated points, n, grows. When n is close to 500, there are very small changes in $\widehat{\mathrm{VOI}}$ which suggest that our estimate for VOI is accurate.

When estimating VOI, we searched after the configuration that resulted in the highest profit for each value of λ. For an efficient search, we designed the acquisition function so that configurations that were expected to yield a high profit were chosen to be evaluated. If the target was to obtain an accurate estimate of VOI, we could instead have designed the acquisition function with this objective in mind. Then the algorithm would proceed until $\widehat{\mathrm{VOI}}$ stabilizes according to some defined measure. The adequacy and development of such an acquisition function are considered as a possible improvement.

We have used (3.25) to draw B realizations from the estimated value function to study sample variation in the process over the conditional mean estimation. For each realization are $\widehat{\mathrm{PV}}$, $\widehat{\mathrm{PoV}}$ and $\widehat{\mathrm{VOI}}$ computed. The estimated mean and standard deviation is found to be

$$\begin{aligned} \overline{\mathrm{VOI}} &= 1339959 \\ \mathrm{SD}(\widehat{\mathrm{VOI}}) &= 218291, \end{aligned} \tag{5.2}$$

by using (3.26) and (3.27), respectively. We note that the mean is much higher than what we estimated. This is because of the bias that we briefly mentioned in Section 3.3.3. Intuitively, it makes sense that $\overline{\mathrm{VOI}} > \widehat{\mathrm{VOI}}$. There are many points that can get a higher sampled value than what we estimated, and thus become the optimal point for a given rate. For instance, the

point $(a_p, a_{CTV}, a_{SES}, \lambda) = (24, 1, 2, 1.5)$ have mean $\mu = 80852277$ and standard deviation $\sigma = 914127$. This point have a $\mathrm{PI} = 0.054$. Thus, when sampling the profit at this point, it is a fair chance that the sample is larger than the optimal configuration in Table 5.1. Since we are drawing in total 15350 points, 1025 points for each rate, we will in most cases get samples that give a larger PoV than what we estimated. This will lead to an overestimation of VOI.

The uncertainty estimate is worth taking into consideration, but it remains only an indication of what the true standard deviation is. Nevertheless, for a decision maker, it will be relevant information when one is to decide whether to buy information or not.

Chapter 6

Closing Remarks

In this thesis, we have used Bayesian optimization with a Gaussian process (GP) surrogate model to estimate the value of information (VOI) in a complex decision situation. The decision situation under consideration was the operation and maintenance (O&M) strategy for an offshore wind farm, and the failure rate for a medium repair failure was the uncertain variable that we want to estimate the value of knowing. By defining the value function as the profit for a given O&M strategy, we can search for favorable strategies by optimizing this function. The results indicated that Bayesian optimization was able to find favorable O&M strategies. The estimated VOI seemed reasonable for the four cases of distributions for the failure rate that were proposed.

6.1 Key Findings

When optimizing the value function, we only provided input-output relations and the feasible range of the variables. Based on this, we were able to find favorable regions in the input space. This suggests that a GP was a suitable surrogate model and that the modified expected improvement acquisition function we used was adequate. Only 499 evaluations in NOWIcob was required before the algorithm terminated. The number of evaluations was reduced by a factor of over 30 compared to the number of feasible points, 15375.

The key results from the estimation of VOI are

- The distributions for λ with the lowest variance resulted in the highest VOI.
- VOI for perfect information was greater than VOI for imperfect information.
- The estimates of VOI suggests that information gathering about the failure rate might be worthwhile, depending on the price of the information.
- The estimate of VOI converged nicely over the number of evaluations.

6.2 Possible Improvements

The proposed distributions for the failure rate was based on literature regarding failure rates for offshore wind turbines. These distributions may be inadequate, so a more appropriate distribution for the failure rate might improve the estimate of VOI.

There are several other uncertain variables involved in the O&M decision problem where the value of information could be estimated. For instance, information about the wave height is valuable since the vessels have an upper limit for wave height that prevents them from operating. One can also include other failure rates, for instance, minor repair failures and look at the VOI for this quantity alone or jointly with the failure rate for medium repair failure.

Some of the assumptions we made when specifying the GP, for instance assuming the same covariance structure over the whole feasible domain, might not be appropriate. By partition the input space into disjoint regions, one might achieve more accurate estimates of the uncertainty.

For a more efficient optimization of the value function, we could let the minimum relative EI, ϵ, be a function of the failure rate. Then we can let points with rates that are unlikely, based on the probability distribution of the failure rate, be subject for less exploration. This will yield a less accurate estimate of the value function for some values of the rate, but it will not affect the overall estimation of VOI significantly.

We only considered a subset of possible decision variables that make up the O&M strategy. A natural extension of this study is therefore to include more decision variables.

Bibliography

Abramowitz, M., Stegun, I. A., 1964. Handbook of mathematical functions: with formulas, graphs, and mathematical tables. Vol. 55. Courier Corporation.

Bratvold, R. B., Bickel, J. E., Lohne, H. P., et al., 2009. Value of information in the oil and gas industry: Past, present, and future. SPE Reservoir Evaluation & Engineering 12 (04), 630–638.

Carroll, J., McDonald, A., McMillan, D., 2016. Failure rate, repair time and unscheduled o&m cost analysis of offshore wind turbines. Wind Energy 19 (6), 1107–1119.

Dinwoodie, I., Endrerud, O.-E. V., Hofmann, M., Martin, R., Sperstad, I. B., 2015. Reference cases for verification of operation and maintenance simulation models for offshore wind farms. Wind Engineering 39 (1), 1–14.

Eidsvik, J., Mukerji, T., Bhattacharjya, D., 2015. Value of information in the earth sciences: Integrating spatial modeling and decision analysis. Cambridge University Press.

Faulstich, S., Hahn, B., Tavner, P. J., 2011. Wind turbine downtime and its importance for offshore deployment. Wind energy 14 (3), 327–337.

Gallala, M. R., 2016. Surrogate-based optimisation using artificial neural networks. Master's thesis, Norwegian University of Technology and Science, Trondheim.

Ginsbourger, D., Le Riche, R., Carraro, L., 2008. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. HAL preprint hal-00260579.

Gramacy, R. B., 2017. A practical introduction to gaussian process regression.
URL `http://bobby.gramacy.com/teaching/gpwebinar/doc.html`

Gramacy, R. B., Lee, H. K. H., 2008. Bayesian treed gaussian process models with an application to computer modeling. Journal of the American Statistical Association 103 (483), 1119–1130.

Gramacy, R. B., Sun, F., 2017. Package 'lagp'.

Gramacy, R. B., Taddy, M., et al., 2010. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models. Journal of Statistical Software 33 (6), 1–48.

Gramacy, R. B., et al., 2007. tgp: an r package for bayesian nonstationary, semiparametric nonlinear regression and design by treed gaussian process models. Journal of Statistical Software 19 (9), 6.

Haraldsdottir, H., Sandstrom, M., 2016. Lifetime analysis of a wind turbine component. Master's thesis, Chalmers University of Technology, Goteborg.
URL http://publications.lib.chalmers.se/records/fulltext/238988/238988.pdf

Hofmann, M., 2011. A review of decision support models for offshore wind farms with an emphasis on operation and maintenance strategies. Wind Engineering 35 (1), 1–15.

Hofmann, M., Sperstad, I. B., Kolstad, M. L., 2017a. Technical documentation of version 3.3 of the nowicob tool. SINTEF Energi. Rapport.

Hofmann, M., Sperstad, I. B., Kolstad, M. L., 2017b. User guide for version 3.3 of the nowicob tool. SINTEF Energi. Rapport.

Jardine, A. K., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical systems and signal processing 20 (7), 1483–1510.

Kahneman, D., Tversky, A., 2013. Prospect theory: An analysis of decision under risk. In: Handbook of the fundamentals of financial decision making: Part I. World Scientific, pp. 99–127.

Katzfuss, M., Guinness, J., 2017. A general framework for vecchia approximations of gaussian processes. arXiv preprint arXiv:1708.06302.

Liu, B., Zhang, Q., Gielen, G. G., 2014. A gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems. IEEE Transactions on Evolutionary Computation 18 (2), 180–192.

Ødegård, H. L., Eidsvik, J., Fleten, S.-E., 2017. Value of information analysis of snow measurements for the scheduling of hydropower production. Energy Systems, 1–19.

Petersen, K. B., Pedersen, M. S., et al., 2008. The matrix cookbook. Technical University of Denmark 7 (15), 510.

Platt, J. C., Burges, C. J., Swenson, S., Weare, C., Zheng, A., 2002. Learning a gaussian process prior for automatically generating music playlists. In: Advances in neural information processing systems. pp. 1425–1432.

Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P. K., 2005. Surrogate-based analysis and optimization. Progress in aerospace sciences 41 (1), 1–28.

Rasmussen, C. E., 2004. Gaussian processes in machine learning. In: Advanced lectures on machine learning. Springer, pp. 63–71.

Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., Wynn, H., 2018. Group kernels for gaussian process metamodels with categorical inputs. arXiv preprint arXiv:1802.02368.

Santner, T. J., Williams, B. J., Notz, W. I., 2013. The design and analysis of computer experiments. Springer Science & Business Media.

Schonlau, M., Welch, W. J., Jones, D. R., 1998. Global versus local search in constrained optimization of computer models. Lecture Notes-Monograph Series 34, 11–25.
URL http://www.jstor.org/stable/4356058

Seyr, H., Muskulus, M., 2016. Value of information of repair times for offshore wind farm maintenance planning. In: Journal of Physics: Conference Series. Vol. 753. IOP Publishing, p. 092009.

Shafiee, M., 2015. Maintenance logistics organization for offshore wind energy: Current progress and future perspectives. Renewable Energy 77, 182–193.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., De Freitas, N., 2016. Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE 104 (1), 148–175.

Smola, A. J., Bartlett, P. L., 2001. Sparse greedy gaussian process regression. In: Advances in neural information processing systems. pp. 619–625.

Snoek, J., Larochelle, H., Adams, R. P., 2012. Practical bayesian optimization of machine learning algorithms. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 2951–2959.
URL http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., Adams, R., 2015. Scalable bayesian optimization using deep neural networks. In: International Conference on Machine Learning. pp. 2171–2180.

Snyman, J. A., 2005. Practical mathematical optimization. Springer.

Sperstad, I. B., McAuliffe, F. D., Kolstad, M., Sjømark, S., 2016. Investigating key decision problems to optimize the operation and maintenance strategy of offshore wind farms. Energy Procedia 94, 261–268.

Welte, T. M., Sperstad, I. B., Halvorsen-Weare, E. E., Netland, Ø., Nonås, L. M., Stålhane, M., 2018. Operation and maintenance modelling. Offshore Wind Energy Technology, 269.

Xiuliang, L., Hongye, S., Jian, C., 2009. Multiple model soft sensor based on affinity propagation, gaussian process and bayesian committee machine. Chinese Journal of Chemical Engineering 17 (1), 95–99.

Appendix

The simulation tool used in this thesis have many input and output parameters that might lead to confusion. Explanation and specification of the most important parameters are therefore provided here. Explanation of the parameters related to failures, wind farm specification, simulation setup, vessels and workforce are given in the tables 6.1, 6.2, 6.3, 6.4, 6.5, respectively. The values of these parameters are specified in the tables 6.6, 6.7, 6.8, 6.9, 6.10, respectively. The descriptions are based on whats provided by Hofmann et al. (2017b). Note that there are two types of crew transfer vessels (CTV), type 1 and 2. We only consider type 1 CTV, but to ease the notation it is simply denoted by CTV.

Table 6.1: Description of some input parameters related to component failures.

Input parameter	Description
Failure type	The different types of failures are categorized based on their severity. We distinguish between manual reset, minor repair, medium repair and annual service.
Annual rate	Each type of failure occur randomly with an annual failure rate.
Costs spare parts	Some failures requires replacement of spare parts, which has an associated cost.
Personnel needed	The number of personnel needed to perform the maintenance task required for a failure.
Active maintenance time	The number of hours required to perform the maintenance task.
Stop at failure	Does the wind turbine stop the energy production because of the failure that occurred?
Stop during repair	Is it required to stop the energy production during the repair?

Table 6.2: Description of some input parameters related to the wind farm.

Input parameter	Description
Turbine type	Each turbine is defined by some properties, such as physical dimensions and cut-in/cut-out wind speeds. Wind speeds outside the cut-in and cut-out interval gives zero power production.
Number of wind turbines	The number of wind turbines in the wind farm under consideration. We will assume that all the wind turbines in the farm are of the same type.
Distance to harbour	The shortest distance, in kilometers, from a wind turbine to the harbour.
Electricity price	The electricity price, measured in GBP per MWh, assumed to be constant throughout the whole simulation period.

Table 6.3: Description of some input parameters related to the simulation setup.

Input parameter	Description
Weather	Time series for wind speed, wind direction, wave height and wave direction. The time series are based on historical data from the wind farm location.
Simulation period	Length of simulation period in years.
Number of MC iterations	The number of Monte Carlo simulations used. For each MC iteration, new weather time series are generated.
Wind speed resolution	The resolution of the wind speed.
Wave height resolution	The resolution of the wave height, in meters.

Table 6.4: Description of some input parameters related to the vessels.

Input parameter	Description
Day rate	Fixed cost per day per vessel in GBP including cost for maintenance of vessel.
Travel speed	The maximum speed, in knots, for the vessels when traveling to/from the wind farm.
Fuel consumption travelling	The amount of fuel, measured in litres per hour, consumed by a vessel.
Personnel space	The maximum number of passengers on a vessel.
Offshore wind speed limit	The upper limit of wind speed, measured in meters per second, so that the vessel still can operate.
Offshore wave height limit	The upper limit of wave height in meters so that the vessel still can operate.

Table 6.5: Description of some input parameters related to the workforce.

Input parameter	Value
Number hours per shift	Scheduled length of a shift.
Number of daily shifts	How many shifts per day.
Shifts start	What time of the day does the shift starts.
Minimum working hours	Minimum length of a shift
Annual salary technicians	The fixed, yearly expense for one technician, in GBP.

Table 6.6: Specification of some input parameters related to component failures. The annual rate for medium repair failures are assumed unknown in this thesis and is therefore denoted by λ in this table.

Failure type	Manual reset	Minor repair	Medium repair	Annual service
Annual rate	7.5	3.6	λ	1
Costs spare parts[GBP]	0	1000	18500	18500
Personnel needed	2	2	3	3
Active maintenance time[hours]	3	7.5	22	60
Stop at failure	Yes	Yes	Yes	No
Stop during repair	Yes	Yes	Yes	Yes

Table 6.7: Specification of some input parameters related to the wind farm.

Input parameter	Value
Turbine type	Vestas V90 3 MW turbine
Number of wind turbines	80
Distance to harbour [km]	50
Electricity price [GBP/MWh]	90

Table 6.8: Specification of some input parameters related to the simulation setup.

Input parameter	Value
Weather	FINO_data
Simulation period [years]	1
Number of MC iterations	5
Wind speed resolution [m/s]	1.0
Wave height resolution [m]	0.1

Table 6.9: Specification of some input parameters related to the vessels.

Input parameter	CTV	SES
Day rate [GBP]	1750	5000
Travel speed [knots]	20	35
Fuel consumption travelling [l/hour]	400	700
Personnel space	12	12
Offshore wind speed limit [m/s]	30	30
Offshore wave height limit [m]	1.5	2

Table 6.10: Specification of some input parameters related to the workforce.

Input parameter	Value
Number hours per shift	12
Number of daily shifts	1
Shifts start	06:00
Minimum working hours	1
Annual salary technicians [GBP]	80000