THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

Victor TRAPPLER

Thèse dirigée par **Arthur VIDARD**, Université Grenoble Alpes
et codirigée par **Elise ARNAUD**
et **Laurent DEBREU**, Chargé de recherche , INRIA

préparée au sein du **Laboratoire Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Contrôle de paramètre en présence d'incertitudes

Parameter control in the presence of uncertainties

Thèse soutenue publiquement le ,
devant le jury composé de :

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Numerical models are widely used to study or forecast natural phenomena and improve industrial processes. However, by essence models only partially represent reality and sources of uncertainties are ubiquitous (discretisation errors, missing physical processes, poorly known boundary conditions). Moreover, such uncertainties may be of different nature. [WHR+03] proposes to consider two categories of uncertainties:

- Aleatoric uncertainties, coming from the inherent variability of a phenomenon, *e.g.* intrinsic randomness of some environmental variables
- Epistemic uncertainties coming from a lack of knowledge about the properties and conditions of the phenomenon underlying the behaviour of the system under study

The latter can be accounted for through the introduction of ad-hoc correcting terms in the numerical model, that need to be properly estimated. Thus, reducing the epistemic uncertainty can be done through parameters estimation approaches. This is usually done using optimal control techniques, leading to an optimisation of a well chosen cost function which is typically built as a comparison with reference observations. An application of such an approach, in the context of ocean circulation modeling, is the estimation of ocean bottom friction parameters in [DL91] and [Bou15].

If parameters to be estimated are not the only source of uncertainties, their optimal control is doomed to overfit the data, *e.g* to artificially introduce errors in the controlled parameter to compensate for other sources. If such uncertainties are of aleatoric nature, then the parameter estimation is only optimal for the observed situation, and may be very poor in other configurations, phenomenon coined as *localized optimisation* in [HB01].

The calibration often takes the form of the minimisation of a function J, that describes a distance between the output of the numerical model and some given observed data, plus generally some regularization terms. In our study, this cost function takes two types of arguments: $k \in \mathbb{K}$ that represents the parameters to calibrate, and $u \in \mathbb{U}$, that represents the environmental conditions. We assume that the environmental conditions are uncertain by nature, and thus will be modelled with a random variable U, to account

for these aleatoric uncertainties. This is then the random variable $J(k, U)$ that we want to minimize "in some sense" with respect to k.

Some of the optimisation under uncertainties methods rely on the optimisation of the moments of $k \mapsto J(k, U)$ (in [LSN04, JLR10]), while other methods are based on multiobjective problems, such as in [Bau12, Rib18]. These approaches may compensate some bad performances by some very good ones, as we are averaging over $\mathbb{U}$.

We propose to compare the value of the objective function to the best value attainable given the environmemtal conditions at this point, with the idea that we want to be as close as possible, and as often as possible, to this optimal value. Introducing the relative regret, that is the ratio of the objective function by its conditional optimum, we can define a new family of robust estimators.

Within this family, choosing an estimator consists in favouring either its robustness, *e.g* its ability to perform well under all circumstances, or on the contrary favour near-optimal performances, transcribing a risk-averse or a risk-seeking behaviour from the user.

CHAPTER 1

INVERSE PROBLEM AND CALIBRATION

Contents

1.1 Introduction

In this chapter we will first lay the ground for developing the general ideas behind calibration, by introducing the notions of models, and forward and inverse problems in Section 1.2. This implies also a short review of notions of probability theory. Calibration will be defined in Section 1.3 as the optimization of an objective function: Maximum likelihood estimation in a frequentist setting, or posterior maximization using Bayes' theorem. In practice, for large-scale applications, the optimization is performed using gradient-descent, and the computational cost of gradient computation can be overcome by adjoint method, as described in Section 1.4. Finally, we are going to discuss two aspects related to calibration, namely model selection in Section 1.5 and the influence of nuisance parameters and model misspecification in calibration in Section 1.6.

1.2 Forward, inverse problems and probability theory

Running a simulation using numerical tools is useful to grasp a better understanding of the physical phenomena, or to forecast them. On the other hand, when observing and comparing the measurements and the output of the numerical simulation, we can quantify the mismatch between the two and tune some parameters involved in the computations. Indeed, these parameters represent different physical quantities or processes that are for example unresolved at the model's scale (friction of the ocean bed, or $k - \epsilon$ turbulence models for instance), or ill-known. A proper estimation of these parameters has to be performed in order to guarantee a meaningful output when evaluating the model.

Model calibration or parameter estimation has been widely treated in the literature, either from a statistical and probabilistic point of view using likelihood-based methods and Bayesian inference, or from a *variational* point of view by defining proper objective functions. To match those two approaches, we will first review the problem from a probabilistic point of view, in order to define properly some appropriate objective functions and introduce tools from optimal control theory to optimize them.

1.2.1 Model space data space and forward problem

In order to describe accurately a physical system, we have to define the notion of models, and will be following [Tar05] approach to define inverse problems. A model represents the link between some parameters and some observable quantities. A simple example is a model that takes the form of a system of ODEs or PDEs, maybe discretized, while the parameters are the initial conditions and the output is one or several time series, describing the time evolution of a quantity at one or several spatial points. An important point is that a model is not only the *forward operator*, but must also include the parameter space.

Definition 1.2.1 – Model: A model $\mathfrak{M}$ is defined as a pair composed of a *forward operator* $\mathcal{M}$, and a *parameter space* Θ

$$\mathfrak{M} = (\mathcal{M}, \Theta) \tag{1.1}$$

The forward operator is the mathematical representation of the physical system, while the parameter space is chosen here to be a subset of a finite dimensional space, so usually Θ will be a subset of $\mathbb{R}^n$.

As we will usually choose Θ as a subset of $\mathbb{R}^n$, for $n \geq 1$, we can define the dimensionality of the model, based on the number of *degrees of freedom* available for the parameters.

Remark 1.2.2: The dimension of a model $\mathfrak{M} = (\mathcal{M}, \Theta)$ is the number of parameters not reduced to a singleton, so if $\Theta \subset \mathbb{R}^n$, the dimension of $\mathfrak{M}$ is $d \leq n$. The dimension of a model $\mathfrak{M}$ is sometimes called the degrees of freedom of $\mathfrak{M}$.

Example 1.2.3: A model with parameter space $\Theta = \mathbb{R}^2 \times [0, 1]$ has dimension 3, while $\Theta = \mathbb{R}^2 \times \{1\}$ has dimension 2.

Now that we have introduced the forward operator and the parameter space, we will focus on the output of the model. Ideally, the data space $\mathbb{Y}$ consists in all the physically acceptable results of the physical experiment. Then, the forward operator $\mathcal{M}$ maps the parameter space $\Theta \subset \mathbb{R}^n$ to the data space $\mathbb{Y}$, as one can expect that all models provide physically acceptable outputs.

1.2.2 Forward problem

Given a model $(\mathcal{M}, \Theta)$, the *forward problem* consists in applying the forward operator to a given $\theta \in \Theta$, in order to get the *model prediction*. The forward problem is then to obtain information on the result of the experiment based on the parameters we chose as input, so deriving a satisfying forward operator $\mathcal{M}$.

$$\begin{array}{rccc} \mathcal{M}: & \Theta & \longrightarrow & \mathbb{Y} \\ & \theta & \longmapsto & \mathcal{M}(\theta) \end{array} \tag{1.2}$$

As said earlier, the forward operator can be a set of ODEs or PDEs, discretized or not. The forward problem is then the attempt to link the causes, i.e. the parameters, to the consequences, i.e. the output in the data space.

1.2.3 Inverse Problem

The inverse problem is the counterpart of the forward problem, and consists in trying to gather more information on the parameters, based on: the result of the experiment or the observation of the physical process and on the knowledge of the forward operator, as illustrated Fig. 1.1.

This is done by directly comparing the output of the forward operator, and trying to reduce the mismatch between the observed data and the model prediction.

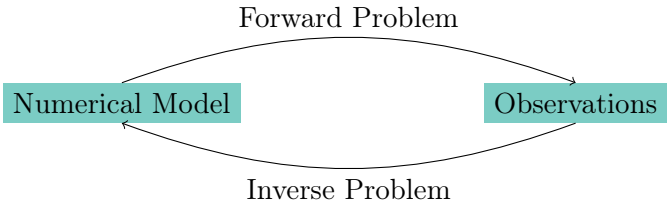

Figure 1.1: Forward and Inverse problem diagram

However, a purely deterministic approach for the inverse problem is doomed to underperform: as most physical processes are not perfectly known, some uncertainties remain in the whole modelling process. Those uncertainties are ubiquitous: the observations available may be corrupted by a random noise coming from the measurement devices and the model may not represent perfectly the reality, thus introducing a systematic bias for instance. Taking into account those uncertainties is crucial to solve the inverse problem.

In that perspective we are going to introduce briefly the usual probabilistic framework, along with common notations that we will use throughout this manuscript. Those notions are well established in the scientific literature, and one can read [Bil08] for a more thorough description.

1.2.4 Notions of probability theory

1.2.4.a Probability measure, and random variables

We are first going through some usual notions of probability theory. Let us consider the usual probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.2.4 – Event probability and conditioning: We call an event an element of the σ-algebra $\mathcal{F}$, and the probability of an event $A \in \mathcal{F}$ is defined as the Lebesgue integral

$$\mathbb{P}[A] = \int_A \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[\{\omega; \omega \in A\}] \tag{1.3}$$

Observing an event $B \in \mathcal{F}$ can bring information upon another event $A \in \mathcal{F}$. In that sense, we introduce the conditional probability of A given B. Let A, $B \in \mathcal{F}$. The event A given B is written $A \mid B$ and its probability is

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \tag{1.4}$$

Formally, an event can be seen as an outcome of some uncertain experiment, and its probability is "how likely" this event will happen.

Let us now introduce a measurable state (or sample) space S, that is the set of all possible events (upon which we can assign a probability).

Definition 1.2.5 – Random Variable, Expectation: A random variable (abbreviated as r.v.) X is a measurable function from $\Omega \longrightarrow S$. A random variable will usually be written with an upper case letter. A realisation or observation x of the r.v. X is the actual image of $\omega \in \Omega$ under X: $x = X(\omega)$. If S is countable, the random variable is said to be *discrete*. When $S \subseteq \mathbb{R}^p$ for $p \geq 1$, X is sometimes called a random vector

The expectation of a r.v. $X : \Omega \to S$ is defined as

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbb{P}(\omega) \tag{1.5}$$

Using the Definition 1.2.5, the probability of an event A can be seen as the expectation of the indicator function of A:

$$\begin{array}{rccc} \mathbb{1}_A : & \Omega & \longrightarrow & \{0,1\} \\ & \omega & \longmapsto & \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A \end{cases} \end{array} \tag{1.6}$$

and it follows that

$$\mathbb{E}[\mathbb{1}_A] = \int_{\Omega} \mathbb{1}_A \, \mathrm{d}\mathbb{P}(\omega) = \int_A \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[A] \tag{1.7}$$

As we defined the notion of a r.v. in Definition 1.2.5 as a measurable function from $\Omega \to S$, we can now focus on the measurable sets through X, by using in a sense the change of variable $x = X(\omega)$.

Definition 1.2.6 – Image (Pushforward) measure: Let $X : \Omega \to S$ be a random variable, and $A \subseteq S$. The image measure (also called pushforward measure) of $\mathbb{P}$ through X is denoted by $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$. This notation can differ slightly depending on the community, so one can find also $\mathbb{P}_X = \mathbb{P} \circ X^{-1} = X_\sharp \mathbb{P}$, the latter notation being used in transport theory. The probability, for the r.v. X to be in A is equal to

$$\mathbb{P}[X \in A] = \mathbb{P}_X[A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_{X^{-1}(A)} \mathrm{d}\mathbb{P}(\omega) = \mathbb{P}[X^{-1}(A)] = \mathbb{P}\left[\{\omega\,;\, X(\omega) \in A\}\right] \tag{1.8}$$

Similarly, for any measurable function h, the expectation taken with respect to a specific random variable X is

$$\mathbb{E}_X[h(X)] = \int_{\Omega} h(X(\omega))\, \mathrm{d}\mathbb{P}_X(\omega) \tag{1.9}$$

In most of this thesis, the sample space will be $S \subseteq \mathbb{R}^p$ for $p \geq 1$, so we are going to introduce useful tools and notations to characterize these particular real random variables.

1.2.4.b Real-valued random variables

We are now going to focus on real-valued random variables, so measurable function from Ω to the sample space $S = \mathbb{R}$.

Definition 1.2.7 – Distribution of a real-valued r.v.: The distribution of a r.v. can be characterized by a few functions:

- The *cumulative distribution function* (further abbreviated as cdf) of a real-valued r.v. X is defined as:

$$F_X(x) = \mathbb{P}\left[X \leq x\right] = \mathbb{P}_X\left[\,]-\infty, x]\,\right] \tag{1.10}$$

  and $\lim_{-\infty} F_X = 0$ and $\lim_{+\infty} F_X = 1$ If the cdf of a random variable is continuous, the r.v. is said to be *continuous* as well.

- The *quantile function* Q_X is the generalized inverse function of the cdf:

$$Q_X(p) = \inf\{q : F_X(q) \geq p\} \tag{1.11}$$

- If there exists a function $f : S \to \mathbb{R}^+$ such that for all measurable sets A

$$\mathbb{P}[X \in A] = \int_A \mathrm{d}\mathbb{P}_X(\omega) = \int_A f(x)\,\mathrm{d}x \tag{1.12}$$

  then f is called the *probability density function* (abbreviated pdf), or *density* of X and is denoted p_X. As $\mathbb{P}[X \in S] = 1$, it follows trivially that $\int_S f(x)\,\mathrm{d}x = 1$. One can verify that if F_X is derivable, then its derivative is the density of the r.v. :

$$\frac{\mathrm{d}F_X}{\mathrm{d}x}(x) = p_X(x) \tag{1.13}$$

Remark 1.2.8: When restricting this search to "classical" functions, p_X may not exist. However, allowing generalized functions such as the *dirac delta function*, provides a way to consider simultaneously all types of real-valued random variables (continous, discrete, and mixture of both). Dirac's delta function can (in)formally be defined as

$$\delta_{x_0}(x) = \begin{cases} +\infty \text{ if } x = x_0 \\ 0 \text{ elsewhere} \end{cases} \qquad \text{and} \quad \int_S \delta_{x_0}(x)\,\mathrm{d}x = 1 \tag{1.14}$$

Example 1.2.9: Let us consider the random variable X that takes the value 1 with probability 0.5, and follows a uniform distribution with probability 0.5 over $[2; 4]$. Its cdf can be expressed as

$$F_X(x) = \begin{cases} 0 \text{ if } x < 1 \\ 0.5 \text{ if } 1 \leq x < 2 \\ 0.5 + \frac{x-2}{8} \text{ if } 2 \leq x < 4 \\ 1 \text{ if } 4 \leq x \end{cases} \tag{1.15}$$

and its pdf (as a generalized function)

$$p_X(x) = \frac{1}{2}\delta_1(x) + \frac{1}{4}\mathbb{1}_{\{2\leq x<4\}}(x) \tag{1.16}$$

The pdf and cdf are shown Fig. 1.2.

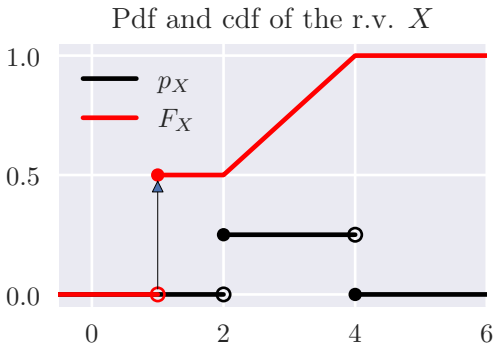

Figure 1.2: Cdf and Pdf of X defined in Example 1.2.9. The arrow indicates Dirac's delta function

Definition 1.2.10 – Moments of a r.v. and L^s spaces: Let X be a random variable. The moment of order s is defined as $\mathbb{E}[X^s]$, and the centered moment of order s is defined as

$$\mathbb{E}[(X - \mathbb{E}[X])^s] = \int (X(\omega) - \mathbb{E}[X])^s \, \mathrm{d}\mathbb{P}(\omega) = \int (x - \mathbb{E}[X])^s \cdot p_X(x) \, \mathrm{d}x \tag{1.17}$$

To ensure that those moments exists, let us define $L^s(\mathbb{P})$ as the space of random variables X such that $\mathbb{E}[|X|^s] < +\infty$. If $X \in L^2(\mathbb{P})$, the centered moment of order 2 is called the variance:

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{V}\mathrm{ar}[X] \geq 0 \tag{1.18}$$

These definition above hold for real-valued random variables, so 1D r.v., but can be extended for random vectors.

1.2.4.c Real-valued random vectors

Most of the definitions for a random variable extend component-wise to random vectors:

Definition 1.2.11 – Joint, marginal and conditional densities: Let $X = [X_1, \cdots, X_p]$ be a random vector from $\Omega \to S \subseteq \mathbb{R}^p$ The expected value of a random vector is the expectation of the components

$$\mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p]] \tag{1.19}$$

The cdf of X at the point $x = [x_1, \dots x_p]$ is

$$F_X(x) = F_{X_1,\dots,X_p}(x_1, \dots, x_p) = \mathbb{P}\left[X_1 \leq x_1, \cdots, X_p \leq x_p\right] = \mathbb{P}\left[\bigcap_{i=1}^{p} \{\omega;\ X_i(\omega) \leq x_i\}\right] \tag{1.20}$$

Similarly as in the real-valued case, we can define the pdf of the random vector, or *joint pdf* by derivating with respect to the variables:

$$p_X(x) = p_{X_1,\dots,X_p}(x_1, \dots, x_p) = \frac{\partial^p F_X}{\partial x_1 \cdots \partial x_p}(x) \tag{1.21}$$

and $\int_S p_{X_1,\dots,X_p}(x_1, \dots, x_p)\, \mathrm{d}(x_1, \dots, x_p) = 1$

For two random vectors X and Y, the (cross-)covariance matrix of X and Y is defined as

$$\mathbb{C}\mathrm{ov}\left[X, Y\right] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T \tag{1.22}$$

and based on this definition, we can extend the notion of variance to vectors. The covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of X, is defined to be

$$\Sigma = \mathbb{C}\mathrm{ov}(X) = \mathbb{C}\mathrm{ov}[X, X] = \mathbb{E}\left[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^T\right] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T \tag{1.23}$$

We can now define the *marginal densities*. For notation clarity, we are going to set $X = [Y, Z]$: the marginal densities of Y and Z are

$$p_Y(y) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}z \quad \text{and} \quad p_Z(z) = \int_{\mathbb{R}} p_{Y,Z}(y, z)\, \mathrm{d}y \tag{1.24}$$

The random variable Y given Z, denoted by $Y \mid Z$ has the conditional density

$$p_{Y|Z}(y \mid z) = \frac{p_{Y,Z}(y, z)}{p_Z(z)} \tag{1.25}$$

allowing us to rewrite the marginals as

$$p_Y(y) = \int_{\mathbb{R}} p_{Y|Z}(y \mid z) p_Z(z)\, \mathrm{d}z = \mathbb{E}_Z\left[p_{Y|Z}(y \mid z)\right] \tag{1.26}$$

$$p_Z(z) = \int_{\mathbb{R}} p_{Z|Y}(z \mid y) p_Y(y)\, \mathrm{d}y = \mathbb{E}_Y\left[p_{Z|Y}(z \mid y)\right] \tag{1.27}$$

1.2.4.d Bayes' Theorem

The classical Bayes' theorem is directly a consequence of the definition of the conditional probabilities in Definition 1.2.4, and for random variables admitting a density in Definition 1.2.11.

Theorem 1.2.12 – Bayes' theorem: Let $A, B \in \mathcal{F}$. Bayes' theorem states that

$$\mathbb{P}[A \mid B] \cdot \mathbb{P}[B] = \mathbb{P}[B \mid A] \cdot \mathbb{P}[A]$$
$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]} \text{ if } \mathbb{P}[B] \neq 0$$

In terms of densities, the formulation is sensibly the same. Let Y and Z be two random variables. The conditional density of Y given Z can be expressed using the conditional density of Z given Y.

$$p_{Y|Z}(y \mid z) = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{p_Z(z)} = \frac{p_{Z|Y}(z \mid y)p_Y(y)}{\int p_{Z,Y}(z, y)\,\mathrm{d}y} \propto p_{Z|Y}(z \mid y)p_Y(y) \tag{1.28}$$

Bayes' theorem is central as it links in a simple way conditional densities. In the inverse problem framework, if Y represents the state of information on the parameter space, while Z represents the information on the data space, $Z \mid Y$ can be seen as the forward problem. Bayes' theorem allow us to "swap" the conditioning, and get information on $Y \mid Z$, that can be seen as the inverse problem.

The influence of one (or a set of) random variable(s) over another can be measured with the conditional probabilities. Indeed, if the state of information on a random variable does not change when observing another one, the observed one carries no information on the other. This notion of dependence (and independence) is first defined on events and extended to random variables

Definition 1.2.13 – Independence: Let $A, B \in \mathcal{F}$. Those two events are said independent if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Quite similarly, two real-valued random variables Y and Z are said to be independent if $F_{Y,Z}(y, z) = F_Y(y)F_Z(z)$ or equivalently, $p_{Y,Z}(y, z) = p_Y(y)p_Z(z)$. Speaking in terms of conditional probabilities, this can be written as $p_{Y|Z}(y, z) = p_Y(y)$. If Y and Z are independent, $\mathbb{C}\mathrm{ov}[Y, Z] = 0$. The converse if false in general.

We discussed so far the different quantities that characterize random variables. Let us consider now two random variables which share the same sample space: $X, X' : \Omega \to S$. There exists various way to compare those two random variables, usually by quantifying some measure of distance between their pdf when they exist. One of the most used comparison tool for random variables is the Kullback-Leibler divergence.

Definition 1.2.14 – KL–divergence and entropy: The Kullback-Leibler divergence, introduced in [KL51] is a measure of dissimilarity between two distributions, based on information-theoretic considerations. Let X, X' be r.v. with the same sample space S, and p_X and $p_{X'}$ their densities, such that $\forall A \in \mathcal{F}$, $\int_A p_X(x)\,\mathrm{d}x = 0 \implies$

$\int_A p_{X'}(x)\,\mathrm{d}x = 0$. The KL-divergence is defined as

$$D_{\mathrm{KL}}\left(p_X \| p_{X'}\right) = \int_S p_X(x) \log \frac{p_X(x)}{p_{X'}(x)}\,\mathrm{d}x \tag{1.29}$$
$$= \mathbb{E}_X\left[-\log p_{X'}(X)\right] - \mathbb{E}_X\left[-\log p_X(X)\right] \tag{1.30}$$
$$= H\left[X', X\right] - H\left[X\right] \tag{1.31}$$

$H[X]$ is called the (differential) entropy of the random variable X, and $H[X', X]$ the cross-entropy of X' and X. Using Jensen's inequality, one can show that for all X and X' such that the KL-divergence exists, $D_{\mathrm{KL}}\left(p_X \| p_{X'}\right) \geq 0$ with equality iff they have the same distribution, a desirable property when measuring dissimilarity. However, the KL–divergence is not a distance function, as it is not symmetric in general, and it does not verify the triangle inequality.

1.2.4.e Important examples of real random variables

One of the most well known distribution is the normal distribution, also called Gaussian distribution, that appears in various situations, but most notably in the central limit theorem.

Example 1.2.15 – The Normal Distribution: Let X be a r.v. from Ω to $\mathbb{R}$. If X follow the normal distribution of mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, we write $X \sim \mathcal{N}(\mu, \sigma^2)$, and its pdf is

$$p_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{1.32}$$

For the multidimensional case, let X be a r.v. from Ω to $\mathbb{R}^p$, that follows a normal distribution of mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, where Σ is semi-definite positive. In that case, $X \sim \mathcal{N}(\mu, \Sigma)$ the density of the random vector X can be written as

$$p_X(x) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-1} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{1.33}$$

where $|\Sigma|$ is the determinant of the matrix Σ, and $(\cdot)^T$ is the transposition operator. As the covariance matrix appears through its inverse, another encountered parametrization is to use the precision matrix Σ^{-1}. Examples of pdf of Gaussian normal distributions are displayed Fig. 1.3.

When adding independent squared samples of the normal distribution, the resulting random variable follows a χ^2 distribution.

Example 1.2.16 – The χ^2 distribution: Let $X_1, X_2, \ldots, X_\nu$ be ν independent random variables, such that for $1 \leq i \leq \nu$, $X_i \sim \mathcal{N}(0, 1)$ We define the random

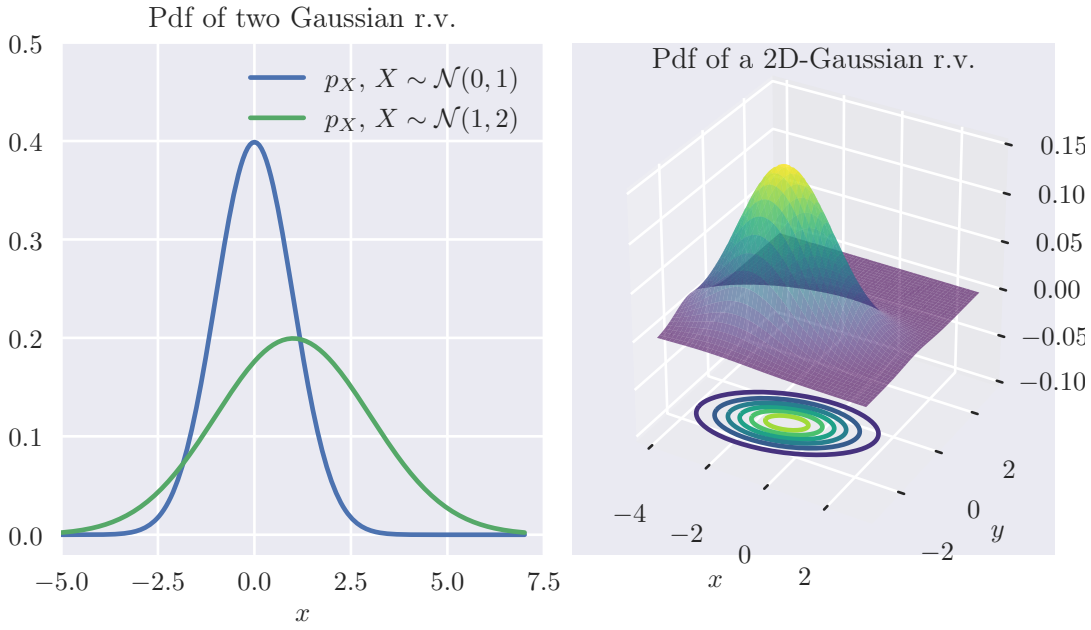

Figure 1.3: Probability Density functions of 1D Gaussian distributed r.v. (left), and density of a 2D Gaussian r.v. (right)

Revoir figure de droite

variable X as

$$X = \sum_{i=1}^{\nu} X_i^2 \tag{1.34}$$

By definition, the random variable X follows a χ^2 distribution with ν degrees of freedom: $X \sim \chi^2_\nu$. The quantile of order β is written $\chi^2_\nu(\beta)$ and verifies

$$\mathbb{P}[X \leq \chi^2_\nu(\beta)] = \beta \tag{1.35}$$

The pdf of such a r.v. is displayed Fig. 1.4, for different degrees of freedom.

1.3 Parameter inference

1.3.1 From the physical experiment to the model

We can represent both the reality and the computer simulation as models. The physical system (the reality) that is observed can be represented by a model $\mathfrak{M} = (\mathscr{M}, \Theta_{\text{real}})$, so by a forward operator $\mathscr{M}$, and a parameter space Θ_{real}. Observing the physical system means to get access to $y \in \mathbb{Y}$ that is the image of an *unknown* parameter value $\vartheta \in \Theta_{\text{real}}$ through the forward operator, so $y = \mathscr{M}(\vartheta) \in \mathbb{Y} \subseteq \mathbb{R}^p$.

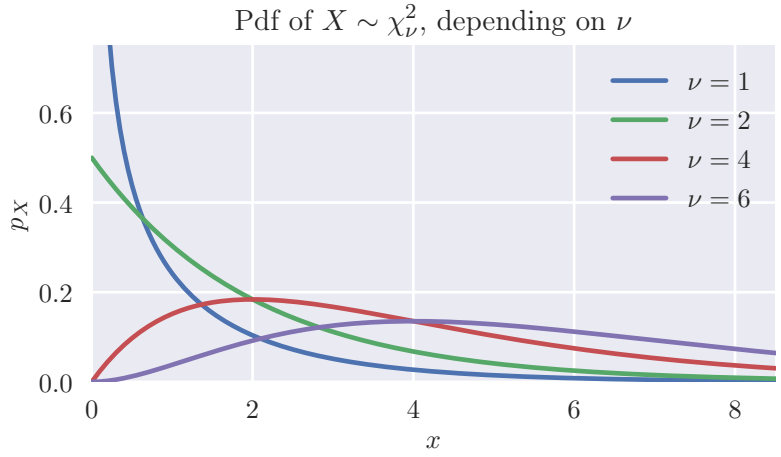

Figure 1.4: Probability density functions of χ^2_ν random variables, for different degrees of freedom

On the other hand, let us assume that a numerical model of the reality has been constructed, by successive various assumptions, discretizations and simplifications giving $(\mathcal{M}, \Theta)$. The main objective of calibration is to find $\hat{\theta}$ such that the forward operator applied to $\hat{\theta}$: $\mathcal{M}(\hat{\theta})$ represents as accurately as possible the physical system, and thus matches as closely the data $\mathscr{M}(\vartheta) = y$. This is illustrated Fig. 1.5.

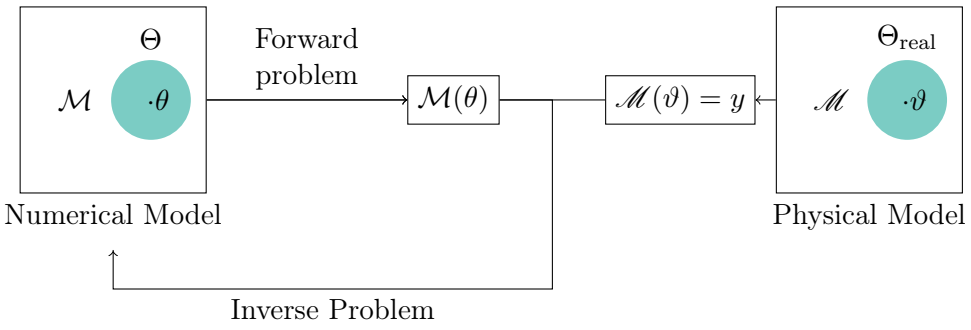

Figure 1.5: Forward and inverse problem using models as defined Definition 1.2.1

First, let us assume that $\vartheta \in \Theta \subseteq \Theta_{\text{real}}$. In [KO01, HKC+04], the authors rewrite the link between the reality and the model at this value as

$$\mathscr{M}(\vartheta) = \mathcal{M}(\vartheta) + \epsilon(\vartheta) \in \mathbb{Y} \subseteq \mathbb{R}^p \tag{1.36}$$

The difference $\epsilon(\vartheta) = \mathscr{M}(\vartheta) - \mathcal{M}(\vartheta)$ is the error between the physical model and the model, called sometimes the misfit, or the residual error. This error is unknown and encompasses different sources of uncertainties, such as measurement errors, or model bias (with respect to the reality). To deal with this unknown, we are going to model it as a sample of a random variable, leading us to treat the obtained data as a random sample as well.

From the diverse assumptions we can make upon this sampled random variable, we can then treat the calibration procedure as a parameter estimation problem of a random variable. The estimated parameter will be written $\hat{\theta}$, and the subscript will denote additional information on the estimator. In this thesis, we focus on extremum estimators. Those estimators are defined as the optimizer of a given objective function J, $\hat{\theta} = \arg\min J$. In the next sections, we will see the probabilistic origins of a few classical objective functions.

1.3.2 Frequentist inference, MLE

1.3.2.a Formulation of the MLE

As mentioned before, we can model the observations as a random variable, say Y (uppercase to highlight its random nature), and assume that this r.v. belongs to a family of parametric distributions, whose densities are

$$\{y \mapsto p_Y(y;\theta); \theta \in \Theta\} \tag{1.37}$$

The choice has been made to keep explicit the dependency on θ. For instance, we can use the hypothesis that the residual are normally distributed with a given covariance matrix Σ. As we assume that $\mathbb{Y} \subseteq \mathbb{R}^p$, Y is a random vector distributed as

$$Y \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.38}$$

whose one sample is $y = \mathscr{M}(\vartheta)$.

Now, instead of looking at the densities of Eq. (1.37) as functions taking as arguments the samples in $\mathbb{Y}$, we may look at it as a function of θ, as the observations $y \in \mathbb{Y}$ do not vary. We can then define the likelihood function and its associated extremum estimator.

Definition 1.3.1 – Likelihood function, MLE: The probability density function of the observations for a set of parameters is called the likelihood of those parameters given the observations, and is written $\mathcal{L}$:

$$\mathcal{L}(\cdot; y) : \theta \mapsto p_Y(y;\theta) = \mathcal{L}(\theta; y) \tag{1.39}$$

$$= (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y)\right) \tag{1.40}$$

If $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, the likelihood can be written as the product of 1D Gaussians:

$$\mathcal{L}(\theta; y) = \left(\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i}\right) \exp\left(\sum_{i=1}^{p} -\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.41}$$

$$= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(\mathcal{M}(\theta)_i - y_i)^2}{2\sigma_i^2}\right) \tag{1.42}$$

with $y = [y_1, \dots, y_p]$ and $\mathcal{M}(\theta) = [\mathcal{M}(\theta)_1, \dots \mathcal{M}(\theta)_p]$. Based on the likelihood function, we can define the *Maximum Likelihood Estimator*, or *MLE*, that maximizes the likelihood defined above:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y) \tag{1.43}$$

For practical and numerical reasons, the maximization of the likelihood is often replaced by the minimization of the negative log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} -\log \mathcal{L}(\theta; y) = \arg\min_{\theta \in \Theta} -\sum_{i=1}^{p} \log p_{Y_i|\theta}(y_i \mid \theta) \tag{1.44}$$

where

$$-\log \mathcal{L}(\theta; y) = \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \tag{1.45}$$

As the optimization is performed on θ, we can remove the constant terms of the objective function, and rewrite the cost function as a L^2 norm in Eq. (1.46).

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \arg\min_{\theta \in \Theta} \frac{1}{2}(\mathcal{M}(\theta) - y)^T \Sigma^{-1} (\mathcal{M}(\theta) - y) \\ &= \arg\min_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}} \end{aligned} \tag{1.46}$$

Frequentist inference and Maximum Likelihood estimation boils down to Generalized non-linear least-square regression, that minimizes the squared Mahalabonis distance between $\mathcal{M}(\theta)$ and y. This is only true as we assumed a Gaussian form of the errors in Eq. (1.38). Other choices of the sampling distribution Eq. (1.38) will result in different objective functions. To reduce the sensitivity on outliers, some authors such as [RSNN15] introduce Student or Laplace distributed errors, or specifically designed norm such as the Huber norm [Hub11].

If the covariance matrix is diagonal, the residual errors are then uncorrelated, thus indepedent due to their Gaussian nature as defined in Eq. (1.38). The likelihood can be rewritten as the product of densities evaluated at the different samples y_i, obtained from their true distribution Y. A direct link can be written between the KL–divergence and the MLE: The KL–divergence between the true density p_Y and the parametric sampling distribution $p_Y(\cdot; \theta)$ is

$$D_{\text{KL}}\left(p_Y \| p_Y(\cdot; \theta)\right) = \mathbb{E}_Y\left[\log p_Y(Y)\right] - \mathbb{E}_Y\left[\log p_Y(Y; \theta)\right] \tag{1.47}$$

As the first term does not depend on θ, minimizing this expression is equivalent to minimizing the second part of the equation, so

$$\arg\min_{\theta \in \Theta} D_{\text{KL}}\left(p_Y \| p_Y(\cdot; \theta)\right) = \arg\min_{\theta \in \Theta} -\mathbb{E}_Y\left[\log p_Y(Y; \theta)\right] \tag{1.48}$$

The true distribution of the observation is unknown, but samples y_i are available. Using the empirical KL-divergence denoted $D_{\text{KL}}^{\text{empirical}}$, and replacing the theoretical expectation with the empirical one, the equation above becomes:

$$\arg\min_{\theta\in\Theta} D_{\mathrm{KL}}^{\text{empirical}}(p_Y \| p_Y(\cdot;\theta)) = \arg\min_{\theta\in\Theta} \frac{1}{p}\sum_{i=1}^{p} -\log p_Y(y_i;\theta) = \hat{\theta}_{\mathrm{MLE}} \tag{1.49}$$

Thus, the MLE minimizes the empirical KL-divergence between the true distribution of the observations and the sampling distribution of the observation (that depends on θ).

The MLE possesses desirable asymptotic properties, such as asymptotic normality when the number of observations grows large [Rei13]. Those properties permit the construction of asymptotic confidence interval, and to perform hypothesis testing, especially for model selection. This aspect will be further developed in Section 1.5.

So far, the only information assumed on θ is its parameter space Θ. In the case where some belief on θ is present before the calibration, we can incorporate this information through the Bayesian framework.

1.3.3 Bayesian Inference

In Bayesian inference, the uncertainty present on θ is modelled by considering it as a random variable. Instead of having a precise value for θ, albeit unknown, we assume that we have a *prior distribution* on θ, denoted p_θ, that represents the initial state of belief upon the parameter, prior to any experiment and observations. The choice of this prior distribution will be discussed later. Using the experiment, whose sampling distribution is given by the likelihood, the prior distribution is updated to reflect the new state of belief upon the parameter. The Gaussian likelihood in Eq. (1.38) for the frequentist approach can be almost be rewritten as is in the Bayesian setting, just by conditioning Y with θ. Eq. (1.38) becomes

$$Y \mid \theta \sim \mathcal{N}(\mathcal{M}(\theta), \Sigma) \tag{1.50}$$

and the likelihood is the pdf $\mathcal{L}(\theta; y) = p_{Y|\theta}(y \mid \theta)$. Using Bayes' theorem, the *posterior distribution* of the parameters given the observed data is

$$p_{\theta|Y}(\theta \mid y) = \frac{p_{Y|\theta}(y \mid \theta)p_\theta(\theta)}{p_Y(y)} = \frac{\mathcal{L}(\theta; y)p_\theta(\theta)}{p_Y(y)} \tag{1.51}$$

The denominator can be seen as a normalizing constant, ensuring that $\int_\Theta p_{\theta|Y} = 1$. But it can also be seen as a measure of how well does the model explain the data obtained. This interpretation will be extended in Section 1.5

> **Definition 1.3.2 – Model Evidence:** The model evidence, (or marginal likelihood, integrated likelihood) is defined as the distribution of the data marginalised over the parameters.
>
> $$p_Y(y) = \int_\Theta p_{Y,\theta}(y,\theta)\,\mathrm{d}\theta = \int_\Theta p_{Y|\theta}(y \mid \theta)p_\theta(\theta)\,\mathrm{d}\theta \tag{1.52}$$
>
> This quantity depends implicitly on the underlying mathematical model $\mathfrak{M} = (\mathcal{M}, \Theta)$. Comparing evidence of different models allows for the comparison of those different models. However, computing the model evidence requires the expensive evaluation of

an integral over the whole parameter space, and no analytical form is available except for trivial cases. Specific techniques for this evaluation are reviewed in [FW11].

When the model $(\mathcal{M}, \Theta)$ and the data y is fixed, the model evidence is constant with respect to the calibration parameter θ. The posterior distribution is thus often written and evaluated up to a multiplicative constant.

$$p_{\theta|Y}(\theta \mid y) \propto \mathcal{L}(\theta; y) p_\theta(\theta) \tag{1.53}$$

1.3.3.a Posterior inference

This posterior distribution is central in Bayesian analysis, as it gathers all the information we have on the parameter, given the observed data. Given Eq. (1.51), evaluating the posterior density at a point requires the evaluation of the model evidence, that is an expensive integral. To bypass this evaluation, several techniques have been developed to get samples from a unnormalized arbitrary function. One of the most well-known method is based on the construction of a Markov-chain whose stationary state is the searched posterior. Classical MCMC algorithms such as Metropolis-Hastings requires the use of a proposal density, and then to accept or reject the proposal based on the posterior distribution evaluated at the point.

A lot of refinement of these methods are available in the literature in order to better tackle the high-dimensionality of the parameter space, or to improve the mixing of the sampled MC chain. One important adaptation to mention is Hamiltonian Monte-Carlo [Han01, Bet17], that improves the performance of the chain by using the value of the gradient of the log-posterior distribution. Obtaining this gradient (although for a different purpose) is discussed in Section 1.4.

For time-dependent systems, Bayesian framework is particularly well-suited to treat observations sequentially, especially because Bayesian updating is done via multiplication. Bayes' theorem is the basis of many data assimilation methods, such as Kalman filter or various particle filters, that are often used for state estimation.

1.3.3.b Bayesian Point estimates

The whole posterior distribution aggregates a lot of information on the problem. However, as mentioned above, a certain work has to be done in order to get independent samples. Instead, one can try to find a point $\theta \in \Theta$ that summarizes as best this distribution. Consequently, the chosen estimate is often an indicator of the central tendency. In that sense, we wish to get a value that is quite close to all sampled values from the posterior [LC06].

Let us define a function L that measures a distance in the parameter space: $L : \Theta \times \Theta$. For a candidate θ', the measured risk with respect to a sample from the posterior $\theta_{\text{sample}} \sim \theta \mid Y$ is $L(\theta', \theta_{\text{sample}})$. The *Bayesian risk* for θ' is then the expectation of this Bayesian loss functions L under the posterior distribution: $\mathbb{E}_{\theta|Y}\left[L(\theta', \theta) \mid y\right]$. A Bayesian

point estimate is defined as a minimizer of the Bayesian risk:

$$\hat{\theta}_L = \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y} \left[L(\theta', \theta) \mid y \right] \tag{1.54}$$

Obviously, different loss functions will lead to different Bayesian point estimates, and we are going to evoke two of them.

Posterior mean

By defining L as the squared error $L(\theta', \theta) = (\theta' - \theta)^2$, we can define the Mean Squared Error (MSE) as $\text{MSE} : \theta' \mapsto \mathbb{E}_{\theta|Y} \left[(\theta' - \theta)^2 \mid y \right]$. Finally, the value corresponding to the Minimum Mean Squared Error is

$$\hat{\theta}_{\text{MMSE}} = \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y} \left[(\theta' - \theta)^2 \mid y \right] \tag{1.55}$$

Simple algebraic manipulations show that the minimizer is in fact the posterior mean:

$$\hat{\theta}_{\text{MMSE}} = \mathbb{E}_{\theta|Y}[\theta \mid y] = \int_{\Theta} \theta \cdot p_{\theta|Y}(\theta \mid y) \, \mathrm{d}\theta \tag{1.56}$$

In order to compute $\hat{\theta}_{\text{MMSE}}$, it is easier to compute directly the mean of the posterior samples obtained via posterior inference, than to solve the minimization problem in Eq. (1.55).

Posterior Mode: the MAP

Taking $L(\theta', \theta) = -\delta_\theta(\theta')$, the dirac delta function defined in Eq. (1.14), one can show that the minimizer of $\mathbb{E}_{\theta|Y} \left[L(\theta', \theta) \mid y \right]$ is the mode of the posterior distribution, and is called the *Maximum A Posteriori* (MAP):

$$\begin{aligned} \hat{\theta}_{\text{MAP}} &= \arg\min_{\theta' \in \Theta} \mathbb{E}_{\theta|Y} \left[\delta_\theta(\theta') \mid y \right] = \arg\min_{\theta' \in \Theta} -p_{\theta|Y}(\theta' \mid y) \\ &= \arg\max_{\theta' \in \Theta} p_{\theta|Y}(\theta' \mid y) = \arg\max_{\theta' \in \Theta} \mathcal{L}(\theta'; y) p_\theta(\theta') \end{aligned} \tag{1.57}$$

One interesting fact about the MAP, is that its evaluation does not require the full knowledge of the posterior distribution, nor samples to evaluate the integral of Eq. (1.56). We can resort to classical optimization techniques for this evaluation. Similarly to the likelihood, taking the negative logarithm leads to the following minimization problem.

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta' \in \Theta} -\log \mathcal{L}(\theta'; y) - \log p_\theta(\theta') \tag{1.58}$$

1.3.3.c Choice of a prior distribution

As seen in the application of Bayes' theorem in Eq. (1.51), the prior has a preponderant role in the formulation of the posterior distribution. Indeed, this prior distribution

represents the current state of knowledge on the value of the parameter, before any experiment. This comes usually from an expert opinion, or some reasonable assumptions about the nature of θ.

Let us assume for instance that we have a Gaussian prior for θ: $\theta \sim \mathcal{N}(\theta_b, B)$ where B is called the background covariance error matrix and θ_b is called the *background value* that acts as a plausible reference value. Assuming a Gaussian form for the errors as well with covariance matrix Σ, the MAP can be written as

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta \in \Theta} \frac{1}{2}\|\mathcal{M}(\theta) - y\|^2_{\Sigma^{-1}} + \frac{1}{2}\|\theta - \theta_b\|^2_{B^{-1}} \tag{1.59}$$

Adding a Gaussian prior for the parameter comes down to adding a L^2 regularization term to the optimization problem, also called Tikhonov regularization [TA77]. This expression is very analoguous to the state estimation in the 3D-Var method in Data assimilation. Other choices of priors lead to other regularizations, such as the lasso regularization [Tib11] that is a consequence for choosing θ that follows a priori a Laplace distribution of mean 0.

The choice of a prior distribution has an influence on the inference of the parameter and its point estimation. Where there is no knowledge on the parameter beforehand, one can try to choose a non-informative prior in order to try to mitigate its effect. One can for instance choose a "flat" prior over the parameter space, but this can lead to *improper prior*, in the sense that they do not integrate to 1. However, improper priors do not necessarily lead to improper posterior, allowing for the usual Bayesian analysis of the quantity. For instance, if $\Theta = \mathbb{R}^n$, the prior $p_\theta(\theta) \propto 1$ is improper, but the MAP estimation is equivalent to the MLE.

All in all, when looking for the MAP or the MLE, parameter estimation boils down to the minimization of a well chosen objective function, that measures the misfit between the output of the numerical model and the observations. This cost function will be written J in the following, to match the notation of data assimilation. In this context of calibration, we can then summarize the estimation as a minimization problem, where J represents some kind of distance between $\mathcal{M}(\theta)$ and the observations.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} J(\theta) \tag{1.60}$$

1.4 Calibration using adjoint-based optimization

Point estimates in this context take the form of extremum estimators, that is an extremum of some given objective function J. This function takes the form of the log-likelihood for the MLE, or the log-posterior for the MAP, but other misfits can be considered, such as optimal transport based metrics. The formulation is then quite simple, but the problem of efficient optimization remains. For differentiable problems, most of minimization instances are solved using gradient based methods, such as gradient descent, quasi-newton methods. However, this implies to be able to compute efficiently the gradient of the cost function J with respect to the parameter: $\nabla_\theta J$. The straightforward way,

is to compute the gradient using finite differences. Let us suppose that $\theta = (\theta_1, \cdots \theta_n)$, and e_i is 0 for all its component except the ith one which is 1. The gradient can be approximated by the usual 1st order forward finite-difference scheme, as displayed in Eq. (1.61).

$$\nabla_\theta J \approx \left[\frac{J(\theta + \epsilon e_1) - J(\theta)}{\epsilon}, \frac{J(\theta + \epsilon e_2) - J(\theta)}{\epsilon}, \ldots, \frac{J(\theta + \epsilon e_n) - J(\theta)}{\epsilon} \right] \quad \text{for } \epsilon \ll 1 \tag{1.61}$$

In addition to the run of the model at θ, we have to evaluate the model n times, for each one of the coordinate of θ. If this is feasible in practice for low dimensional problems, this is impossible for large problems that cumulate more than hundreds of parameters. Nevertheless, different methods can be used to compute the gradient, atleast approximately for optimization purpose: for instance, [Bou15] uses Simultaneous Perturbation Stochastic Approximation to approximate the gradient using only one additional run, indepedently on the number of parameters.

In geophysical applications, parameter estimation and the subsequent optimization is usually performed by deriving the adjoint equation in order to get the exact gradient for a relatively reasonable cost. This gradient is used afterward in optimization methods such as conjuguate gradient, or BFGS for instance. Adjoint methods are thus very popular in large-scale optimization of Computational Fluid Dynamics codes, as the additional cost of implementation is often worth the gain in the short term. This situation is common in data assimilation, as shown in [DL91, DL92, HMR+10, CMMV13], or in shape optimization of airfoils in [HB01].

To derive the adjoint equations, we will first rewrite the cost function as a function of the forward operator and the parameter: $J(\theta) = J(\mathcal{M}(\theta), \theta)$: The estimation of the parameter can be written as the following constrained optimisation problem:

$$\begin{aligned} &\min_{\theta \in \Theta} J(\theta) = J(y, \theta) \\ &\text{such that } \mathcal{F}(y, \theta) = 0 \end{aligned} \tag{1.62}$$

where the constraint on $\mathcal{F}$ signifies that the model is admissible, i.e. that $y = \mathcal{M}(\theta) \in \mathbb{Y}$.

Differentiating the Eq. (1.62) with respect to θ using the chain rule gives

$$\begin{aligned} \nabla_\theta J &= \frac{\partial J}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial J}{\partial \theta} \\ \nabla_\theta \mathcal{F} &= \frac{\partial \mathcal{F}}{\partial y}\frac{\partial y}{\partial \theta} + \frac{\partial \mathcal{F}}{\partial \theta} \end{aligned} \tag{1.63}$$

In those equations, the partial derivatives with respect to θ are quite easily obtainable, while the real challenge is to obtain the derivative with respect to the state variable: $\frac{\partial}{\partial y}$.

To treat the constrained optimization in Eq. (1.62), let us introduce the Lagrange multiplier $\lambda \in \mathbb{Y}$, so that we can write the Lagrangian $\mathscr{L}$

$$\mathscr{L}(\theta, y, \lambda) = J(y, \theta) - \lambda^T \mathcal{F}(y, \theta) \tag{1.64}$$

is then

$$\min_{\theta, y, \lambda} \mathscr{L}(\theta, y, \lambda) \tag{1.65}$$

The first-order condition of optimality for the Lagrangian: $\frac{\partial \mathscr{L}}{\partial \theta} = \frac{\partial \mathscr{L}}{\partial y} = \frac{\partial \mathscr{L}}{\partial \lambda} = 0$ translates into the optimality condition, adjoint equation and the state equation: When differentiating with respect to the adjoint variable, we retrieve the state equation:

$$\frac{\partial \mathscr{L}}{\partial \lambda} = -\mathcal{F}(y, \theta) = 0 \qquad \text{(State equation)}$$

When differentiating with respect to the state variable, the equation that verifies the adjoint variable is called the adjoint equation

$$\frac{\partial \mathscr{L}}{\partial y} = \frac{\partial J}{\partial y} - \lambda^T \frac{\partial \mathcal{F}}{\partial y} = 0 \qquad \text{(Adjoint equation)}$$

Finally, when λ verifies the adjoint equation: $\left(\frac{\partial \mathcal{F}}{\partial y}\right)^T \lambda = \left(\frac{\partial J}{\partial y}\right)^T$, the gradient of the cost function can be expressed using the partial derivative *with respect to* θ of the cost function and of the forward model, and the adjoint variable:

$$\frac{\partial \mathscr{L}}{\partial \theta} = \nabla_\theta J = \frac{\partial J}{\partial \theta} - \lambda^T \frac{\partial \mathcal{F}}{\partial \theta} = 0 \qquad \text{(Optimality condition)}$$

So, to get $\nabla_\theta J$, as the partial derivatives with respect to the control variable are relatively easy to obtain, the challenge lies in solving the adjoint equation. Albeit tedious, one can derive those equations by writing the tangent linear model of the original model, and implement a dedicated solver for the adjoint variables. A more common and way simpler approach is to derive the adjoint equations directly from the computer code implemented to solve the model, by using Automatic differentiation tools, such as TAPENADE [HP13]. Those programs directly translate the source code into a program that solves the original model equations and the adjoint equations, and outputs the gradient along with the cost function.

1.5 Model selection

So far, we have discussed the calibration of a specific model $(\mathcal{M}, \Theta)$ given some observations, thus solving an inverse problem and finding $\hat{\theta}$ as an extremum of an objective function. But different models may be considered to explain the data. Those models may differ by their forward operator, by their parameter space, or by both at the same time.

But changing models also means changing the potential "best" fit attainable, in terms of minimum reached by the objective function. More complex models usually provide a better fit of the model but to the cost of a higher dimension in the parameter space. At the same time, more complex models may exhibit an overfitting behaviour.

> **Example 1.5.1:** Figure 1.6 shows a curve-fitting problem using polynomial functions, where the y_i's are realisations of $Y_i \sim \mathcal{N}(i, 1)$ for $i = 0$ to 10. For a problem of curve fitting using polynomial functions, which is illustrated Fig. 1.6, increasing the degree of the polynomial used (thus the dimensionality of the model) decreases the minimum value of J reached. However, the increase in the degree leads also to some

oscillations between the sampled points, as the fitting procedure looks to account for the deviations due to the random origin of the y_i's

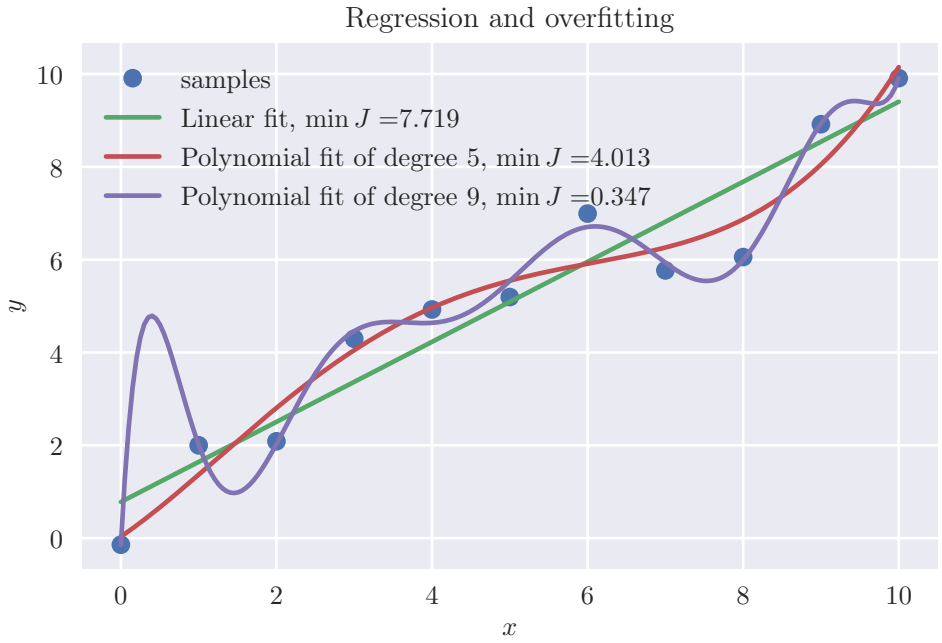

Figure 1.6: Overfitting phenomenon, and reduction of the cost function

We can then look to can reduce the complexity of the model, without decreasing significantly its performance.

We are first going to consider the case of nested models: models that share the same forward model, but whose parameter spaces are nested. Model selection in this case is a way to reduce the dimension of the model, by reducing the parameter space.

Finally, tools introduced in this section bring up model comparison. A calibrated model $(\mathcal{M}, \{\hat{\theta}\})$ is optimal given an objective function, but we can show that values close to the calibrated value $\hat{\theta}$ may also be of interest, as the decrease of performance may not be statistically detectable. The "perturbed" model $(\mathcal{M}, \{\hat{\theta} + \varepsilon\})$ for small ε may accurately describe the data as well.

1.5.1 Likelihood ratio test and relative likelihood

Generally speaking, more complex models have a better ability to represent the data, but all the parameters included in the model may not be relevant for the modelling. It can be interesting to test if a "simpler" model would give similar performances, or at least show a decrease in performances that is not statistically significant. One of the most well-known test is the Likelihood-ratio test, that tests if two *nested models* are equivalent: Let us consider two nested models: $\mathfrak{M}_1 = (\mathcal{M}_1, \Theta_1)$, $\mathfrak{M}_2 = (\mathcal{M}_2, \Theta_2)$, such that $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$ and $\Theta_2 \subsetneq \Theta_1$. In this case, $\mathfrak{M}_2$ represents the simpler model, with

a reduced parameter space, while $\mathfrak{M}_1$ is the more general model. Recalling the notion of model dimension in Remark 1.2.2, $\mathfrak{M}_1$ has dimension r, and $\mathfrak{M}_2$ has dimension d with $r > d$. As $\mathfrak{M}_1$ is more general, one can expect better performances.

The likelihood ratio is defined as the ratio of the largest values taken by the likelihood on their respective parameter space, value that is assumed to be attained at $\hat{\theta}_1$ and $\hat{\theta}_2$.

$$\Lambda(y) = \frac{\sup_{\theta\in\Theta_2} \mathcal{L}(\theta; y)}{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; y)} = \frac{\mathcal{L}(\hat{\theta}_2; y)}{\mathcal{L}(\hat{\theta}_1; y)} \leq 1 \tag{1.66}$$

Based on this quantity, we can test whether the smaller model is sufficient to explain the data as good as the larger model. The two hypothesis for this test are

- $\mathcal{H}_0$: The two models are statistically equivalent: the difference between the maximal values of the likelihood is not statistically significant. This corresponds to Λ close to 1
- $\mathcal{H}_1$: the two models are statistically different: the larger model performs better than the reduced one. This corresponds to Λ significantly smaller than 1.

Under the null hypothesis, $-2\log\Lambda$, (sometimes called the deviance) follows asymptotically (as the number of observations becomes large) a χ^2 distribution defined in Example 1.2.16. The number of degrees of freedom of the χ^2 distribution is given by the difference of dimensionality between the two models:

$$-2\log\Lambda(y) \xrightarrow{\mathrm{d}} \chi^2_{r-d} \tag{1.67}$$

By denoting $\chi^2_{r-d}(1-\nu)$ the quantile of order $1-\nu$ of the χ^2 distribution with $r-d$ degrees of freedom, the asymptotic rejection region of level ν is:

$$\mathrm{RejReg}_\nu = \left\{ y \mid -2\log\Lambda(y) > \chi^2_{r-d}(1-\nu) \right\} \tag{1.68}$$

Or by reformulating using the log-likelihoods and objective functions $l(\theta; y) = \log\mathcal{L}(\theta; y) = -J(\theta)$

$$\mathrm{RejReg}_\nu = \left\{ y \mid (\sup_{\theta\in\Theta_1} l(\theta; y) - \sup_{\theta\in\Theta_2} l(\theta; y)) > \frac{1}{2}\chi^2_{r-d}(1-\nu) \right\} \tag{1.69}$$

$$= \left\{ y \mid J(\hat{\theta}_2) - J(\hat{\theta}_1) > \frac{1}{2}\chi^2_{r-d}(1-\nu) \right\} \tag{1.70}$$

As a basis for comparison, when $\Theta \subset \mathbb{R}$, $r-d=1$ and $\chi^2_1(1-0.05) = 3.84$. When the data falls into the rejection region (*i.e.* $J(\hat{\theta}_2)$ significantly larger than $J(\hat{\theta}_1)$), the null hypothesis is rejected, and the models can be asserted significantly different.

Conversely, the rejection region defined above allows us to define an asymptotic confidence interval for the parameter. Let us introduce the *Relative Likelihood* ([Kal85]) which is the ratio of the likelihood evaluated at a point θ to the maximal value of the likelihood:

$$R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\mathrm{MLE}}; y)} = \frac{\mathcal{L}(\theta; y)}{\sup_{\theta'\in\Theta} \mathcal{L}(\theta'; y)} \tag{1.71}$$

This ratio allows for comparing the plausibility of the value θ, compared to the MLE. The likelihood interval of level $p \in]0, 1]$ is defined as

$$\mathcal{I}_{\text{Lik}}(p) = \left\{ \theta \mid R(\theta) = \frac{\mathcal{L}(\theta; y)}{\mathcal{L}(\hat{\theta}_{\text{MLE}}; y)} \geq p \right\} \tag{1.72}$$

p can be set to an arbitrary threshold, but it can also be chosen specifically in order to avoid the rejection region of a likelihood ratio test with certain confidence. When comparing the models $(\mathcal{M}, \{\theta\})$ and $(\mathcal{M}, \Theta)$, let $R(\theta)$ be their likelihood ratio. The complement of the rejection region Eq. (1.70) written using Eq. (1.72) is

$$\mathcal{I}_{\text{Lik}}\left(\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\nu)\right)\right) = \left\{\theta \mid R(\theta) \geq \exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\nu)\right)\right\} \tag{1.73}$$

The values of the calibrated parameters in this set generate models that are statistically equivalent to the model comprising the MLE as its calibrated parameter.

For 1 dimensional models, and the confidence level of .05, the threshold of Eq. (1.73) is $\exp\left(-\frac{1}{2}\chi^2_{\dim(\Theta)}(1-\nu)\right) = \exp\left(-\frac{1}{2}\chi^2_1(.95)\right) \approx 0.15$, and at a level .10, $\exp\left(-\frac{1}{2}\chi^2_1(.90)\right) \approx 0.26$.

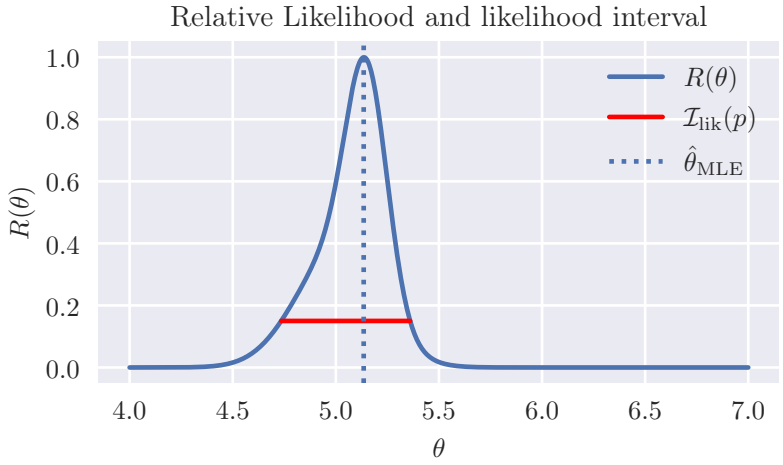

Figure 1.7: Example of relative likelihood, and associated likelihood interval

Due to the likelihood ratio test and the relative likelihood, we can see that even though $\hat{\theta}_{\text{MLE}}$ is the optimizer of the likelihood function, other values close to it may not be discarded, provided that the log-likelihood does not drop off too much.

1.5.2 Criteria for model comparison

The likelihood ratio test presented above is defined for nested models. In the more general case, we can also associate each model with a single numerical value that measures the balance between "fit" and complexity of the model. This takes usually the form of

$$\text{Crit}(\mathfrak{M}) = -2\log\mathcal{L}(\hat{\theta}_{\text{MLE}}) + \text{Complexity penalization} \tag{1.74}$$

where $\mathcal{L}$ is the likelihood function for the model $\mathfrak{M} = (\mathcal{M}, \Theta)$, and $\mathcal{L}(\hat{\theta}_{\text{MLE}}) = \max \mathcal{L}$. The role of the complexity penalization is to avoid overfitting, and is often directly linked to the dimension of the parameter space Θ. Two quite popular examples of criteria are the AIC (Akaike Information Criterion) introduced in [Aka74] and the BIC (Bayesian Information Criterion) in [Sch78].

Then, to compare two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the magnitude of the difference $\text{Crit}(\mathfrak{M}_1) - \text{Crit}(\mathfrak{M}_2)$ is compared to some thresholds as shown in [BA04]. The difference shows whether a model should be preferred, or if no substantial evidence exists for either model. These criteria, as well as the likelihood ratio test, are based on the evaluation of the likelihood at its maximal value. Another approach is to marginalize the likelihood with respect to the prior distribution of the calibration parameter, giving Bayesian model comparison.

1.5.3 Bayesian model comparison

Given a model $\mathfrak{M}$, the model evidence is the likelihood marginalized over the parameter space as introduced in Definition 1.3.2, and will be written $p_{Y|\mathfrak{M}}$. This evidence represents how likely have the data y been generated using the statistical model $\mathfrak{M}$.

For two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ the Bayes' factor is defined as the ratio of the evidence of the two models:

$$\text{BF}(\mathfrak{M}_1, \mathfrak{M}_2) = \frac{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_1)}{p_{Y|\mathfrak{M}_1}(y \mid \mathfrak{M}_2)} \tag{1.75}$$

where

$$p_{Y|\mathfrak{M}_i}(y \mid \mathfrak{M}_i) = \int_{\Theta_i} p_{Y,\theta|\mathfrak{M}_i}(y, \theta \mid \mathfrak{M}_i)\, \mathrm{d}\theta = \int_{\Theta_i} p_{Y|\theta,\mathfrak{M}_i}(y \mid \theta, \mathfrak{M}_i) p_{\theta|\mathfrak{M}_i}(\theta \mid \mathfrak{M}_i)\, \mathrm{d}\theta \tag{1.76}$$

Quite similarly to the BIC and AIC, the logarithm of the Bayes' factor is usually compared to specific values, allowing us to conclude roughly on how strong does the data favors $\mathfrak{M}_1$. Again, the logarithm of Bayes' factor is compared with specified threshold ([KR95, BA04]).

1.6 Parametric model misspecification

We introduced earlier the mathematical model $(\mathcal{M}, \Theta)$, and based our analysis on the fact that the "target model", i.e. the reality is $(\mathscr{M}, \Theta_{\text{real}} = \Theta)$, so the parameter spaces are the same. However, between the reality and the numerical model, various simplifications are introduced, thus the reality is not often completely *representable* by the numerical model: we have then misspecified models.

> **Definition 1.6.1 – Misspecified model:** Let Y be the random variable associated with the observations, and p_Y its pdf. Let $\{p_{Y|\theta}; \theta \in \Theta\}$ the parametric family of

densities, among which we are looking to find p_Y. The model is said to be misspecified, if $p_Y \notin \{p_{Y|\theta}; \theta \in \Theta\}$.

We can also define this misspecification in terms of numerical models defined in this chapter: let $(\mathscr{M}, \Theta_{\text{real}})$ be the physical system under study and $(\mathcal{M}, \Theta)$ the numerical model that is to be calibrated with respect to the observations $y = \mathscr{M}(\vartheta)$. $(\mathcal{M}, \Theta)$ is said to be misspecified, if $\vartheta \notin \Theta$

In practice, in addition to the simplifications and the complexity of the reality, the parameter space Θ does not contain necessarily all the parameters needed to run the forward model, but represents the space of the parameters of interest, or calibration parameters. In addition to them, some other parameters are at play, that we are going to call the *environmental parameters*, or *uncertain parameters* written $u \in \mathbb{U}$. These parameters come from instance from the external forcings.

Bayesian framework and more specifically Bayesian update of the prior by the likehihood puts the emphasis on the update of the information on the *parameter of interest*. However the environmental parameters are assumed to have an inherent variability. In that sense, it may not be worth spending time and resources to infer these parameter values, as they are bound to change. Moreover, we can only get information on the environmental conditions used to generate the observations.

In terms of models, each choice of $u \in \mathbb{U}$ gives a different model $\mathfrak{M}(u) = \{\mathcal{M}(\cdot, u), \Theta\}$. Let us assume that we can model the uncertain parameters as a random variable U. Let us consider that we chose a specific $u_0 \in \mathbb{U}$, and that we are given some observation $y = \mathscr{M}(\vartheta)$. We can formulate an inverse problem, and an objective function $J : \theta \mapsto J(\theta, u_0)$, that we wish to minimize with respect to θ. Some estimators still carry nice properties. The MLE for instance, defined Section 1.3.2 can still be written as the minimizer of the empirical KL–divergence. We assume that we can write the sampling distribution as $p_{Y|\theta,U}$, and

$$\hat{\theta}_{\text{MLE}}(u_0) = \arg\min_{\theta \in \Theta} D_{\text{KL}}^{\text{empirical}} \left(p_Y \| p_{Y|\theta,U}(\cdot \mid \theta, U = u_0) \right) \tag{1.77}$$

and can be seen as the "best" value given $U = u_0$. However, the asymptotic properties of the MLE are slightly different as described in [Whi82]. So when the model is misspecified, minimizing the same cost function still makes sense.

However, the calibration will depend on the chosen u_0: $\hat{\theta}(u_0) = \arg\min_{\theta \in \Theta} J(\theta, u_0)$, and there is no guarantee that $\hat{\theta}(u_0)$ will minimize $J(\cdot, u_1)$ for $u_0 \neq u_1$, as illustrated Fig. 1.8.

u_0 and u_1 are close to each other, but $\hat{\theta}(u_0)$ and $\hat{\theta}(u_1)$ are not. However, as the cost function shows similar values at those points, choosing either one would lead to a satisfactory cost given $U = u_0$ or $U = u_1$. If $U = u_2$ is considered as well, the modeller may have a preference and choose $\hat{\theta}(u_1)$ as the final estimator.

In terms of model selection, the asymptotic distribution of the likelihood ratio statistic defined Section 1.5.1 is also slightly different. Instead of following a χ^2_r distribution, where r is the number of dimensions for the test, $-2 \log \Lambda$ will asymptotically have the

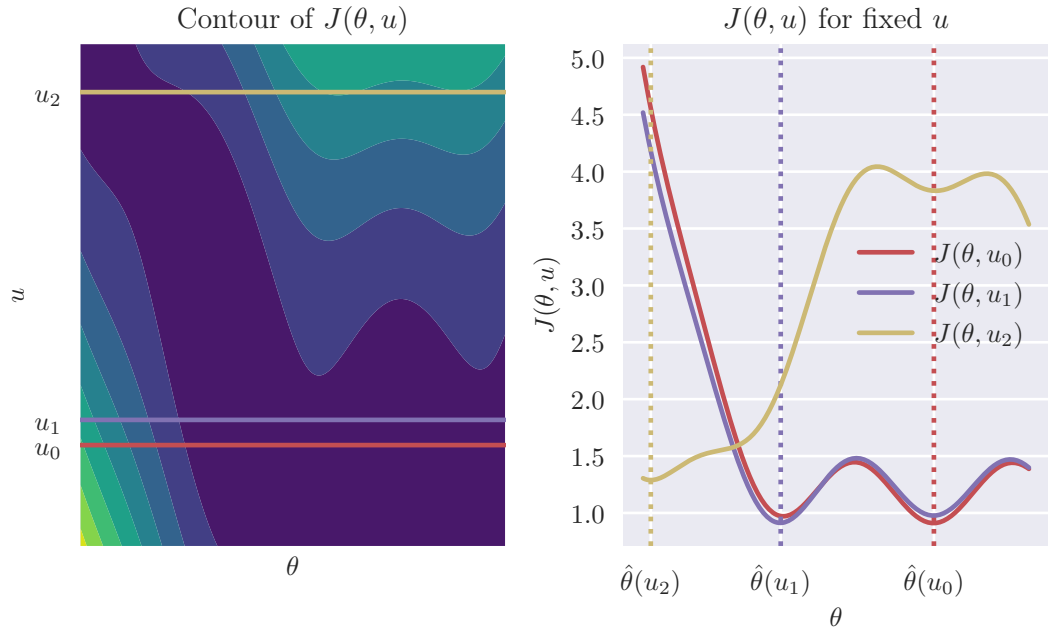

Figure 1.8: Effect of the misspecification on the minimizer.

same distribution as a weighted sum of r random variables, where every one have a χ^2_1 distribution and whose weights are the eigenvalues of a matrix involving the Jacobian and the Hessian of the log-likelihood [Ken82].

This random misspecification leads to some issues in the calibration of the model, and it asks for a notion of robustness with respect to the environmental parameters.

1.7 Partial conclusion

In this chapter, starting from a probabilistic point of view, we established the usual tools encountered in model calibration: the misfit between the data and the numerical model is measured by a cost function J, that can be minimized using for instance gradient descent. From this optimization, we can define a "acceptable" region for the estimate. In other words, values in this set yield an misfit that is not different enough to be completely discarded.

Adding an environmental variable as a random nuisance parameter introduces a parametric misspecification of the model: each realization of this underlying random variable will yield a different estimation. In Chapter 2, we will discuss the notion of robustness under this random misspecification, and introduce a family of robust estimators, inspired by model selection.

CHAPTER 2

ROBUST ESTIMATORS IN THE PRESENCE OF UNCERTAINTIES

Contents

In the previous chapter, we introduced the problem of calibration of a numerical model with respect to a *calibration parameter* θ. This takes the form of the optimisation of an objective function. We also raised the problem of parametric misspecification of the numerical model with respect to the reality: $u \in \mathbb{U}$. Moreover, this misspecification is modelled by a random variable U with known distribution. One desirable property is that the calibrated model shows relatively good performances when the environmental variables varies, or in other words, we want the calibrated model to be *robust* with respect to the varying environmental parameters. In this chapter, we are going to introduce some criteria that aim at solving this *robust optimization problem*. The actual computation of those estimates will be discussed in the next chapter.

2.1 Defining robustness

2.1.1 Classifying the uncertainties

In the Bayesian formulation of the problem, the uncertainty on the calibration parameter is modelled through the prior distribution, while the uncertain parameter, u has its own distribution. While mathematically similar, those two representations actually encompasses a significant difference: we are actively trying to reduce the uncertainty of the calibration parameter by Bayesian update, while the uncertainty on the environmental parameter is seen as a nuisance.

In that context, the very notion of uncertainty can be roughly split in two, as described in [WHR+03]:

- Aleatoric uncertainties, coming from the inherent variability of a phenomenon, *e.g.* intrinsic randomness of some environmental variables
- Epistemic uncertainties coming from a lack of knowledge about the properties and conditions of the phenomenon underlying the behaviour of the system under study

According to this distinction, the epistemic uncertainty can be reduced by investigating the effect of the calibration parameter θ upon the physical system, and choose it accordingly to an objective function. The uncertain variable u on the other hand is uncertain in the aleatoric sense, and cannot be controlled directly, as its value is doomed to change. This is why we model it using a random variable U.. This distinction illustrated Fig. 2.1 is a bit simplistic, as [KD09] point out that deciding the type of uncertainties is up to the modeller, who decides on which parameters inference is worth doing.

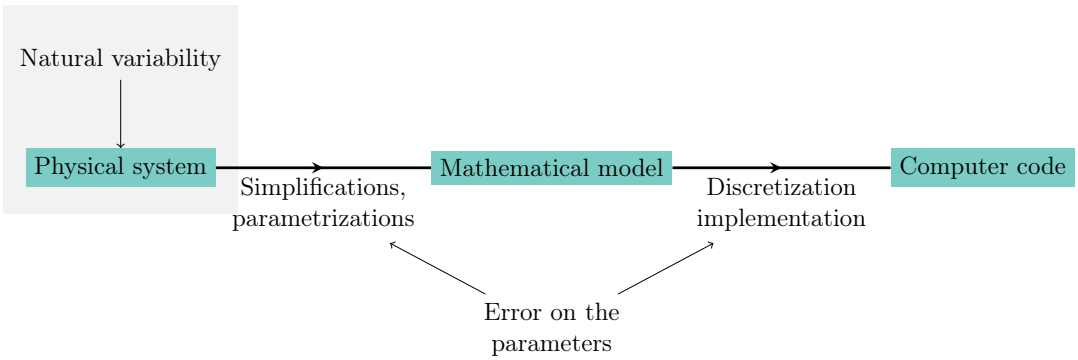

Figure 2.1: Sources of uncertainties and errors in the modelling. The natural variability of the physical system can be seen as aleatoric uncertainties, and the errors on the parameters as epistemic uncertainties

2.1.2 Robustness and/or reliability

The notion of *robustness* is dependent on the context in which it is used. In this work, the term "robust" qualifies a model that behaves still nicely under uncertainties, or to put it in an other way, that is insensitive up to certain extent to some perturbations. Moreover, robustness is often linked and sometimes confused to the semantically close notion of *reliability*. In [LBM+16] we can find summarized in Table 2.1 the difference between these notions, by defining optimality as the deterministic counterpart of robustness, and admissibility as the counterpart of reliability.

	No objective	Objective with deterministic inputs	Objective with uncertain inputs
Unconstrained		Optimal	Robust
Deterministic constraints	Admissible	Optimal and admissible	Robust and admissible
Uncertain constraints	Reliable	Optimal and reliable	Robust and reliable

Table 2.1: Types of problems, depending on their deterministic nature for the constraints or the objective. Shaded cells correspond to problems comprising an uncertain part. Reproduced from [LBM+16]

Other definitions of robustness can be encountered in the literature, and will not be treated in this work: Bayesian approaches are sometimes criticized for their use of subjective probabilities that represent the state of beliefs, especially on the choice of prior distributions. In that sense, robust Bayesian analysis aims at quantifying the sensitivity of the choice of the prior distribution on the resulting inference and relative Bayesian quantities derived. In the statistical community, robustness is often implied as the non-sensitivity on the outliers in the sample set.

2.1.3 Robustness under parameteric misspecification

Given a family of models $\{(\mathcal{M}(\cdot, u), \Theta), u \in \mathbb{U}\}$ and some observations $y \in \mathbb{Y}$ sampled from a random variable Y, we can derive a problem of parameter estimation for each

$u \in \mathbb{U}$. As detailed in Chapter 1, we can formulate the likelihood $\mathcal{L}$ and the posterior distribution, and then compute the MLE and the MAP.

Not taking into account the uncertainty on u may be an issue in the modelling, especially if the influence of this variable is non-negligible. Choosing a specific $u \in \mathbb{U}$ leads to *localized optimization* [HB01] and *overcalibration*, that is choosing a value $\hat{\theta}$ that is optimal for the given situation (which is induced by u). This value does not carry the optimality to other situations, or in Layman's term according to [ALMP+12], being lured by "fool's gold". In geophysics and especially in hydrological models, this overcalibration may lead to the appearance of abberations in the predictions as those uncertainties become prevalent sources of errors. In hydrology, uncertainties are the principal culprit of the existence of "Hydrological monsters" [KRTK10], that are calibrated models that perform really badly.

There are two main ways to tackle this problem. Since the environmental parameter is random by nature with known distribution, we can introduce it directly in the probabilistic inference framework, by appending u to the calibration parameter and to consider (θ, u) for the inference. This will be treated Section 2.2.

Another way, that we are calling the *variational* approach, is to consider instead the loss function $(\theta, u) \mapsto J(\theta, u)$ that we want to minimize, as introduced in the previous chapter. Due to the uncertainty on u, we can then study the family of random variables indexed by $\theta \in \Theta : \{J(\theta, U); \theta \in \Theta\}$. This will be addressed Section 2.3.

2.2 Frequentist and Bayesian inference

In probabilistic inference, the environmental parameters are sometimes called *nuisance* parameters, and different ways have been studies to remove their influence. We will first detail likelihood-based methods to deal with them and then the extension to Bayesian framework.

2.2.1 Profile and integrated Likelihood

From a frequentist approach, we define the joint likelihood $\mathcal{L}(\theta, u; y) = p_{Y|\theta,U}(y \mid \theta, u)$. Using a Gaussian assumption, the sampling distribution, $Y \mid \theta, U$ is

$$Y \mid \theta, U \sim \mathcal{N}(\mathcal{M}(\theta, U), \Sigma) \tag{2.1}$$

where Σ is a covariance matrix.

There are two common ways to get rid of the nuisance parameters: one by *profiling*, one by *marginalization*. Profiling implies to perform first a maximization of the likelihood with respect to the nuisance parameters:

$$\mathcal{L}_{\text{profile}}(\theta; y) = \max_{u \in \mathbb{U}} \mathcal{L}(\theta, u; y) \tag{2.2}$$

and

$$\hat{\theta}_{\text{prMLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}_{\text{profile}}(\theta; y) \tag{2.3}$$

In other words, considering the most favorable case of the likelihood given the nuisance parameters. Comparing the MLE over $\Theta \times \mathbb{U}$ for the original joint likelihood and the profile MLE on Θ for the profile likelihood, it is straightforward to verify that their components on Θ coincide as

$$\max_{(\theta,u)\in\Theta\times\mathbb{U}} \mathcal{L}(\theta, u; y) = \max_{\theta\in\Theta} \mathcal{L}_{\text{profile}}(\theta; y) \tag{2.4}$$

The resulting estimator does not take into account the uncertainty upon u, and can perform quite badly when the likelihood presents sharp ridges [BLW99].

Another alternative is to define the *integrated*, or *marginalized* likelihood as

$$\begin{aligned}
\mathcal{L}_{\text{integrated}}(\theta; y) &= \int_{\mathbb{U}} \mathcal{L}(\theta, u; y) p_U(u)\, \mathrm{d}u && (2.5)\\
&= \int_{\mathbb{U}} p_{Y|\theta,U}(y \mid \theta, u) p_U(u)\, \mathrm{d}u && (2.6)\\
&= \int_{\mathbb{U}} p_{Y,U|\theta}(y, u \mid \theta)\, \mathrm{d}u && (2.7)\\
&= p_{Y|\theta}(y \mid \theta) && (2.8)
\end{aligned}$$

and by maximizing this function,

$$\hat{\theta}_{\text{intMLE}} = \arg\max_{\theta\in\Theta} \mathcal{L}_{\text{integrated}}(\theta; y) \tag{2.9}$$

Example 2.2.1: In order to illustrate the difference between those two methods, the profile and integrated likelihood have been computed for the following likelihood:

$$Y \mid \theta, U \sim \mathcal{N}(\theta + u^2, 2^2) \tag{2.10}$$

and the observations $y = (y_1, \cdots, y_{10})$ have been generated using $\theta + u^2 = 1$. We set $\Theta = [-5, 5]$ and $\mathbb{U} = [-2, 2]$. The likelihood evaluated on $\Theta \times \mathbb{U}$ is displayed Fig. 2.2, with the integrated and profile likelihood. We can see that there is not unicity of the maximizer for the profile likelihood: $\mathcal{L}_{\text{profile}}(\theta; y)$ is constant for $\theta \in [-3, 1]$. This is due to the fact that the observations can have been generated with any θ and u verifying $\theta + u^2 = 1$. For the integrated likelihood however, there is a unique maximum, attained for $\hat{\theta}_{\text{intMLE}} \approx 0.8$.

2.2.2 Joint posterior distribution

Similarly as in Section 1.3.3, we can incorporate information on θ by introducing a prior distribution p_θ, and we can derive the posterior distribution using Bayes' theorem. We assume that U and θ are independent: $p_{\theta,U} = p_\theta \cdot p_U$. The likelihood of the data given θ and u is

$$\mathcal{L}(\theta, u; y) = p_{Y|\theta,U}(y \mid \theta, u) \tag{2.11}$$

The joint posterior distribution can be written as:

$$p_{\theta,U|Y}(\theta, u \mid y) = \mathcal{L}(\theta, u; y) p_\theta(\theta) p_U(u) \frac{1}{p_Y(y)} \tag{2.12}$$

$$\propto \mathcal{L}(\theta, u; y) p_\theta(\theta) p_U(u) \tag{2.13}$$

Here, the posterior is used to do inference on θ and u jointly. In order to suppress the dependency in u, we integrate with respect to U and get the marginalized posterior $p_{\theta|Y}$:

$$p_{\theta|Y}(\theta \mid y) = \int_{\mathbb{U}} p_{\theta,U|Y}(\theta, u \mid y) \, \mathrm{d}u \tag{2.14}$$

$$= \int_{\mathbb{U}} p_{\theta|Y,U}(\theta \mid y, u) p_{U|Y}(u \mid y) \, \mathrm{d}u \tag{2.15}$$

We can then define the *marginalized maximum a posteriori* (MMAP) [DGR02] as the maximizer of this marginalized posterior:

$$\hat{\theta}_{\mathrm{MMAP}} = \arg\max_{\theta \in \Theta} p_{\theta|Y}(\theta \mid y) \tag{2.16}$$

or, by taking the negative logarithm to get a minimization problem, can be written

$$\hat{\theta}_{\mathrm{MMAP}} = \arg\min_{\theta \in \Theta} -\log p_{\theta|Y}(\theta \mid y) \tag{2.17}$$

Unfortunately, neither the integration with respect to the nuisance parameter in Eq. (2.14) nor the subsequent optimization is analytically easy. Assuming that we are able to get i.i.d. samples $\{(\theta_i, u_i)\}_{1 \leq i \leq n_{\mathrm{samples}}}$ from the posterior distribution using MCMC methods for instance, by discarding the u components, the samples $\{\theta_i\}_{1 \leq i \leq n_{\mathrm{samples}}}$ are distributed according to the marginal posterior $p_{\theta|Y}$, and thus can be used to get the MMAP. More direct techniques, such as [DGR02], introduce methods in order to estimate iteratively the MMAP, through sampling of the joint posterior.

Example 2.2.2: Using the same data as in Example 2.2.1, we add a prior distribution of θ as a centered normal distribution, truncated on Θ. On Fig. 2.2 we can see the influence of the prior distribution, as it nudges the MMAP $\hat{\theta}_{\mathrm{MMAP}}$ toward 0, compared to the integrated likelihood.

2.3 Variational approach

We discussed so far the calibration problem with nuisance parameters in the formulation of the likelihood or the posterior distribution. However, in data assimilation for instance, problems of parameter estimation are often formulated directly by introducing a cost function:

$$\begin{array}{rccc} J: & \Theta \times \mathbb{U} & \longrightarrow & \mathbb{R}^+ \\ & (\theta, u) & \longmapsto & J(\theta, u) \end{array} \tag{2.18}$$

This function, in a calibration context, is measuring the misfit between the data y and the forward operator, and can be written as the negative log-likelihood, or the

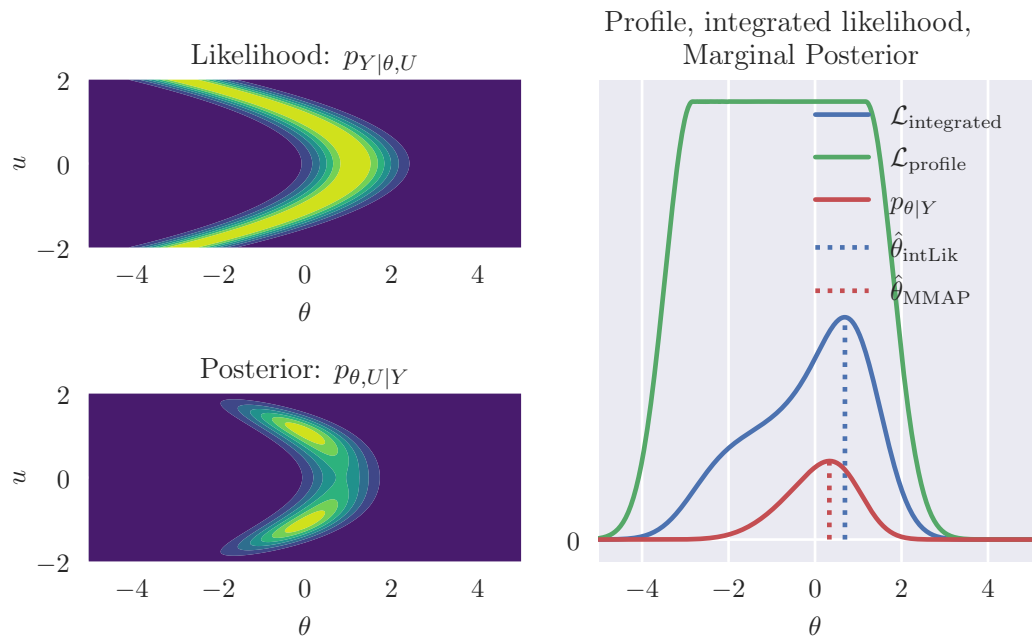

Figure 2.2: Joint likelihood and posterior (left). Profile and integrated likelihood for an uniform nuisance parameter and marginal posterior distribution (right)

negative log posterior distribution. Stil, the objectives and criteria introduced in the following are not specific to this context, and J can represent other properties that ought to be reduced such as instability or drag in airfoil design optimization for instance. This general problem is sometimes quoted as *Optimization under uncertainties* (OUU) ([Coo18, SCIP14])

All in all, $J(\theta, u)$ represent the cost of taking the decision $\theta \in \Theta$ when the environmental variable is equal to u. We are going to make several assumptions on this function:

- Θ is convex and bounded
- For all $\theta \in \Theta$ and $u \in \mathbb{U}$, $J(\theta, u) > 0$
- For all $\theta \in \Theta$, $J(\theta, \cdot)$ is measurable
- For all $\theta \in \Theta$, $J(\theta, U) \in L^p(\mathbb{P}_U)$ and $p \geq 2$. So for each θ, mean and variance exist and are finite.

As the function represents a cost, *i.e.* an undesirable property, we are interested in minimising in some sense this random variable, which depends on θ. Most of existing methods to approach such a problem require first to remove the dependency on the uncertain variable, by defining a *robust* (and deterministic) counterpart of the minimization problem under uncertainty, that we can solve using classical methods of optimization.

2.3.1 Decision under complete uncertainty

We will first detail some estimators that can be argued robust, even though the random nature of U is not directly taken into account. This complete uncertainty is modelled by assuming that no information is available on u, except that $u \in \mathbb{U}$.

2.3.1.a Global optimization

A global optimization criterion, as its name suggests, advocates for minimizing the cost function over the whole space $\Theta \times \mathbb{U}$, giving this optimization problem:

$$\min_{(\theta,u)\in\Theta\times\mathbb{U}} J(\theta, u) \tag{2.19}$$

Rearranging slightly this problem, the θ-component of the minimizer can be written as

$$\hat{\theta}_{\text{global}} = \arg\min_{\theta\in\Theta} \min_{u\in\mathbb{U}} J(\theta, u) \tag{2.20}$$

The global minimum is the equivalent of profile likelihood maximization, when J is the negative log-likelihood. This method exhibits some flaws: we are optimizing the cost function only over the most favourable cases of the environmental parameter, thus there is no guarantee on the behaviour of J outside of those optimistic situations. It then makes sense to "separate" θ and u in the optimization.

2.3.1.b Worst-case optimization

As global optimization is inhenrently optimistic, we can easily derive a criterion which is pessimistic in the sense that we want to minimize over the *least favourable* cases, thus minimizing the loss in the worst-case scenarios. The optimization problem in this case becomes

$$\min_{\theta\in\Theta} \max_{u\in\mathbb{U}} J(\theta, u) \tag{2.21}$$

This criterion is sometimes called Wald's Minimax criterion [Wal45], and the associated estimator is

$$\hat{\theta}_{\text{WC}} = \arg\min_{\theta\in\Theta} \max_{u\in\mathbb{U}} J(\theta, u) \tag{2.22}$$

Minimizing in the worst-case sense also possesses some flaws, especially from a computational point of view. First, the maximum on $\mathbb{U}$ may not exist, especially if $\mathbb{U}$ is unbounded: we could make the model perform as badly as possible by taking extreme values of u. Additionally, if it exists, the resulting estimator is most likely very conservative as only the worse cases are considered.

2.3.1.c Regret maximin

One other approach, called Savage's maximin regret [Sav51] is to compare the current loss to the best performance given the uncertain variable u. The translated loss is called the *regret* and is defined as

$$r(\theta, u) = J(\theta, u) - \min_{\theta \in \Theta} J(\theta, u) \tag{2.23}$$

Using the regret as the new loss function, we can optimize it in the worst-case sense, as introduced in Section 2.3.1.b, and the minimum is attained at $\hat{\theta}_{\text{rWC}}$:

$$\hat{\theta}_{\text{rWC}} = \arg\min_{\theta \in \Theta} \max_{u \in \mathbb{U}} r(\theta, u) \tag{2.24}$$

Example 2.3.1: Figure 2.3 shows global, worst-case and regret optimization for the analytical cost function

$$J(\theta, u) = \left(1 + u(\theta + 0.1)^2\right)\left(1 + (\theta - u)^2\right) \tag{2.25}$$

We can see how the worst-case minimization (in blue) and Savage's maximin regret (in green) compare in this example. Maximin regret will favour values of θ giving a loss that are never too far from the optimal value available, in contrast to the worst-case that focuses on the absolute loss.

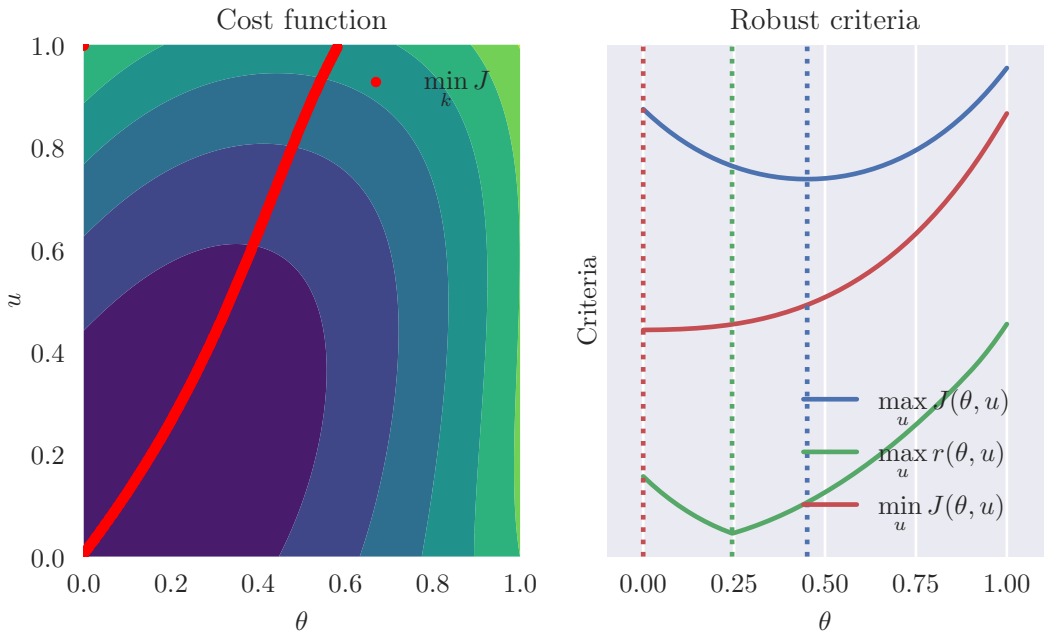

Figure 2.3: Illustration of global optimization, worst-case, and regret worst-case. The red points on the contour of the cost function are the location of the minimizers.

So far, we did not use the fact that u was a realisation of a random variable, and did not take advantage of the knowledge we have upon it. In the next sections, we will see how to incorporate the knowledge of the distribution of U in the estimations.

2.3.2 Robustness based on the moments of an objective function

In the presence of uncertainties, choosing a parameter value θ can also be seen as making a choice under risk. Let $J : \Theta \times \mathbb{U} \to \mathbb{R}^+$ be an objective function, and assume that for all $\theta \in \Theta$, $J(\theta, \cdot)$ is a measurable function. J can be seen as the opposite of the *utility* function, often encountered in game theory or econometrics. Because of the random nature of U, we can define a family of real random variables $\{J(\theta, U) \mid \theta \in \Theta\}$, indexed by $\theta \in \Theta$. In [BS07], the authors proposes an *aggregation approach*, based on the integration with respect to the uncertain variable. An example of this aggregation is the integration of the successive powers of the cost function, in order to get the moments of the associated random variable, that we will detail Sections 2.3.2.a to 2.3.2.c The aggregated objective is then minimized with respect to the control variable.

2.3.2.a Expected loss minimization, central tendency

One of the simplest approach when facing such a problem is to look to optimize a central tendency of those random variables. The mean value being an obvious candidate, we define the expected loss as

$$\mu(\theta) = \mathbb{E}_U \left[J(\theta, U) \right] = \int_{\mathbb{U}} J(\theta, u) p_U(u) \, \mathrm{d}u \tag{2.26}$$

The expected loss $\mu(\theta)$ is sometimes called the conditional mean given θ. Taking the average of the loss function is very common in many problems of classification and regression [Bis06].

The conditional mean is minimized, giving $\hat{\theta}_{\text{mean}}$. Assuming that $J(\theta, u) \propto -\log \mathcal{L}(\theta, u; y)$, we have

$$\hat{\theta}_{\text{mean}} = \arg\min_{\theta \in \Theta} \mu(\theta) = \arg\min_{\theta \in \Theta} \int_{\mathbb{U}} J(\theta, u) p_U(u) \, \mathrm{d}u \tag{2.27}$$

$$= \arg\min_{\theta \in \Theta} - \int_{\mathbb{U}} \log \mathcal{L}(\theta, u; y) p_U(u) \, \mathrm{d}u \tag{2.28}$$

$$= \arg\min_{\theta \in \Theta} - \int_{\mathbb{U}} \log \left(p_{Y|\theta,U}(y \mid \theta, u) \right) p_U(u) \, \mathrm{d}u \tag{2.29}$$

Taking the average of a loss function is the basis of *stochastic programming*. However, the integral Eq. (2.26) is intractable analytically, so instead of computing it exactly, one usually resorts to minimizing the empirical mean risk. For $1 \leq i \leq n_U$, let u_i be i.i.d. samples from U. We can then use those samples to approximate μ: the empirical mean is

$$\mu^{\text{emp}}(\theta) = \frac{1}{n_U} \sum_{i=1}^{n_U} J(\theta, u_i) \tag{2.30}$$

and the minimization problem

$$\min_{\theta\in\Theta} \frac{1}{n_U} \sum_{i=1}^{n_U} J(\theta, u_i) \tag{2.31}$$

is called the *sample average problem* [JNLS09], or *empirical risk minimization* problem in Machine Learning (see e.g. [Vap92]) Other indicators of central tendency can be considered for optimization, such as the mode or the median of the cost function.

Despite some similarities with the integrated likelihood introduced Eq. (2.5), $\hat{\theta}_{\text{mean}}$ and $\hat{\theta}_{\text{intMLE}}$ are not equal in general, as shown Fig. 2.4.

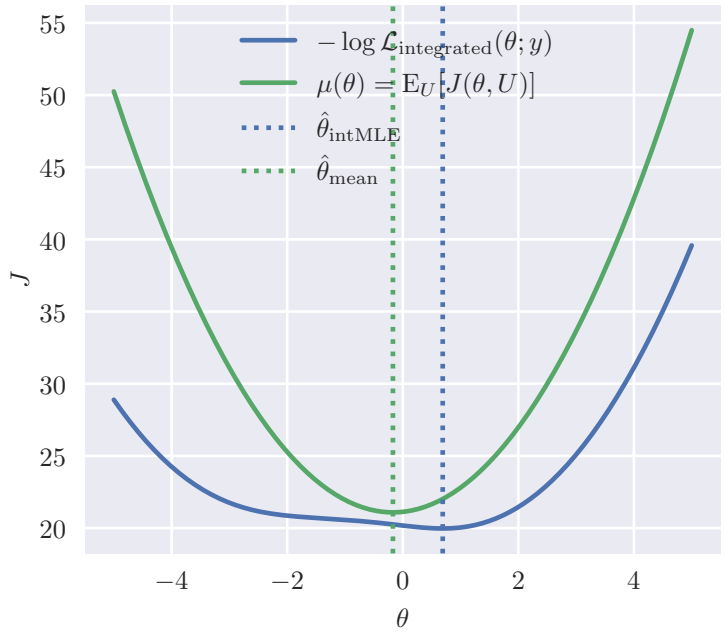

Figure 2.4: Difference between the negative logarithm of the integrated likelihood, and the mean loss of $J = -\log \mathcal{L}$ and the subsequent difference in estimators

A low expected value is to be taken with caution, as it refers to a behaviour *in the long run*. Indeed, the mean values is equivalent to averaging over all the outcomes, but there can be a compensation effect, where "good surprises" balance the "bad surprises". An example is the following problem:

$$J(\theta_1, U) \sim \mathcal{N}(2, 2^2) \tag{2.32}$$
$$J(\theta_2, U) \sim \mathcal{N}(3, 1^2) \tag{2.33}$$

and we have to choose either $\hat{\theta} = \theta_1$ or $\hat{\theta} = \theta_2$. It is clear that $\mathbb{E}_U[J(\theta_1, U)] < \mathbb{E}_U[J(\theta_2, U)]$. However, making the decision $\hat{\theta} = \theta_2$ leads to less extreme values:

$$\mathbb{P}_U[J(\theta_1, U) > 5] = 0.06681 > \mathbb{P}_U[J(\theta_2, U) > 5] = 0.02275 \tag{2.34}$$

Depending on the application, such a behaviour could be prohibitive. This difference in these probabilities is explained by the difference in the variance of the random variable $J(\theta, U)$. Accounting for the variance in the objective function is discussed in Section 2.3.2.b.

2.3.2.b Variance, multiobjective optimization

In Section 2.3.2.a, we used the mean as a measure of the central tendency that we want to minimize. Jointly with the central tendency, information about the dispersion of the random variable may also be relevant, in order to predict how much deviation should be expected around the mean. Let us define the variance of the cost function:

$$\sigma^2(\theta) = \mathbb{V}\mathrm{ar}\left[J(\theta, U)\right] \tag{2.35}$$

and minimizing this variance yields

$$\hat{\theta}_{\mathrm{var}} = \min_{\theta \in \Theta} \sigma^2(\theta) \tag{2.36}$$

As the exact variance computation require the evaluation of an expensive integral, this problem can be tackled using sample averaging, and the minimization problem becomes

$$\min_{\theta \in \Theta} \frac{1}{n_U - 1} \sum_{i=1}^{n_U} \left(J(\theta, u_i) - \mu^{\mathrm{emp}}(\theta)\right)^2 \tag{2.37}$$

Figure 2.5 shows the conditional mean and conditional standard deviation for the cost function J defined Eq. (2.25).

Minimizing the variance is often irrelevant without additional constraints, as it could just point toward really high values of the cost function, but steady with respect to θ. Taking both objectives: low mean value and low variance together to the following multiobjective optimization problem:

$$\min_{\theta \in \Theta} \left(\mu(\theta), \sigma(\theta)\right) \tag{2.38}$$

This problem can be tackled in different ways using multiobjective optimization. To compare θ_1 and θ_2, we can compare component-wise the objective vectors $(\mu(\theta_i), \sigma(\theta_i))$ for $i = 1, 2$. If $\mu(\theta_1) \leq \mu(\theta_2)$ and $\sigma(\theta_1) \leq \sigma(\theta_2)$, θ_2 is said to be *dominated* by θ_1. The Pareto frontier is defined as the set of points in Θ that cannot be dominated by any other points. For points on this front, you cannot decrease further one of the objective without increasing the other. On Section 2.3.2.b is illustrated the Pareto frontier for a multiobjective problem Eq. (2.38). The red point corresponding to θ_1 is dominated by the green point θ_0 on the frontier, but not by the green point of θ_2. A solution of the multiobjective problem can then be chosen within the Pareto frontier.

Instead of finding the Pareto frontier, the multiobjective problem is often "scalarized" by adding the weighted objectives [MA10], provided that such an operation makes sense

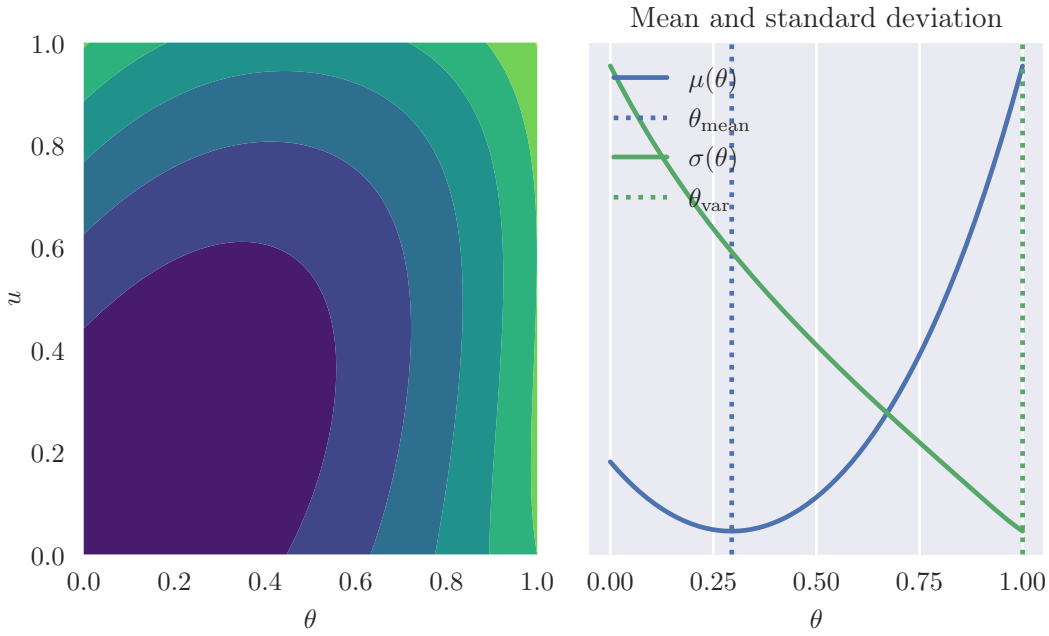

Figure 2.5: Illustration of conditional mean and conditional standard deviation, as a function of θ. Those quantities have been rescaled to share the same range on the right plot.

in regards to the units of the quantities, justifying the use of the standard deviation instead of the variance.

$$\min_{\theta\in\Theta} \lambda\mu(\theta) + (1-\lambda)\sigma(\theta) = \min_{\theta\in\Theta} \lambda\mathbb{E}_U[J(\theta,U)] + (1-\lambda)\sqrt{\mathbb{V}\text{ar}[J(\theta,U)]} \tag{2.39}$$

where $\lambda \in [0,1]$ is chosen to reflect the preference toward one or another objective.

2.3.2.c Higher moments in optimization

Higher moments can also be considered as additional criteria, especially in Portfolio optimization [LYW06, BKJ07]. The skewness coefficient measures the asymmetry in the distribution, and is the (normalized) centered moment of order 3:

$$\text{sk}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{2.40}$$

where $\mu = \mathbb{E}[X]$ and $\sigma = \sqrt{\mathbb{V}\text{ar}[X]}$.

Adding the skewness in the optimization translates to a preference toward a risk-averse or a risk-seeking approach. Indeed, as the main goal is the optimization of a cost function, deviations of the value of the random variable toward lower values is more desirable than deviations toward larger values.

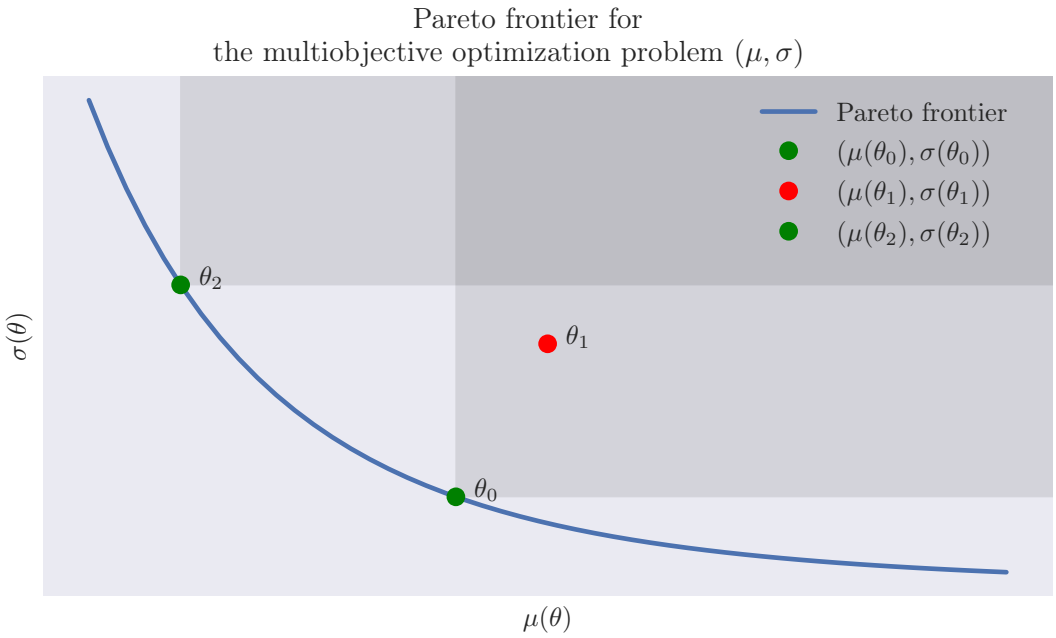

Figure 2.6: Illustration of the Pareto frontier for the multiobjective problem of Eq. (2.38). The shaded regions corresponds to the domain dominated by each points

This is illustrated Fig. 2.7: all three of the random variables displayed have the same mean and variance. If the skewness coefficient is negative, the distribution presents a heavier left tail than right. In other words, a sample taken from this distribution has a higher probability of being a "good surprise". On the other hand, if a big deviation occurs for a sample from a right-skewed distribution, it is more probable to be a large deviation toward large values of the sample space, hence the term "bad surprise".

In order to have a finer tuning on the "risk-averse" or "risk-seeking" properties of the wanted solution, some authors propose to directly minimize with respect to θ the difference between the cdf of the r.v. $J(\theta, U)$ and a target cdf, giving *Horsetail matching* [CJW17, Coo18].

Other extensions have been developed around the cdf of the random variable, especially in portfolio optimization. Indeed, integrating and comparing the cdf allows to introduce a *domination order* between random variables. These concepts of Stochastic Dominance [OR97] are then used to take decisions under uncertainties.

2.4 Regret-based families of estimators

All the methods introduced above required first to eliminate in some sense the dependency on the environmental parameter, in order to transform the random variable $J(\theta, U)$ into an objective that depends solely on θ, and to optimize this deterministic

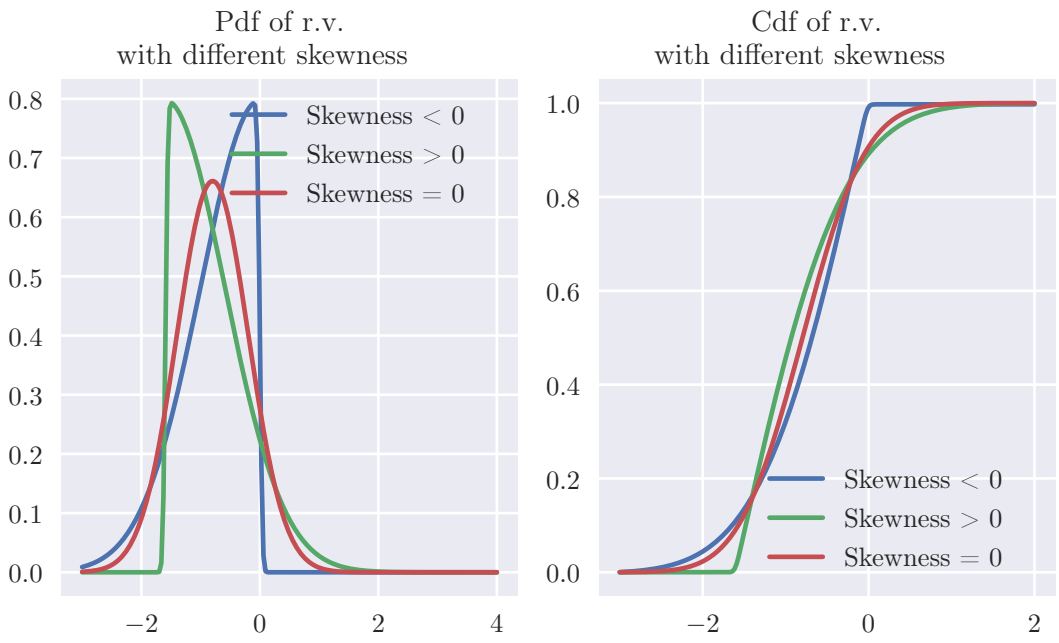

Figure 2.7: Pdf and cdf of random variables with same mean, variance but different skewness

counterpart. For a given $\theta \in \Theta$, this elimination is done by aggregating all the possible outcomes $J(\theta, u)$ when u is a sample of U.

We propose now to reverse these steps, by first optimizing the cost function with respect to θ, and, from the set of minimizers that depend on u obtained, derive an estimator. The rationale behind this permutation is that every situation induced by a realisation u is to be taken separately, quite similarly as Savage's regret introduced Section 2.3.1.c. In turn, this avoids aggregation (and in a sense compensation) between the different u.

The work detailed in this section is largely based on [TAVD20], which was submitted for review in February 2020.

2.4.1 Conditional minimum and minimizer

We assume that U is a continuous random variable, with a compact support.

Definition 2.4.1 – Conditional minimum, minimizer: Let $J : \Theta \times \mathbb{U}$ be a cost function, and let us assume that for each $u \in \mathbb{U}$, $\min_{\theta\in\Theta} J(\theta, u)$ exists and is attained at a unique point. We denote

$$J^*(u) = \min_{\theta\in\Theta} J(\theta, u) \tag{2.41}$$

the *conditional minimum* of J given u, and

$$\theta^*(u) = \arg\min_{\theta\in\Theta} J(\theta, u) \tag{2.42}$$

is defined as the *conditional minimizer*

As u is thought to be a realization of a random variable U, we can consider the two random variables $\theta^*(U)$ and $J^*(U)$. The conditional minimum $J^*(U)$ is then a random variable describing the best performances of the calibration, if we could optimize the cost function for each realization of U.

Similarly, let us assume that the conditional minimizer is well defined for all $u \in \mathbb{U}$. We can study the image random variable through this mapping, that we will denote $\theta^*(U)$. This random variable in itself gives already information on the "identifiability" of a robust estimate, depending on the information carried by its distribution.

Example 2.4.2: Let $\Theta = \mathbb{U} = [0, 1]$, and $U \sim \text{Unif}(\mathbb{U})$, and the following cost functions:

$$J_1(\theta, u) = (1 + u) + (\theta - 0.5)^2 \tag{2.43}$$
$$J_2(\theta, u) = (\theta - u)^2 + 1 \tag{2.44}$$

We have

$$\theta_1^*(u) = \arg\min_{\theta\in\Theta} J_1(\theta, u) = 0.5 \tag{2.45}$$
$$\theta_2^*(u) = \arg\min_{\theta\in\Theta} J_2(\theta, u) = u \tag{2.46}$$

In the first case, $\theta_0 = \arg\min_\theta J(\theta, U)$, so $\theta_1^*(U)$ is a degenerate random variable almost surely equals to θ_0. In other words, the minimizer is not dependent on the value taken by the environmental parameter. The minimum value attained J^* might be dependent though. On the other hand, for J_2, as $\Theta = \mathbb{U}$ and $U \sim \text{Unif}(\mathbb{U})$, $\theta_2^*(U)$ is uniformly distributed on Θ, no value shows a better affinity of being a minimizer than the other.

In general, this random variable cannot be classified as continuous or discrete without further study. However, in the following, we are going to assume that it is a *continuous random variable*. The entropy of the random variable $\theta^*(U)$ can be seen as a measure of the sensitivity of the calibration when the environmental variable varies. Per the continuity assumption, this entropy can be estimated by various methods (see for instance [BDGV97]). This distribution of the minimizers can be used for global optimization, as outlined in [HS11]. Furthermore, the authors provide an analytical expression of the pdf of the minimizers, and the nature of the infinite product is discussed:

$$p_{\theta^*}(\theta) = \int_{\mathbb{U}} p_U(u) \prod_{\substack{\tilde{\theta}\in\Theta \\ \tilde{\theta}\neq\theta}} \mathbb{1}_{\{J(\tilde{\theta},u)>J(\theta,u)\}} \, \mathrm{d}u \tag{2.47}$$

However, except for simple analytical problems, this pdf cannot be obtained analytically, and needs to be estimated.

The estimation of p_{θ^*} can be performed by different methods, depending on the assumptions we can make upon $\theta^*(U)$. Let $\{u_i\}_{1\leq i\leq n_{\text{samples}}}$ be n_{samples} i.i.d. samples of U, and $\{\theta^*(u_i)\}_{1\leq i\leq n_{\text{samples}}}$ the corresponding minimizers, as defined Eq. (2.42). Among the methods of density estimation, one of the easiest to implement and widespread methods is *Kernel Density Estimation* (KDE). Given the samples u_i and the minimizers $\theta^*(u_i)$ for $1 \leq i \leq n_{\text{samples}}$, the isotropic KDE is given by

$$\hat{p}_{\theta^*}(\theta^*) = \frac{1}{n_{\text{samples}} h^{\dim\Theta}} \sum_{i=1}^{n_{\text{samples}}} \mathcal{K}\left(\frac{\theta^* - \theta^*(u_i)}{h}\right) \tag{2.48}$$

where $h > 0$ is the bandwidth (that measure the influence of each sample), and $\mathcal{K}$ is a kernel of dimension $\dim\Theta$, usually defined as the product of one-dimensional kernels $\mathcal{K}_{1\text{D}}$: $\mathcal{K}(\theta) = \prod_{j=1}^{\dim\Theta} \mathcal{K}_{1\text{D}}(\theta_j)$. Several choices of 1D kernels are available, and one of the most common one is the Gaussian Kernel: $\mathcal{K}_{1\text{D}}(\theta_j) = (2\pi)^{-1/2}\exp(-\theta_j^2/2)$. Figure 2.8 shows the estimated density $\hat{p}_{\theta^*}$ using KDE and Scott's rule for the bandwidth [Sco79], along with the histogram of the minimizers.

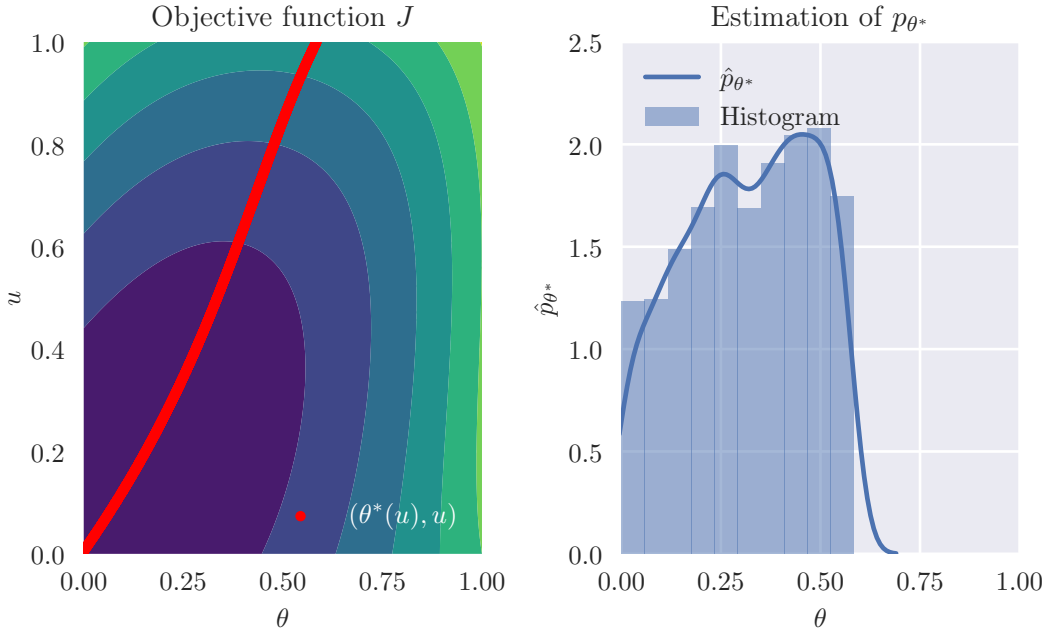

Figure 2.8: Density estimation of the minimizers of J

Finally, when we have an estimation of the density of θ^*, and if it exists, we can compute its mode, that we are going to call the *Most Probable estimator*:

$$\hat{\theta}_{\text{MPE}} = \arg\max_{\theta\in\Theta} p_{\theta^*}(\theta) \tag{2.49}$$

This mode can be sought directly using appropriate algorithms, such as the Mean-shift algorithm [Yiz95], based on the KDE, or clustering methods, such as the Expectation-Maximization algorithm introduced [DLR77].

Choosing $\hat{\theta}_{\mathrm{MPE}}$ means to select the value that is "most often" the minimiser of the cost function. However, we have no indications on its performances when it is *not* optimal, and how often this non-optimality happens. In Section 2.4.2, we are going to introduce the notions of regret (additive and relative), in order to try to be "near optimal" with high probability.

2.4.2 Regret and model selection

In this section, we will first focus on cost functions defined as the negative log-likelihood in order to link the additive regret and the likelihood ratio test.

2.4.2.a Cost as the negative log-likelihood

Let the cost function be the negative log-likelihood:

$$J(\theta, u) = -\log p_{Y|\theta,U}(y \mid \theta, u) = -\log \mathcal{L}(\theta, u) \tag{2.50}$$

We can then link the notion of Wald's regret introduced earlier to the likelihood ratio test:

$$\begin{aligned} r(\theta, u_0) &= J(\theta, u_0) - \min_{\theta' \in \Theta} J(\theta', u_0) = J(\theta, u_0) - J^*(u_0) && (2.51)\\ &= \max_{\theta' \in \Theta} \log \mathcal{L}(\theta', u_0) - \log \mathcal{L}(\theta, u_0) && (2.52)\\ &= -\log \frac{\mathcal{L}(\theta, u_0)}{\max_{\theta' \in \Theta} \mathcal{L}(\theta', u_0)} = -\log \frac{\max_{\theta' \in \{\theta\}} \mathcal{L}(\theta', u_0)}{\max_{\theta' \in \Theta} \mathcal{L}(\theta', u_0)} && (2.53) \end{aligned}$$

As the model $(\mathcal{M}(\cdot, u_0), \Theta)$ is misspecified, using the misspecified likelihood ratio test introduced Section 1.6

For a candidate $\theta \in \Theta$ we can test for the following hypotheses:

- $\mathcal{H}_0$: the model $(\mathcal{M}(\cdot, u_0), \{\theta\})$ is statistically equivalent to $\{\mathcal{M}(\cdot, u_0), \Theta\}$
- $\mathcal{H}_1$: the models are statistically different

regret r is to be compared with half the quantile of the r.v. $X(u_0)$ defined as

$$X(u_0) = \sum_{i=1}^{\dim \Theta} c_i(u_0) \Xi_i \quad \text{with } \Xi_i \sim \chi^2_1 \text{ i.i.d.} \tag{2.54}$$

where $\{c_i(u_0)\}_{1 \leq i \leq \dim \Theta}$ are coefficients linked to the eigenvalues of the Fisher information matrix as evoked in Section 1.6.

The null hypothesis $\mathcal{H}_0$ is rejected at a level $\eta \in]0;1[$ if

$$r(\theta, u_0) = J(\theta, u_0) - J^*(u_0) > \beta \tag{2.55}$$

Where β is half the $1-\eta$ quantile of the random variable $X(u_0)$ defined Eq. (2.54).

Using this rejection region, we can construct a likelihood interval (as defined Eq. (1.72)), which depends on u_0:

$$\mathcal{I}_{\text{Lik}}(u_0; \beta) = \{\theta \in \Theta \mid J(\theta, u_0) - J^*(u_0) \leq \beta\} \tag{2.56}$$

So, for $\theta \in \mathcal{I}_{\text{Lik}}(u_0; \beta)$, the model $(\mathcal{M}, \{\theta\})$ is acceptable at the η-level, per the Likelihood ratio test.

From a computational point of view, the coefficients $\{c_i(u_0)\}$ are hard to obtain, and depend on u_0. A first approximation would be to suppose that $X(u_0) \sim \chi^2_{\dim \Theta}$, *i.e.* to apply the "well-specified" likelihood ratio test. In the more general case, we can choose $\beta > 0$ in a more arbitrary way in order to avoid the computations of the coefficients $\{c_i(u_0)\}_i$, as we are going to see Section 2.4.2.b.

2.4.2.b Interval and probability of acceptability

We assumed before that J was the negative log-likelihood. In the more general case, J represents a loss function, that we want to minimize. The generalization of Eq. (2.56) is what we are calling the *interval of acceptability*.

Definition 2.4.3 – Interval of acceptability: By analogy with the likelihood interval defined Eq. (1.72) and Eq. (2.56), we can construct a set for an arbitrary threshold $\beta \geq 0$ such that

$$\mathcal{I}_\beta(u) = \{\theta \in \theta \mid J(\theta, u) \leq J^*(u) + \beta\} \tag{2.57}$$

As J may not stem from a likelihood, we call $\mathcal{I}_\beta(u)$ the *interval of acceptability*. In other words, we say that $\theta \in \Theta$ is β-acceptable for $U = u$ if $\theta \in \mathcal{I}_\beta(u)$.

Figure 2.9 shows a cost function evaluated for different fixed u_i. The β-acceptable intervals for those environmental variables are plotted below the curves.

Now, for a given θ, we can define the set of $u \in \mathbb{U}$ such that $\theta \in \mathcal{I}_\beta(u)$, i.e.

$$R_\beta(\theta) = \{u \in \mathbb{U} \mid \theta \in \mathcal{I}_\beta(u)\} \tag{2.58}$$

$$= \{u \in \mathbb{U} \mid J(\theta, u) \leq J^*(u) + \beta\} \tag{2.59}$$

This set can be measured with respect to the distribution of U, giving

$$\Gamma_\beta(\theta) = \mathbb{P}_U\left[R_\beta(\theta)\right] \tag{2.60}$$

Loosely speaking, for a given θ, $\Gamma_\beta(\theta)$ is the probability that the model $(\mathcal{M}(\cdot, U), \{\theta\})$ is "statistically equivalent" to the "full model" $(\mathcal{M}(\cdot, U), \Theta))$, at a certain level linked to the value of β.

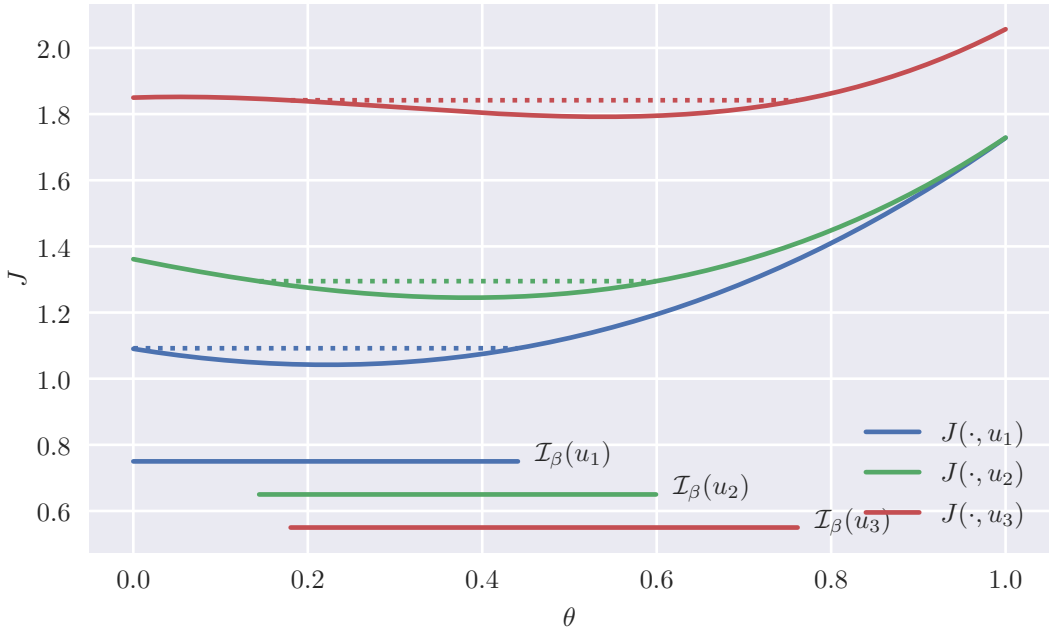

Figure 2.9: Different acceptable regions corresponding to different $u \in \mathbb{U}$

Figure 2.10 shows the regions of acceptability for different β, and the associated Γ_β.

Γ_β is then maximized with respect to its argument, in order to get the value θ which has the highest probability of being acceptable at the level β. For different $\beta \geq 0$, we can define the family of additive-regret estimators.

Definition 2.4.4 – Additive-regret family of estimators: For $\beta \geq 0$, we define the family of robust estimators as the maximizers of Eq. (2.60):

$$\left\{ \hat{\theta}_{\mathrm{AR},\beta} = \arg\max_{\theta \in \Theta} \Gamma_\beta(\theta) \mid \beta > 0 \right\} \tag{2.61}$$

Among this family of estimators, we can then choose a particular value, either by setting a threshold β arbitrarily, or by choosing it so that the probability of being acceptable $\max \Gamma_\beta$ reaches a particular value. This will be discussed later in Section 2.4.4.

2.4.3 Relative-regret

2.4.3.a Absolute and relative error

We examined before regret that can be qualified as *additive* as this is the difference betdifference between J and J^* that is compared to fixed thresholds. However, we can argue that the relative magnitude of the cost function has an importance in the comparison. For illustration purposes, let us consider the situation described Table 2.2, with $\Theta = \{\theta_1, \theta_2\}$ and $\mathbb{U} = \{u_1, u_2\}$, and $\mathbb{P}[U = u_{1,2}] = 1/2$.

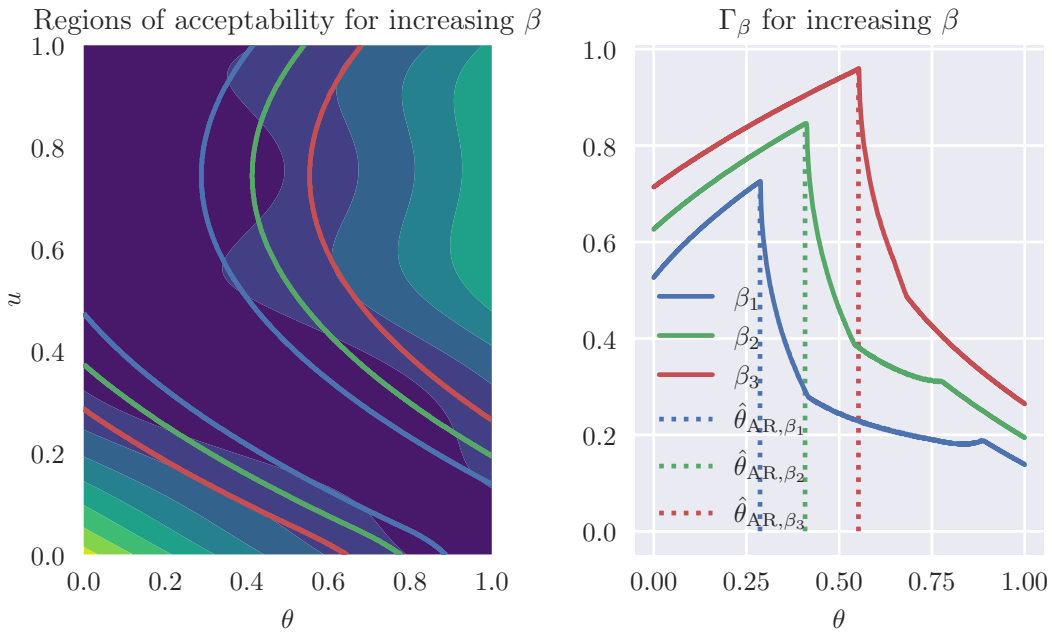

Figure 2.10: Boundaries of regions of acceptability for increasing β, and Γ_β. The colored lines on the left plot are the boundaries of those regions

J	u_1	u_2	$\mathbb{E}[J(\cdot, U)]$
θ_1	10000	110	5055
θ_2	10100	10	5055

$J - J^*$	u_1	u_2
θ_1	0	100
θ_2	100	0

$\frac{J-J^*}{J^*}$	u_1	u_2
θ_1	0	10
θ_2	0.01	0

Table 2.2: Illustration of a cost function, expected loss, additive regret and relative error

In this situation, for both u_1 and u_2, the maximal additive regret is $\max_\theta J(\theta, u) - J^*(u) = 100$, so no clear preference could be inferred toward one or another value. However, choosing θ_1 over θ_2 means to choose to improve the performance of an already pretty bad situation (10000 instead of 10100), while increasing tenfold the loss for the situation $U = u_2$.

From the example developed Table 2.2, an alternative to the additive regret may be considered, as the difference in magnitude of the cost function between $U = u_1$ and $U = u_2$ is probably due to the effect of a misspecification. So to take into account this difference, we are now going to consider the *relative regret* J/J^*, instead of the *absolute regret* $J - J^*$, and derive a family of estimators in a similar fashion.

2.4.3.b Relative-regret estimators family

Analogously as the additive regret defined before we are going first to define the notions of acceptability in the case of the relative regret.

Definition 2.4.5 – α-acceptability: Let $\alpha \geq 1$. A point (θ, u) is said to be α-acceptable if $J(\theta, u) \leq \alpha J^*(u)$. We define the α-acceptable interval as

$$\mathcal{I}_\alpha(u) = \{\theta \in \Theta \mid J(\theta, u) \leq \alpha J^*(u)\} \tag{2.62}$$

Then, for a given α and θ, we can define the set of $u \in \mathbb{U}$ such that θ is α-acceptable:

$$R_\alpha(\theta) = \{u \in \mathbb{U} \mid \theta \in \mathcal{I}_\alpha(u)\} = \{u \in \mathbb{U} \mid J(\theta, u) \leq \alpha J^*(u)\} \tag{2.63}$$

$R_\alpha(\theta)$ is a measurable subset of $\mathbb{U}$, and by integrating this set with respect to $\mathbb{P}_U$, we get

$$\Gamma_\alpha(\theta) = \mathbb{P}_U\left[R_\alpha(\theta)\right] \tag{2.64}$$

the probability of being α-acceptable. Using this function, we can define an estimator as the value which maximizes this probability. And by varying the threshold α, we get the family of Relative-regret estimators.

Definition 2.4.6 – Relative-regret family of estimators: Given α, the value of θ that maximizes the probability of being α-acceptable is called the relative-regret (RR) estimator $\hat{\theta}_{\mathrm{RR},\alpha}$.

We define the family of relative regret estimators as the set of those estimators:

$$\left\{\hat{\theta}_{\mathrm{RR},\alpha} = \arg\max_{\theta \in \Theta} \Gamma_\alpha(\theta) \mid \alpha > 1\right\} \tag{2.65}$$

For different increasing α, the corresponding regions of acceptability were represented Fig. 2.11, along with the functions Γ_α.

Among those estimators and the associated quantities, two limiting cases appear. One particular choice is to set α to 1. In this case, we have

$$\mathcal{I}_1(u) = \{\theta^*(u)\} \tag{2.66}$$

$$R_1(\theta) = \{u \in \mathbb{U} \mid J(\theta, u) = J^*(u)\} = \{u \in \mathbb{U} \mid \theta = \theta^*(u)\} \tag{2.67}$$

We have then that $\Gamma_1(\theta)$ is non-zero if the set $R_1(\theta)$ has non-zero measure with respect to $\mathbb{P}_U$. In other words, $\Gamma_1(\theta)$ is non-zero if θ is the minimizer of $J(\cdot, u)$ for a non-negligible subset of $\mathbb{U}$. If we consider that Θ is a discrete space (due to a discretization for instance), $\theta^*(U)$ is a discrete random variable. $\Gamma_1(\theta)$ is then the probability mass function (discrete parallel of the pdf) of the discrete r.v. $\theta^*(U)$. In this case, $\hat{\theta}_{\mathrm{RR},\alpha=1} = \hat{\theta}_{\mathrm{MPE}}$. A similar argument can be made for the additive regret and $\beta = 0$.

We can see that the thresholds act like a relaxation of the optimality condition, as for $\alpha = 1$ and $\beta = 0$, we are measuring the probability of being optimal, while increasing those values means to measure the probability of being nearly optimal.

Another choice is to set the threshold large enough so that the probability of being acceptable reaches a unique maximum, being 1. In this situation, the regret is bounded almost surely. Let β_{inf} and α_{inf}, which verify

$$\beta_{\text{inf}} = \inf\{\beta \geq 0 \mid \max \Gamma_\beta = 1\} \text{ and } \alpha_{\text{inf}} = \inf\{\alpha \geq 1 \mid \max \Gamma_\alpha = 1\} \tag{2.68}$$

To put it differently, α_{inf} and β_{inf} are the smallest thresholds where there exists a value in Θ acceptable almost surely. This value shares similarities with the minimizer of Savage's regret: $\hat{\theta}_{\text{rWC}}$ introduced Section 2.3.1.c, as it minimizes almost surely (*i.e.* for all u in a non-negligible set) the regret, either additive or relative.

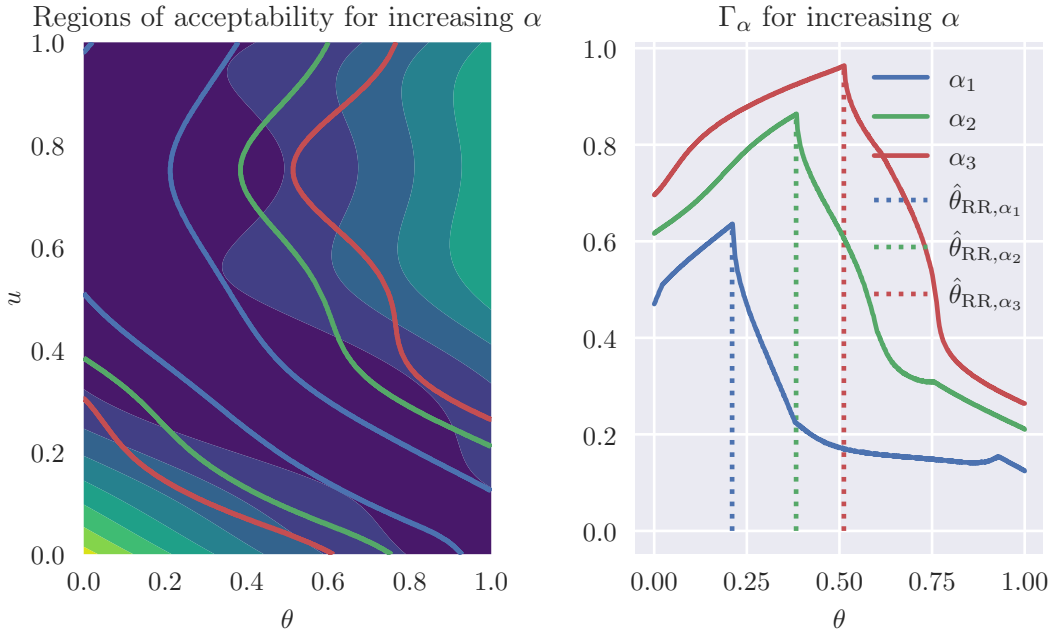

Figure 2.11: Regions of acceptability for relative regret and increasing α. The colored lines on the left plot are the boundaries of those regions

For both regret-based approach, the interval of acceptability at a fixed u grows with the threshold, and the sharper the minimum is (in the sense of a large curvature), the faster it grows. A more telling illustration of the differences between relative and additive regret is shown Fig. 2.12. The regions of acceptability of a cost function J have been plotted, along with an interval $\mathcal{I}(u)$ and the region R for a given θ for both regrets. For u around 0, J is quite flat, but also has very low values. In this case, the interval $\mathcal{I}_\beta(u)$ is also large. But for u around 1, the cost function presents higher values, and a sharper minimum (*i.e.* a higher curvature). Additive regret in this case puts stronger confidence on the value of the parameter as indicated by the smaller interval of acceptability.

For the relative regret, the situation is reversed. Although sharp, the large value attained by the minimum $J^*(u)$, for u around 1 leads to a large interval of acceptability, meaning that we can deviate a bit more from this minimum, as the situation $u \approx 1$ is already pretty bad. For $u \approx 0$, which is close to the global minimum of J, the interval is smaller, as this criterion does not favour a large deviation from this global minimum.

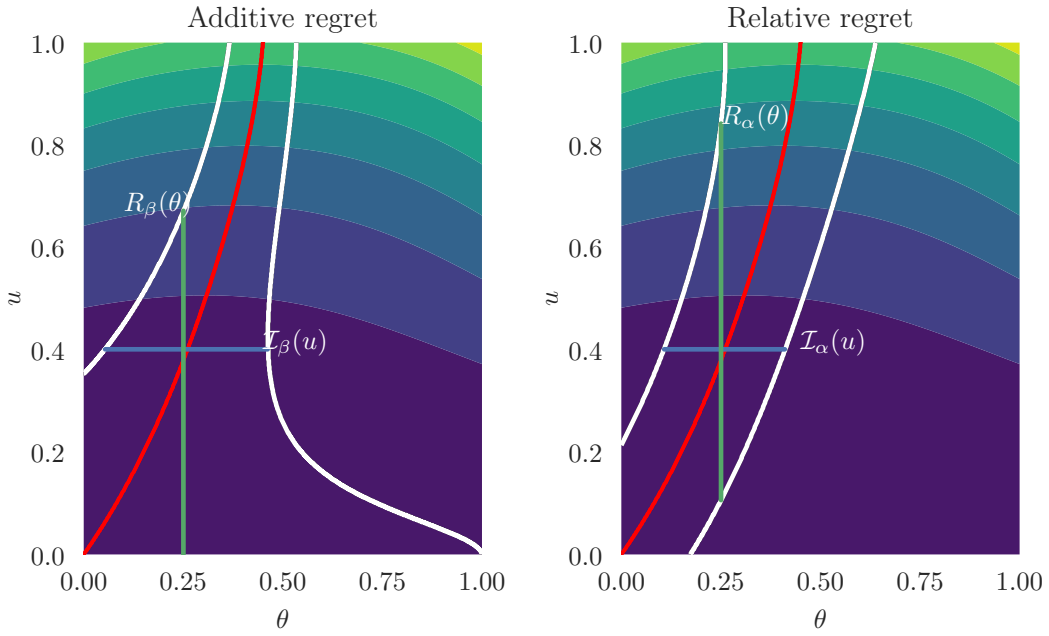

Figure 2.12: Comparison of the regions of acceptability for additive and relative regret

2.4.4 The choice of the threshold

We are now going to focus exclusively on the relative-regret, and the associated treshold α, but similar arguments can be made for the absolute regret and the threshold β.

In order to have an insight on the potential robustness of an estimator $\hat{\theta}_{\mathrm{AR},\alpha}$ both values can be studied together: the threshold α and the maximal probability reached $\max \Gamma_\alpha$

Finding a relevant threshold can be a thorny issue, especially with no further information on J. Setting it too large will lead to large α-acceptable intervals, and Γ_α may reach 1 for several different values. On the other hand, choosing a threshold too small may give a maximal acceptability probability too low to assess the robustness of the chosen solution. summarize, we can contemplate three starting points:

- Set a probability $0 < p \leq 1$ we want to reach, and define $\alpha_p = \inf_{\alpha \geq 1}\{\max_{\theta \in \Theta} \Gamma_\alpha(\theta) \geq p\}$, so the problem can be thought as the estimation and the optimization of quantile of a particular random variable.
- Set a relaxation parameter $\alpha > 1$. From this, the maximal probability is $p_\alpha = \max_{\theta \in \Theta} \Gamma_\alpha = \max_{\theta \in \Theta} \mathbb{P}_U \left[J(\theta, U) \leq \alpha J^*(U) \right]$. This task can be seen as the estimation and optimization of a well chosen probability.
- Study the evolution of one quantity with respect to the other, in order to find a balance between the probability of being acceptable, and the relaxation needed to reach it.

Specific techniques may be applied in order to perform the first two approaches efficiently, and some will be introduced in the next chapter. The third one however requires the knowledge of the cost function on the whole joint space $\Theta \times \mathbb{U}$, in order to compute $\max \Gamma$ for a various number of thresholds, thus may not be adapted for costly computer simulations.

2.5 Partial Conclusion

As shown through this chapter, when optimizing under uncertainties, a lot of criteria can be defined in order to satisfy the idea of *robustness*, depending on the interpretation of this term. A summary of those introduced can be found Table 2.3. Some criteria are commonly encountered in optimization under uncertainty, such as expected loss minimization. We introduce also new families of estimators, which aim at maximizing a probability based on the regret, either additive or relative.

Objective name	Objective to *minimize* wrt θ
Profile Likelihood	$-\log\left(\max_{u\in\mathbb{U}} p_{Y\mid\theta,U}(y \mid \theta, u)\right)$
Integrated Likelihood	$-\log \int_{\mathbb{U}} p_{Y\mid\theta,U}(y \mid \theta, u)\,\mathrm{d}u = -\log p_{Y\mid\theta}(y \mid \theta)$
Marginal maximum a posteriori	$-\log p_{\theta\mid Y}(\theta \mid y)$
Global Optimum	$\min_{u\in\mathbb{U}} J(\theta, u)$
Worst-case	$\max_{u\in\mathbb{U}} J(\theta, u)$
Regret worst-case	$\max_{u\in\mathbb{U}} \{J(\theta, u) - \min_{\theta'\in\Theta} J(\theta', u)\}$
Mean	$\mathbb{E}_U[J(\theta, U)]$
Mean and variance	$\lambda\mathbb{E}_U[J(\theta, U)] + (1-\lambda)\sqrt{\mathbb{V}\mathrm{ar}_U[J(\theta, U)]}$
Most Probable Estimate	$-\log p_{\theta^*}(\theta)$
Additive-regret	$-\mathbb{P}_U\left[J(\theta, U) \leq J^*(U) + \beta\right] = -\Gamma_\beta(\theta)$
Relative-regret	$-\mathbb{P}_U\left[J(\theta, U) \leq \alpha J^*(U)\right] = -\Gamma_\alpha(\theta)$

Table 2.3: Summary of single objective robust estimators

Obviously, other criteria can be defined, that satisfy other robustness requirements. Furthermore, we did not treat the possibility of combining some of those objectives by using them to set constraints. An example is the minimization of the variance, under the constraint that the mean value does not exceed a certain threshold T:

$$\begin{aligned} &\min \mathbb{V}\mathrm{ar}_U\left[J(\theta, U)\right] \\ &\text{s.t. } \mathbb{E}_U\left[J(\theta, U)\right] \leq T \end{aligned}$$

All of the criteria introduced above require costly numerical procedure, such as integration and optimization. Solving these robust estimation problems is then expensive in term of computer resources, as one would need to run the forward model a very large number of times, in order to get accurate numerical integration or optimization. In the next chapter we will discuss methods based on surrogate modelling, that can be used to

solve efficiently such problems, in order to avoid make the best use of evaluations of the numerical model $\mathcal{M}$ on the space $\Theta \times \mathbb{U}$.

CHAPTER 3

ADAPTATIVE DESIGN ENRICHMENT FOR CALIBRATION USING GAUSSIAN PROCESSES

Contents

In this chapter, we will focus on the use of *surrogate models* to solve robust optimization problems, according to some of the criteria introduced in the previous chapter.

3.1 Computational bottleneck, curse of dimensionality and surrogate models

Numerical models are usually very expensive to run in terms of computer resources. Indeed, for most realistic physical simulations, the programs have to solve systems of PDEs over large grids. Even though the computations are optimized and parallelized to take best advantage of high-performance computers, the time required to compute the quantities of interest may range from a few seconds to days. In that sense, methods that require a large number of runs of the model for exhaustivity should be avoided. Also,

Another common issue encoutered is the

3.1.0.a Dimension Reduction

[BHRV17, Rib18]

3.1.1 Surrogate models

In this chapter, we will focus exclusively on Kriging, or Gaussian Process regression. In this chapter, after defining the usual kriging equations for Gaussian Process Regression, we are going to introduce a few classic and useful criteria in Section 3.3 for global optimization and/or exploration of an unknown function f. Afterwards, we are going to develop on the case of robust optimisation, by splitting the input space in Θ and $\mathbb{U}$. We are then going to introduce other strategies to solve for robust optimisation criteria, as introduced previously.

3.2 Gaussian process regression

In the following, we will introduce a generic function f, that maps a space $\mathbb{X}$ to $\mathbb{R}$. Depending on the application, $\mathbb{X} = \Theta$ or $\mathbb{X} = \Theta \times \mathbb{U}$. This function is unknown, and supposedly expensive to evaluate, but it has already been evaluated on a set of points x_i [RW06]

3.2.1 Random processes

Let us assume that we have a map f from a p dimensional space to $\mathbb{R}$:

$$\begin{array}{rccc} f: & \mathbb{X} \subset \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) \end{array} \tag{3.1}$$

This function is assumed to have been evaluated on a design of n points, $\mathcal{X} = \{(x_i, f(x_i))\}_{1 \leq i \leq n}$, called the *initial design* For notational simplicity, we write $x \in \mathcal{X}$ if $(x, f(x)) \in \mathcal{X}$. As this function is unknown, there is (epistemic) uncertainty on the values outside of the initial design. This uncertainty can be reduced by directly evaluating the function. This uncertainty is modelled by random processes as defined in the following

Definition 3.2.1 – Random process: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathbb{X} \subset \mathbb{R}^p$. A random process Z is a collection of random variables indexed on $\mathbb{X}$, so for each $x \in \mathbb{X}$, $Z(x)$ is a random variable:

$$\begin{array}{rcl} Z : \mathbb{X} & \longrightarrow & (\Omega \to \mathbb{R}) \\ x & \longmapsto & Z(x) \end{array} \tag{3.2}$$

A sample from this random process, that is $Z(\cdot)(\omega)$ for $\omega \in \Omega$ will be shortened as $Z(\cdot, \omega)$ for notational purpose, and is called a *sample trajectory*, or a *sample path*.

From a Bayesian point of view, such a random process Z acts as a prior on the function f, or in other words f is seen as a particular sample path of Z. Evaluating the function at an additional point $x \notin \mathcal{X}$ provides new information on the random process, and we can update our belief on f. In this work, we are going to focus exclusively on a specific type of random process, namely, the Gaussian process (abbreviated as GP), but other types of random process can be encountered in the literature: Student t-processes in [SWG14] are introduced as alternatives to GP, or various graphical models such as Gaussian and Markov Random Fields in [Bis06, Li09].

Definition 3.2.2 – Gaussian process: Let Z be a random process on $\mathbb{X}$, i.e. a collection of random variables indexed by $\mathbb{X}$. It is defined as a Gaussian process if any finite number of those random variables have a multivariate joint Gaussian distribution. In that case, Z is uniquely defined by its mean function m_Z and its covariance function C_Z:

$$m_Z(x) = \mathbb{E}\left[Z(x)\right] \tag{3.3}$$

$$C_Z(x, x') = \mathbb{C}\mathrm{ov}[Z(x), Z(x')] \tag{3.4}$$

and we write $Z \sim \mathrm{GP}(m_Z, C_Z)$

Based on the initial design $\mathcal{X}$, we can construct the mean function m_Z, that acts as a surrogate for the unknown function f. Similarly,

3.2.2 Linear Estimation

Given a random process Z as a prior on f, we want to construct a surrogate $\hat{Z}$, using the intial design $\mathcal{X} = \{(x_i, f(x_i)\}_{1 \leq i \leq n}$. This surrogate will be constructed as a linear

estimation: A linear estimation $\hat{Z}$ of f at an unobserved point $x \notin \mathcal{X}$ can be written as

$$\hat{Z}(x) = \left[w_1 \dots w_n\right] \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \mathbf{W}^T f(\mathcal{X}) = \sum_{i=1}^{n} w_i(x) f(x_i) \tag{3.5}$$

Using those *kriging weights* $\mathbf{W}$, a few additional conditions must be added, in order to obtain the Best Linear Unbiased Estimator:

- Non-biased estimation: $\mathbb{E}[\hat{Z}(x) - Z(x)] = 0$
- Minimal variance: $\min\ \mathbb{E}[(\hat{Z}(x) - Z(x))^2]$

The non-biasedness condition using Eq. (3.5) can be rewritten

$$\mathbb{E}[\hat{Z}(x) - Z(x)] = 0 \iff m\left(\sum_{i=1}^{n} w_i(x) - 1\right) = 0 \iff \sum_{i=1}^{n} w_i(x) = 1 \iff \mathbf{1}^T\mathbf{W} = 1 \tag{3.6}$$

For the minimum of variance, we introduce the augmented random vectors $\mathbf{Z}_n(x) = [Z(x_1), \dots Z(x_n), Z(x)]$ and $\mathbf{Z}_n = [Z(x_1), \dots Z(x_n)]$, and the variance can be expressed as:

$$\mathbb{E}[(\hat{Z}(x) - Z(x))^2] = \mathbb{C}\text{ov}\left[[\mathbf{W}^T, -1] \cdot \mathbf{Z}_n(x)\right] \tag{3.7}$$

$$= [\mathbf{W}^T, -1]\mathbb{C}\text{ov}\left[\mathbf{Z}_n(x)\right][\mathbf{W}^T, -1]^T \tag{3.8}$$

In addition, we have

$$\mathbb{C}\text{ov}\left[\mathbf{Z}_n(x)\right] = \begin{bmatrix} \mathbb{C}\text{ov}\left[\mathbf{Z}_n^T\right] & \mathbb{C}\text{ov}\left[\mathbf{Z}_n^T, Z(x)\right] \\ \mathbb{C}\text{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T & \mathbb{V}\text{ar}\left[Z(x)\right] \end{bmatrix} \tag{3.9}$$

Once expanded, the kriging weights solve then the following optimisation problem:

$$\min_{\mathbf{W}}\ \mathbf{W}^T\mathbb{C}\text{ov}\left[\mathbf{Z}_n\right]\mathbf{W} + \mathbb{V}\text{ar}\left[Z(x)\right] \tag{3.10}$$

$$- \mathbb{C}\text{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T\mathbf{W} - \mathbf{W}^T\mathbb{C}\text{ov}\left[\mathbf{Z}_n^T, Z(x)\right] \tag{3.11}$$

$$\text{s.t. } \mathbf{1}^T\mathbf{W} = 1 \tag{3.12}$$

This leads to

$$\begin{bmatrix} \mathbf{W} \\ m \end{bmatrix} = \begin{bmatrix} \mathbb{C}\text{ov}\left[\mathbf{Z}_n\right] & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{C}\text{ov}\left[\mathbf{Z}_n^T, Z(x)\right]^T \\ 1 \end{bmatrix} \tag{3.13}$$

$$= \begin{bmatrix} C(x_1, x_1) & \cdots & C(x_1, x_n) & 1 \\ C(x_2, x_1) & \cdots & C(x_2, x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(x_n, x_1) & \cdots & C(x_n, x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} C(x_1, x) \\ C(x_2, x) \\ \vdots \\ C(x_n, x) \\ 1 \end{bmatrix} \tag{3.14}$$

and

$$\hat{Z}(x) = \begin{bmatrix} C(x_1,x) & C(x_2,x) & \dots & C(x_n,x) \end{bmatrix} \left(\begin{bmatrix} C(x_1,x_1) & \cdots & C(x_1,x_n) \\ C(x_2,x_1) & \cdots & C(x_2,x_n) \\ \vdots & \ddots & \vdots \\ C(x_n,x_1) & \cdots & C(x_n,x_n) \end{bmatrix}^{-1} \right)^T \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix} \tag{3.15}$$

One main interesting point of GP regression, is that more than the surrogate $\hat{Z} = m_Z$, we have a measure of the uncertainty on the estimation:

$$Z(x) \sim \mathcal{N}\left(m_Z(x), \sigma_Z^2(x)\right) \qquad \text{with } \sigma_Z^2(x) = C_Z(x,x) \tag{3.16}$$

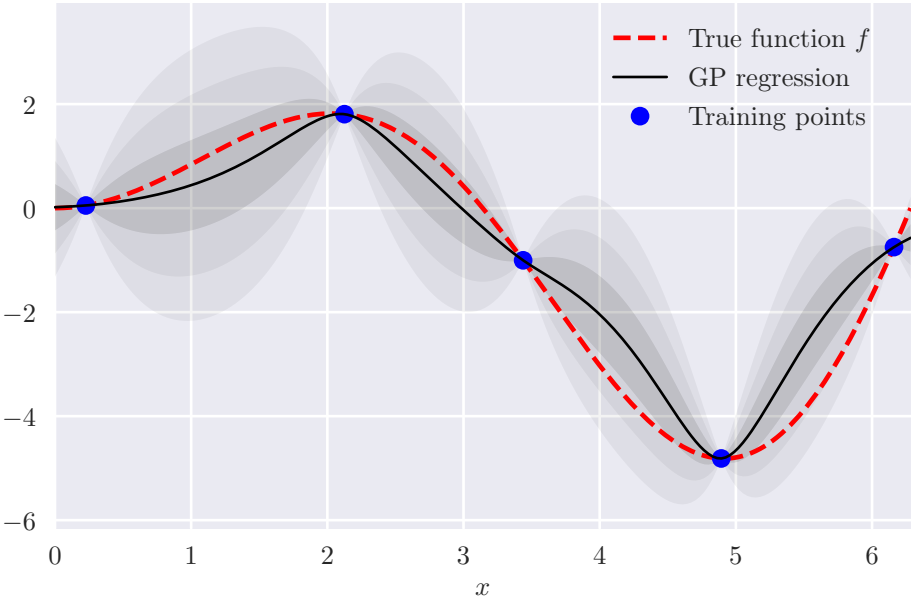

Figure 3.1: Example of a Gaussian Process, as a surrogate for a function f evaluated at 4 inputs. The shaded regions correspond to the regions $m_Z \pm i \cdot \sigma_Z$ for $i = 1, 2, 3$.

3.2.3 Covariance functions

In the previous section, we described the equations to solve to get the surrogate $\hat{Z}$ of f based on GP. The coefficients of the linear estimation are based on the covariance function C_Z. A covariance is said to be stationary, if for all $x, x' \in \mathbb{X}$, the covariance of the GP between those two points depends only on the difference $h = x - x' = (h_1, \cdots, h_{\dim \mathbb{X}})$. In that case, we will write $C_Z(x, x') = C_Z(h)$ For multidimensional problems, covariance functions are usually chosen as the product of $1D$ covariance functions:

$$C_Z(h) = s^2 \prod_{i=1}^{\dim \Theta} C_i(h_i; l_i) \tag{3.17}$$

These covariance functions introduce an additional parameter l of dimension $\dim \mathbb{X}$, and a variance parameter s^2. l called the *length scale*, that measure the influence of each variable on its vicinity. If the length scales are all equals, the covariance kernel is said *isotropic*. Otherwise, the kernel is *anisotropic*.

A few common stationary $1D$-covariance functions are introduced Table 3.1.

Name	$C(h;l)$	Regularity of sample paths
Gaussian	$\exp\left(-\frac{h^2}{2l^2}\right)$	C^∞
Exponential	$\exp\left(-\frac{\lvert h\rvert}{l}\right)$	C^0
Matérn 3/2	$\left(1+\sqrt{3}\frac{h}{l}\right)\exp\left(-\sqrt{3}\frac{h}{l}\right)$	C^1
Matérn 5/2	$\left(1+\sqrt{5}\frac{h}{l}+\frac{5}{3}\frac{h^2}{l^2}\right)\exp\left(-\sqrt{5}\frac{h}{l}\right)$	C^2

Table 3.1: Common covariance functions

One main difference that motivates one or the other covariance function is the assumption upon the regularity of the sample paths. For example, if the unknown function f is assumed to be infinitely differentiable, a Gaussian kernel is suited for the modelling. One common choice is the Matérn kernel of order 5/2, so that the samples paths are twice-differentiable.

Those $(\dim \mathbb{X}+1)$ hyperparameters have to be estimated based on the training set $\mathcal{X}$. This is usually done by MLE (see for instance [RBGH19]), or by cross-validation (see [GDB+09]).

3.2.4 Initial design

3.2.5 Gaussian Process validation

3.3 Stepwise Enrichment strategies for Gaussian Processes

For a unknown function f, a GP is initially constructed based on a design $\mathcal{X} = \{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$, that consists of n points of $\mathbb{X}$, and their corresponding evaluations. This GP is denoted $Y \mid \mathcal{X}$ and is defined as

$$Y \mid \mathcal{X} \sim \mathcal{N}(m_{Y|\mathcal{X}}(x), \sigma^2_{Y|\mathcal{X}}(x)) \tag{3.18}$$

For notational convenience, the conditioning with respect to $\mathcal{X}$ will be omitted if the experimental design is clear from the context.

Stepwise Uncertainty Reduction is based on the construction of a measure of uncertainty, which is problem dependent.

is to define a criterion, say κ_n, that measures in a way the uncertainty upon a certain objective associated with the GP and f, and to maximize this criterion, in order to select

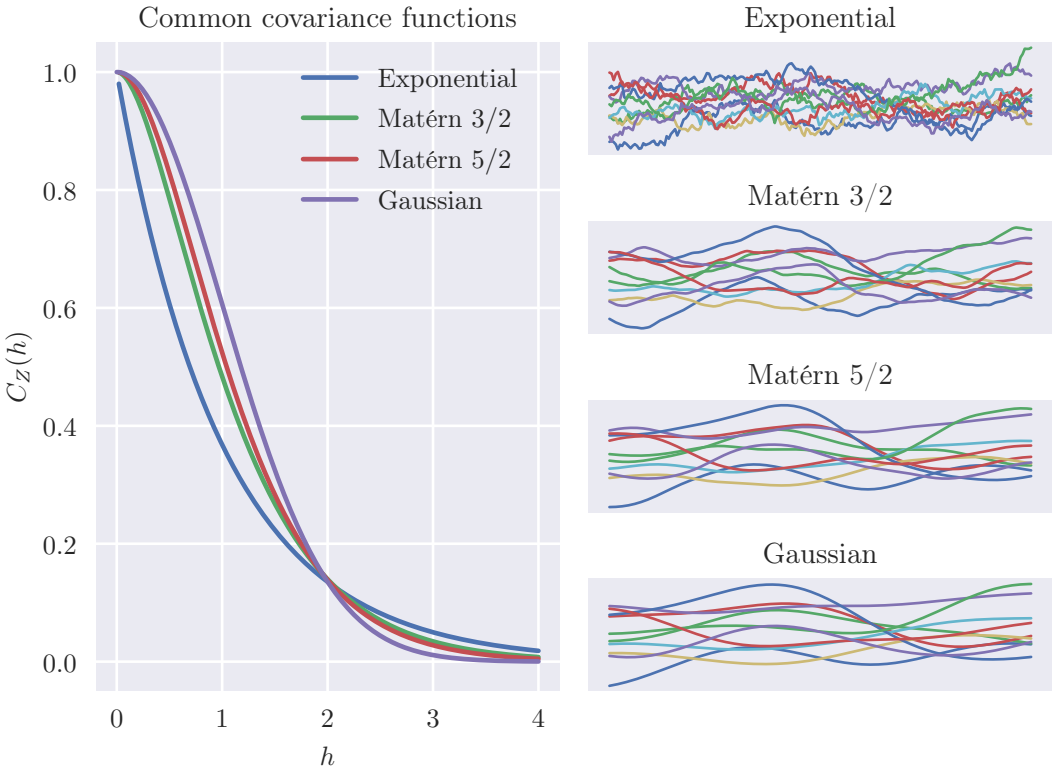

Figure 3.2: Common covariance functions for GP regression. The right plots show (unconditioned) sample paths for those different covariance functions with same length scale.

the next point:

$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \kappa_n(x) = \arg\max_{x \in \mathbb{X}} \kappa(x; Y \mid \mathcal{X}) \tag{3.19}$$

3.3.1 Exploration based criteria

We are going to introduce first some common criteria of enrichment, that aim at exploring the input space $\mathbb{X}$.

3.3.1.a Maximum of variance

A measure of uncertainty on the GP is $\max_{x \in \mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)$, the maximum value of the prediction variance on the space. A simple criterion is to select and evaluate the point corresponding to this maximum of variance:

$$x_{n+1} = \arg\max_{x \in \mathbb{X}} \kappa_n(x) = \arg\max_{x \in \mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x) \tag{3.20}$$

Algorithm 1 SUR strategy

Require: Initial design $\mathcal{X}_0$, criterion function κ
  Fit Y, a GP using the design $\mathcal{X}_0$
  $n \leftarrow 0$
  while Stopping criterion not met or max budget evaluation not reached **do**
    $x_{n+1} \leftarrow \arg\max_{x \in \mathbb{X}} \kappa(x; Y \mid \mathcal{X}_n)$
    Evaluate $f(x_{n+1})$
    $\mathcal{X}_{n+1} \leftarrow \mathcal{X}_n \cup \{(x_{n+1}, f(x_{n+1}))\}$
    $n \leftarrow n + 1$
  end while

This criterion by its simplicity is easy to implement, as the prediction variance is cheap to compute given a GP, and does not depend directly on the evaluations of the function $f(x_i)$, uniquely on the distance between the inputs points and the covariance parameters.

3.3.1.b Integrated Mean Square Error

[SSW89] The prediction variance is directly given by $\sigma^2_{Y|\mathcal{X}}$ and represents the uncertainty on the Gaussian regression. To summarize this uncertainty on the whole space $\mathcal{X}$, we define the Integrated Mean Square Error (IMSE) as

$$\text{IMSE}(Y \mid \mathcal{X}) = \int_{\mathbb{X}} \sigma^2_{Y|\mathcal{X}}(x)\,\mathrm{d}x \tag{3.21}$$

For practical reasons, we can consider to integrate the MSE only on a subset $\mathfrak{X} \subset \mathcal{X}$ that yields

$$\text{IMSE}_{\mathfrak{X}}(Y \mid \mathcal{X}) = \int_{\mathcal{X}} \sigma^2_{Y|\mathcal{X}}(x)\mathbb{1}_{\mathfrak{X}}(x)\,\mathrm{d}x = \int_{\mathfrak{X}} \sigma^2_{Y|\mathcal{X}}(x)\,\mathrm{d}x \tag{3.22}$$

Unfortunately, exact evaluation of this integral is impossible, so it needs to be approximated using numerical integration, such as Monte-carlo or quadrature rules:

$$\text{IMSE}_{\mathfrak{X}}(Y \mid \mathcal{X}) \approx \sum_{i=1}^{n_{\text{quad}}} w_i \sigma_{Y|\mathcal{X}}(x_i) \tag{3.23}$$

For a given $x \in \mathbb{X}$ and an outcome $y = f(x) \in \mathbb{Y}$, the augmented design is defined as $\mathcal{X} \cup \{(x, y)\}$, and the IMSE of the augmented design is $\text{IMSE}(Y \mid \mathcal{X} \cup \{(x, y)\})$. Before the actual experiment though, y is unknown, but we can model it by its distribution given by the GP (per Eq. (3.18)). So for a given candidate x, the mean prediction error we will get when evaluating x is given by

$$\mathbb{E}_{Y(x)}\Big[\,\text{IMSE}\left(Y \mid \mathcal{X} \cup \{(x, Y(x))\}\right)\Big] \tag{3.24}$$

where the expectation is to be taken with respect to the random variable $Y(x)$. As each scenario requires to fit a GP, and to compute the IMSE, a precise evaluation is quite expensive. A strategy ffound for instance in [VVW06] is to take M possible outcomes

for $Y(x)$, corresponding to evenly spaced quantiles of the its distribution. It is maybe important to note that the hyperparameters of the GP should not be reevaluated when augmenting the design, in order to get comparable values for the IMSE.

To enrich the design with the best point, that reduces the most the expected prediction error, a simple 1-step strategy is to minimize the expectation of Eq. (3.24).

$$x_{n+1} = \arg\min_{x \in \mathbb{X}} \mathbb{E}_{Y(x)} \Big[\mathrm{IMSE}\big(Y \mid \mathcal{X} \cup \{(x, Y(x))\}\big) \Big] \tag{3.25}$$

Et si au lieu d'estimer l'espérance, on choisit un échantillon. Stochastic simulation ?

3.3.2 Optimization oriented criteria

The criteria we detailed above aim at reducing the epistemic uncertainty modelled through the Gaussian Process. In other words, we try to improve our knowledge on the unknown function globally. We are now going to evoke a few criteria which are driven by the global optimization of the function.

Those methods usually aim at striking a balance between exploration and *intensification*. We covered exclusive exploration in Section 3.3.1. Let f be the unknown function, and Y be a GP constructed based on an initial design $\mathcal{X} = \{(x_i, f(x_i))\}$.

3.3.2.a Probability of improvement

We are first going to introduce the probability of improvement PI, which is the probability that the GP is smaller than a threshold $f_{\min}$. Due to the Gaussian nature of $Y(x)$, this probability can be written in closed form using $\Phi = F_{\mathcal{N}(0,1)}$ the cdf of the standard gaussian.

$$\mathrm{PI}(x) = \mathbb{P}\left[Y(x) < f_{\min}\right] \tag{3.26}$$

$$= \Phi\left(\frac{m_Y(x) - f_{\min}}{\sigma_Y(x)}\right) \tag{3.27}$$

This threshold can have different forms

- $f_{\min} = \min_i f(x_i)$, so the GP is to be compared with the current minimal value reached by the function
- $f_{\min} = \min_i f(x_i) + \epsilon$ so we introduce a small tolerance ϵ, in order to encourage exploration instead of intensification.

Using the probability of improvement tends to select points quite close to the point evaluated so far, thus does favor intensification at the expense of exploration.

3.3.2.b Expected improvement and EGO

One of the most c criteria, [Moč74][JSW98] Quite related to the probability of improvement, we define the improvement $I(x)$ as the random variable defined as

$$I(x) = [f_{\min} - Y(x)]_+ \tag{3.28}$$

where $[y]_+ = \max(y, 0)$. The *Expected Improvement* EI is

$$\mathrm{EI}(x) = \mathbb{E}[I(x)] = \mathbb{E}\left[[f_{\min} - Y(x)]_+\right] \tag{3.29}$$

Again, a closed form is available to compute the expected improvement, that does not require the evaluation of the expectation Eq. (3.29):

$$\mathrm{EI}(x) = (f_{\min} - m_Y(x))\,\Phi\left(\frac{f_{\min} - m_Y(x)}{\sigma_Y(x)}\right) + \sigma_Y(x)\phi\left(\frac{f_{\min} - m_Y(x)}{\sigma_Y(x)}\right) \tag{3.30}$$

3.3.2.c IAGO

Another criterion worth mentioning is a criterion based on the distribution of the minimizers [VVW06, HS11] Let y_i be a sample path of Y, and let x_i^* the global minimizer of y_i. We denote then X^* the random variable corresponding to the global minimizer of Y. We consider the differential entropy of X^* given the augmented design $\mathcal{X} \cup \{(x, Y(x))\}$. So at each step, we choose the point that gives the smallest expected uncertainty on the location of the global minimizers of the sample paths. The criterion can then be written as

$$\kappa_{\mathrm{IAGO}}(x \mid \mathcal{X}) = -\mathbb{E}_{Y(x)}\Big[H\left[X^* \mid \mathcal{X} \cup \{(x, Y(x))\}\right)\Big] \tag{3.31}$$

3.3.3 Contour and volume estimation

Let us start by introducing diverse tools based around Vorob'ev expectation of closed sets ([EA19, HST12, VV03]).

Let us consider A, a random closed set, such that its realizations are subsets of $\mathbb{X}$, and π_A is its coverage probability, that is

$$\pi_A(x) = \mathbb{P}\left[x \in A\right], x \in \mathbb{X} \tag{3.32}$$

For a given $x \in \mathbb{X}$, the event "x belongs to A" happens with probability $\pi_A(x)$, thus has variance $\pi_A(x)(1 - \pi_A(x))$.

For $\eta \in [0, 1]$, we define the η-level set of π_A, also called *Vorob'ev quantiles*(see [VV03])

$$Q_\eta = \{x \in \mathbb{X} \mid \pi_A(x) \geq \eta\} \tag{3.33}$$

Those sets are decreasing (with respect to the inclusion) when η increases:

$$0 \leq \eta \leq \xi \leq 1 \implies Q_\xi \subseteq Q_\eta \tag{3.34}$$

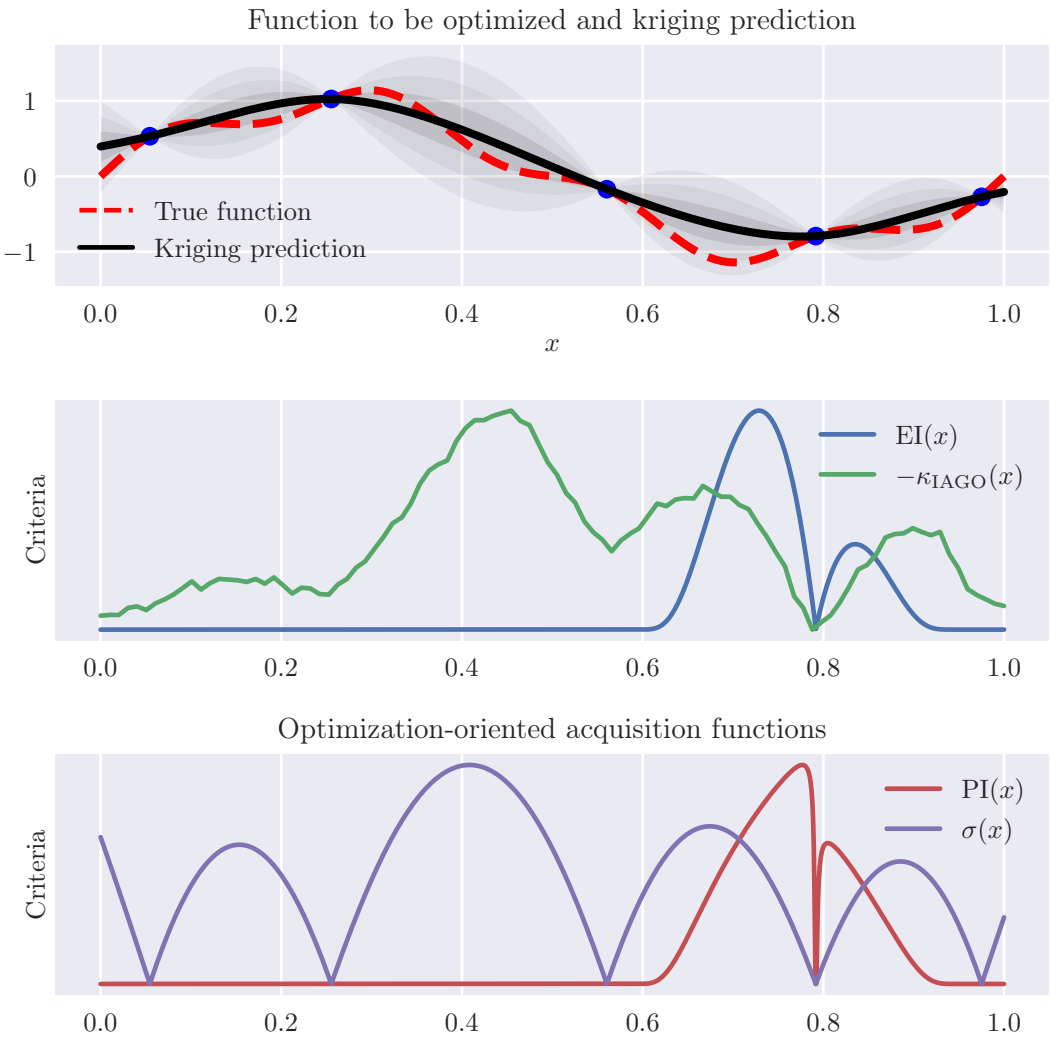

Figure 3.3: Example of optimization criteria. At this iteration, EI and PI aim toward intensification, while IAGO and the maximum of variance favorize exploration

Definition 3.3.1 – Vorob'ev expectation of random closed sets, Vorob'ev deviation: Let A a random closed set of $\mathbb{X}$, and μ a measure on $\mathbb{X}$. We define the Vorob'ev expectation, as the η^*-level set of A that verifies

$$\mathbb{E}[\mu(A)] = \mu(Q_{\eta^*}) \tag{3.35}$$

that is the level set of p, that has the volume of the mean of the volume of the random set A. If this equation does not have solutions, η^* is chosen as

$$\forall \beta < \eta^* \quad \mu(Q_\beta) \leq \mathbb{E}[\mu(A)] \leq \mu(Q_{\eta^*}) \tag{3.36}$$

Furthermore, as we have defined a kind of expectation of a random set, we can define a deviation as

$$\mathbb{E}[\mu\left(Q_{\eta^*} \triangle A\right)] \tag{3.37}$$

with $E \triangle F = (E \setminus F) \cup (F \setminus E)$ being the symmetric difference of the sets E and F

Now, let us assume that the random set A is

$$\kappa(x \mid \mathcal{X}) = \mathbb{E}_{Y(x)}\left[\mathbb{E}\left[\mu\left(Q_{\eta^*_{n+1}} \triangle A\right)\right]\right] \tag{3.38}$$

where $Q_{\eta^*_{n+1}}$ is the Vorob'ev expectation of the random set A, constructed using the GP Y with the augmented design $\mathcal{X} \cup \{(x, Y(x))\}$.

3.3.3.a Margin of uncertainty

Using the level sets, we can construct the η-margin of uncertainty, as introduced in [DSB11], that is the set of points $x \in \mathbb{X}$ that we cannot classify in or out of A with high enough probability. Setting the classical level $\eta = 0.05$ for instance, $Q_{1-\frac{\eta}{2}} = Q_{0.975}$ is the set of points whose probability of coverage is higher than 0.975, while $Q_{\frac{\eta}{2}} = Q_{0.025}$ is the set of points whose probability of coverage is higher than 0.025. Obviously, $Q_{1-\frac{\eta}{2}} \subset Q_{\frac{\eta}{2}}$. The complement of $Q_{\frac{\eta}{2}}$ in $\mathbb{X}$, denoted by $Q^C_{\frac{\eta}{2}}$ is the set of points whose probability of coverage is lower than 0.025. The η-margin of uncertainty $\mathbb{M}_\eta$ is defined as the sets of points whose coverage probability is between 0.025 and 0.975.

$$\mathbb{M}_\eta = \left(Q_{1-\frac{\eta}{2}} \cup Q^C_{\frac{\eta}{2}}\right)^C = Q^C_{1-\frac{\eta}{2}} \cap Q_{\frac{\eta}{2}} = Q_{\frac{\eta}{2}} \setminus Q_{1-\frac{\eta}{2}} \tag{3.39}$$

3.3.4 Robust criteria and GP

So far, we introduced strategies for either optimization or exploration to be applied on a generic space $\mathbb{X}$. From a robust point of view, we are going to consider the cost function J on a joint space $\mathbb{X} = \Theta \times \mathbb{U}$:

We assume that we constructed a GP Y on the joint space $\Theta \times \mathbb{U}$, based on a design of n evaluated points $\mathcal{X}_n = \{((\theta_i, u_i), J(\theta_i, u_i))\}_{1 \leq i \leq n}$, denoted as $(\theta, u) \mapsto Y(\theta, u)$.

As a GP, Y is described by its mean function m_Y and a covariance function $C(\cdot, \cdot)$, while $\sigma^2_Y(\theta, u) = C\left((\theta, u), (\theta, u)\right)$

$$Y(\theta, u) \sim \mathcal{N}\left(m_Y(\theta, u), \sigma^2_Y(\theta, u)\right) \tag{3.40}$$

A surrogate of J using Y is then m_Y.

One main challenge when making the distinction $\mathbb{X} = \Theta \times \mathbb{U}$ is that the objectives on the two spaces are not the same. Most of criteria need first to remove the dependence on U (by projecting the GP for instance in Section 3.3.4.a), and then to minimize a criterion with respect to θ, giving $\theta_{\text{candidate}}$. The next point to evaluate however has to be chosen in $\Theta \times \mathbb{U}$, thus needing to select a point (θ_{n+1}, u_{n+1}), where θ_{n+1} may not be equal to $\theta_{\text{candidate}}$.

Algorithm 2 SUR strategy with distinctive spaces

Require: Initial design $\mathcal{X}_0$, criterion function κ
  Fit Y, a GP using the design $\mathcal{X}_0$
  $n \leftarrow 0$
  while Stopping criterion not met or max budget evaluation not reached **do**
    $\tilde{\theta} \leftarrow \arg\max_{\theta\in\Theta} \kappa(\theta; Y \mid \mathcal{X}_n)$
    $(\theta_{n+1}, u_{n+1}) \leftarrow \arg\max_{(\theta,u)\in\Theta\times\mathbb{U}} \kappa((\theta,u); \tilde{\theta}, Y \mid \mathcal{X}_n)$
    Evaluate $J(\theta_{n+1}, u_{n+1})$
    $\mathcal{X}_{n+1} \leftarrow \mathcal{X}_n \cup \{((\theta_{n+1}, u_{n+1}), J(\theta_{n+1}, u_{n+1}))\}$
    $n \leftarrow n + 1$
  end while

3.3.4.a Expected loss

Recalling the definition of $\hat{\theta}_{\text{mean}} = \arg\min_{\theta\in\Theta} \mathbb{E}_U [J(\theta, U)]$, we can look to minimize the expected In [JLR10], the author define the *projected process* Z, a stochastic process $\theta \to (\Omega \to \mathbb{R})$ as

$$Z(\theta) = \mathbb{E}_U[Y(\theta, U)] = \int_{\mathbb{U}} Y(\theta, u) p_U(u)\, \mathrm{d}u \tag{3.41}$$

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} \text{EI}_Z(\theta) \tag{3.42}$$

$$(\theta_{n+1}, u_{n+1}) = \arg\min \mathbb{V}\text{ar} \tag{3.43}$$

3.3.4.b Profile expected improvement

In the previous chapter, we introduce the conditional minimum $J^*(U)$ and the conditional minimizers $\theta^*(U)$. To follow on that idea, the function $u \mapsto \theta^*(u)$ can be explored: [GBC+14] introduces the *Profile Expected Improvement* PEI, defined as

$$\text{PEI}(\theta, u) = \mathbb{E}\left[[f_{\min}(u) - Y(\theta, u)]_+\right] \text{ with } f_{\min} = \max(\min_i f(x_i), \min_{\theta\in\Theta} m_Y(\theta, u)) \tag{3.44}$$

This writing allow us to see the similarity with the EI criterion: instead of having a fixed threshold, the PEI introduces a criterion that depends on u. Figure 3.4 shows an example of GP enriched using the PEI criterion.

3.3.5 GP of the penalized cost function

We are now going to detail how Gaussian processes can help in recovering the regret-based families of robust estimators:

$$\{\hat{\theta}_{\text{add},\beta} = \max_{\theta\in\Theta} \Gamma_\beta(\theta) = \mathbb{P}_U\left[J(\theta, U) \le J^*(U) + \beta\right] \mid \beta \ge 0\} \tag{3.45}$$

$$\{\hat{\theta}_{\text{rel},\alpha} = \max_{\theta\in\Theta} \Gamma_\alpha(\theta) = \mathbb{P}_U\left[J(\theta, U) \le \alpha J^*(U)\right] \mid \alpha \ge 1\} \tag{3.46}$$

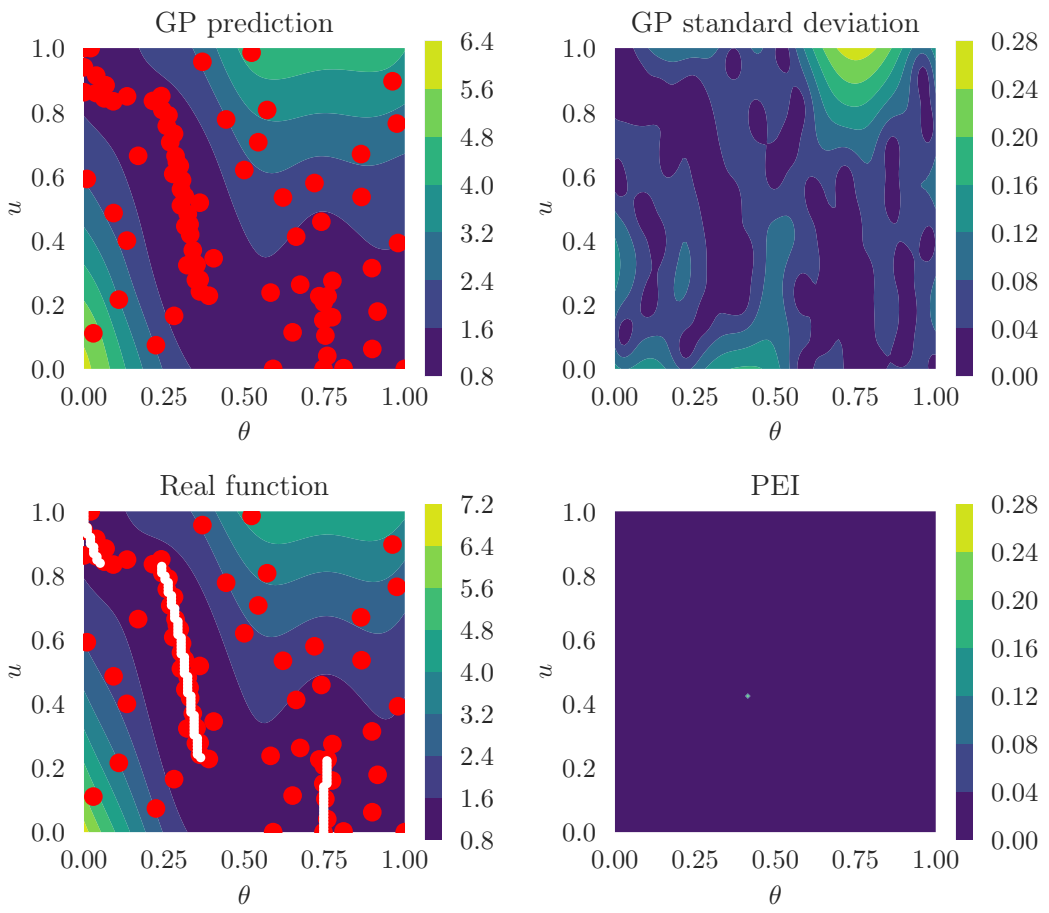

Figure 3.4: GP after 50 additional iterations chosen using PEI

Let us consider now the conditional minimiser:

$$J^*(u) = J(\theta^*(u), u) = \min_{\theta \in \Theta} J(\theta, u) \tag{3.47}$$

Analogous to J and J^*, we define Y^* as

$$Y^*(u) \sim \mathcal{N}\left(m_Y^*(u), \sigma_Y^{2,*}(u)\right) \tag{3.48}$$

where

$$m_Y^*(u) = \min_{\theta \in \Theta} m_Y(\theta, u) = m_Y(\theta^*(u)) \tag{3.49}$$

$$\sigma_Y^{2,*}(u) = \sigma_Y^{2,*}(\theta^*(u)) \tag{3.50}$$

The surrogate conditional minimiser is used in [GBC+14] for instance, but other choices could be considered, such as $m_Y(\theta^*(u)) - \gamma\sigma_Y^{2,*}(\theta^*(u))$. This choice would lead to be more "optimistic" in the estimation of the minimum (i.e. a lower minimum), and in turn,

would have a tendency to overestimate the estimated value of α, thus trending toward more conservative estimates.

The difference defined as $\Delta_\alpha = Y - \alpha Y^*$ is a linear combination of correlated Gaussian processes. Its distribution is thus Gaussian and can be derived by first considering the joint distribution of $Y(\theta, u)$ and $Y^*(u) = Y(\theta^*(u), u)$:

$$\begin{bmatrix} Y(\theta,u) \\ Y^*(u) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_Y(\theta,u) \\ m_Y^*(u) \end{bmatrix}; \begin{bmatrix} C\left((\theta,u),(\theta,u)\right) & C\left((\theta,u),(\theta^*(u),u)\right) \\ C\left((\theta,u),(\theta^*(u),u)\right) & C\left((\theta^*(u),u),(\theta^*(u),u)\right) \end{bmatrix} \right) \tag{3.51}$$

Multiplying by the matrix $\begin{bmatrix} 1 & -\alpha \end{bmatrix}$ yields

$$\Delta_\alpha(\theta,u) \sim \mathcal{N}\left(m_\Delta(\theta,u); \sigma_\Delta^2(\theta,u)\right) \tag{3.52}$$

$$m_\Delta(\theta,u) = m_Y(\theta,u) - \alpha m_Y^*(u) \tag{3.53}$$

$$\sigma_\Delta^2(\theta,u) = \sigma_Y^2(\theta,u) + \alpha^2 \sigma_{Y^*}^2(\theta,u) - 2\alpha C\left((\theta,u),(\theta^*(u),u)\right) \tag{3.54}$$

Decomposing the variance σ_Δ^2 in Eq. (3.54), 3 sources of uncertainty arise:

- σ_Y^2 is the prediction variance of the GP on J, that is directly reduced when additional points are evaluated
- $\sigma_{Y^*}^2$ is the variance of the predicted value of the minimizer.
- Assuming a stationary form of the covariance, the third term is directly dependent on the distance between θ and $\theta^*(u)$. As the covariance term can be written $C((\theta,u),(\theta',u')) = s \prod_{i\in\mathcal{I}_\theta} \rho_{\theta_i}(\|k_i - k_i'\|) \prod_{j\in\mathcal{I}_u} \rho_{u_j}(\|u_j - u_j'\|)$, substituting $\theta^*(u)$ for θ' gives

$$C\left((\theta,u),(\theta^*(u),u)\right) = s \prod_{i\in\mathcal{I}_\theta} \rho_{\theta_i}(\|k_i - k_i^*(u)\|) \prod_{j\in\mathcal{I}_u} \rho_{u_j}(0) \tag{3.55}$$

$$= s \prod_{i\in\mathcal{I}_\theta} \rho_{\theta_i}(\|k_i - k_i^*(u)\|) \tag{3.56}$$

This decomposition highlights the fact that the uncertainty measured at a point (θ, u) using σ_Δ^2 will not be reduced completely by evaluating the function at this point, as only the prediction variance σ_Y^2 will be significantly affected in general. In this case, reducing the uncertainty on a slice of constant θ (candidate) will not result necessarily in an evaluation located on this slice.

3.3.6 Evaluation and optimization of Γ

Let consider $\alpha \geq 1$ fixed. In order to compute $\hat{\theta}_{\text{rel},\alpha}$, we need to estimate and optimize the function Γ_α. For that purpose, we can first explore the space to improve the estimation of Γ_α, and once sufficient knowledge is acquired, use the plug-in estimate $\hat{\Gamma}_\alpha$ for the optimization, to get the wanted estimator.

3.3.6.a Improving the estimation of Γ_α

For a given $\theta \in \Theta$, the coverage probability of the α-acceptable region, i.e. the probability for θ to be α-acceptable is

$$\Gamma_\alpha(\theta) = \mathbb{P}_U \left[J(\theta, U) \leq \alpha J^*(U) \right] \tag{3.57}$$
$$= \mathbb{E}_U \left[\mathbb{1}_{J(\theta,U) \leq \alpha J^*(U)} \right] \tag{3.58}$$

As J is not known perfectly, it can be seen as a classification problem. This classification problem can be approached with a plug-in approach in Eq. (3.59), or a probabilistic one in Eq. (3.60):

$$\mathbb{1}_{J(\theta,u) \leq \alpha J^*(u)} \approx \mathbb{1}_{m_Y(\theta,u) \leq \alpha m_Y^*(u)} \tag{3.59}$$
$$\mathbb{1}_{J(\theta,u) \leq \alpha J^*(u)} \approx \mathcal{P} \left[\Delta_\alpha(\theta, u) \leq 0 \right] = \pi_\alpha(\theta, u) \tag{3.60}$$

Based on those two approximation, we can define two different estimations of Γ_α, namely $\hat{\Gamma}_\alpha^{\text{PI}}$ with the plug-in approach, and Γ_α^π for the probabilistic one.

- For $\hat{\Gamma}_\alpha^{\text{PI}}$, we are going to reduce the expected augmented IMSE of the GP $Y - \alpha Y^*$.
- For $\hat{\Gamma}_\alpha^\pi$, we are going to reduce the expected augmented integrated variance of probability of coverage.

3.3.6.b Plug-in approach

For the plug-in approach, the chosen estimator is defined Eq. (3.61):

$$\hat{\Gamma}_\alpha^{\text{PI}}(\theta) = \mathbb{P}_U \left[m_Y(\theta, u) \leq \alpha m_Y^*(u) \right] \tag{3.61}$$

The outer expectation operator is to be computed numerically, using quadrature rule, or Monte-carlo methods. In general, from a set of samples $\{u_i\}_{1 \leq i \leq n_u}$,

$$\Gamma_\alpha(\theta) \approx \frac{1}{n_u} \sum_{i=1}^{n_u} \mathbb{1}_{m_Y(\theta,u_i) - \alpha m^*(u_i) \leq 0} \tag{3.62}$$

Due to the fact that the GP surrogate is cheap to evaluate, the computation of the outer expectation with respect to U is assumed to be performed without too much problems.

In order to improve the accuracy of this estimator, one need to improve the GP prediction m_Y of the cost function J. In this case, we propose to reduce the IMSE of the GP $Y - \alpha Y^*$. The choice of the IMSE (instead of choosing the point of maximal variance for instance) comes from the decomposition of the variance Eq. (3.54). Let $\hat{\Gamma}_{\alpha,n}$ be the plug-in approximation of Γ_α, constructed using the Gaussian Process surrogate with n points added, according to the augmented IMSE. Figure 3.5 illustrates the L^2 and L^∞ between the truth Γ_α and the estimation $\hat{\Gamma}_{\alpha,n}$.

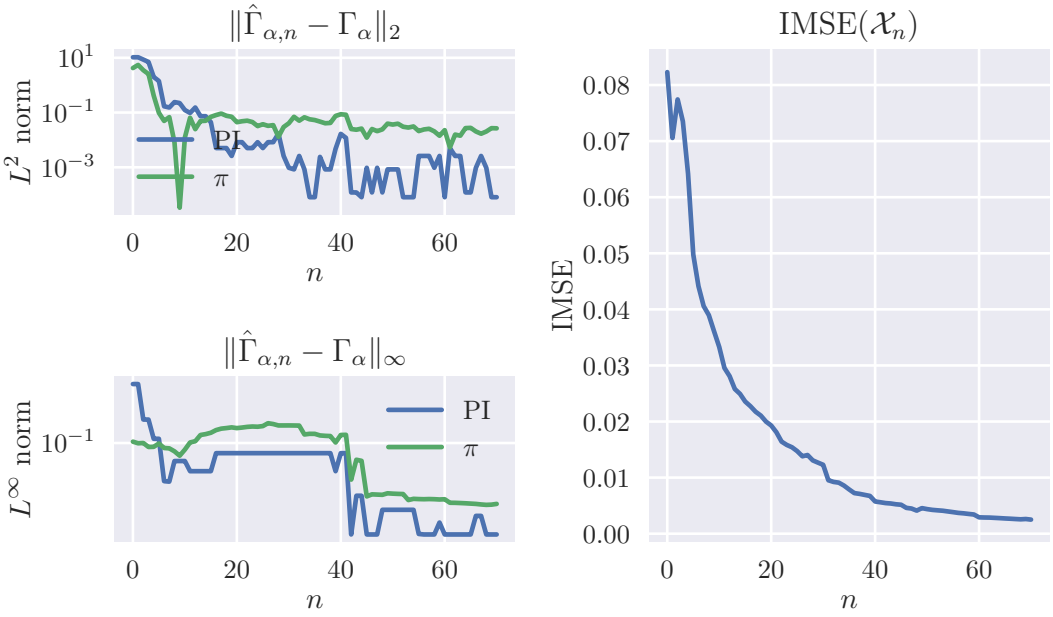

Figure 3.5: Evolution of the L^2 and L^∞ error of the estimation of Γ_α and IMSE of the successive constructed GP. The points added are chosen by the augmented expected IMSE

3.3.6.c Improving the estimation of the probability of coverage

Recalling the definition of the probabilistic approach Eq. (3.60), given the GP Y, we can write an estimator of Γ_α:

$$\hat{\Gamma}^\pi_\alpha(\theta) = \mathbb{E}_U\left[\mathbb{P}_Y\left[\Delta_\alpha(\theta, U) \leq 0\right]\right] = \mathbb{E}_U\left[\mathbb{P}_Y\left[Y(\theta, U) - \alpha Y^*(U) \leq 0\right]\right] \tag{3.63}$$

$$= \mathbb{E}_U\left[\pi_\alpha(\theta, u)\right] \tag{3.64}$$

The set $\{(\theta, u) \mid Y(\theta, u) - \alpha Y^*(u) \leq 0\}$ has a probability of coverage written π_α, which can be computed using the CDF of the standard normal distribution Φ, because Δ_α is a GP, as defined Eqs. (3.52) to (3.54):

$$\pi_\alpha(\theta, u) = \Phi\left(-\frac{m_{\Delta_\alpha}(\theta, u)}{\sigma_{\Delta_\alpha}(\theta, u)}\right) \tag{3.65}$$

Finally, averaging the coverage probability over u yields

$$\hat{\Gamma}^\pi_\alpha(\theta) = \mathbb{E}_U\left[\pi_\alpha(\theta, u)\right] = \int_{\mathbb{U}} \pi_\alpha(\theta, u) p_U(u)\,\mathrm{d}u = \int_{\mathbb{U}} \Phi\left(-\frac{m_{\Delta_\alpha}(\theta, u)}{\sigma_{\Delta_\alpha}(\theta, u)}\right) p_U(u)\,\mathrm{d}u \tag{3.66}$$

The variance of the probability of coverage is $\pi_\alpha(\theta, u)\left(1 - \pi_\alpha(\theta, u)\right)$. Integrating this variance over the whole space $\Theta \times \mathbb{U}$ gives the integrated variance of the probability of

coverage IVPC: [BGL+12]

$$\mathrm{IVPC}(\mathcal{X}) = \int_{\Theta\times\mathbb{U}} \pi_\alpha(\theta,u)\,(1-\pi_\alpha(\theta,u))\,p_U(u)\,\mathrm{d}\theta\,\mathrm{d}u \tag{3.67}$$

Instead of evaluating this integrated variance, on the current design $\mathcal{X}$, we can once again, augment the design at the (unevaluated) point (θ,u), assuming that its evaluation is the random variable $Y(\theta,u)$:

$$\mathrm{IVPC}(\mathcal{X}\cup\{((\theta,u),Y(\theta,u))\} \tag{3.68}$$

Finally, we can define a new learning function, which is the expected IVPC with respect to the random variable $Y(\theta,u)$:

$$\kappa(\theta,u) = \mathbb{E}_{Y(\theta,u)}\Big[\mathrm{IVPC}(\mathcal{X}\cup\{((\theta,u),Y(\theta,u))\}\Big] \tag{3.69}$$

Figure 3.6 shows the evolution of the error in the estimation of Γ_α, with respect to the L^2 and L^∞ norm.

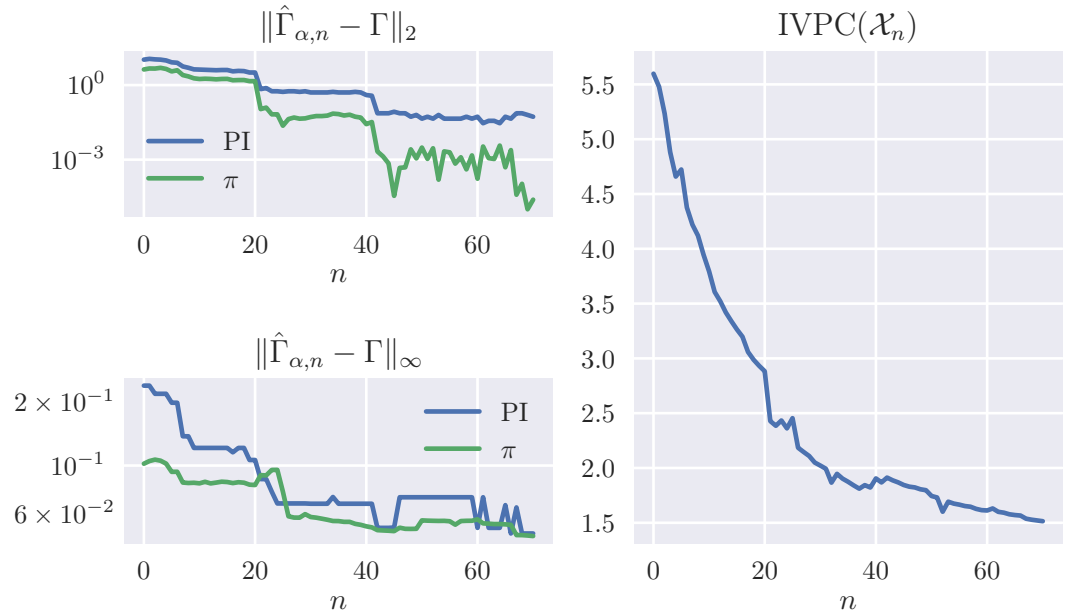

Figure 3.6: Enriching the design according to the criterion of Eq. (3.69)

3.3.7 Estimation of α_p based on GP

Instead of choosing a fixed threshold $\alpha \geq 0$, we can instead look for a threshold such that $\max\Gamma_\alpha$ is large enough. For a level p, we define α_p as the smallest threshold giving a maximal probability of acceptability above p.

$$\alpha_p = \inf_{\alpha\geq 1}\{\max_{\theta\in\Theta}\Gamma_\alpha(\theta)\geq p\} \tag{3.70}$$

As $J^*(u) \neq 0$ for all u, we can define $q_p(\theta)$ as the quantile of level p of the ratio $J(\theta, U)/J^*(U)$:

$$q_p(\theta) = Q_U\left(\frac{J(\theta,U)}{J^*(U)}; p\right) \tag{3.71}$$

$$\iff p = \mathbb{P}_U\left[\frac{J(\theta,U)}{J^*(U)} \leq q_p(\theta)\right] \tag{3.72}$$

α_p verifies then

$$\alpha_p = \min_\theta q_p(\theta) \tag{3.73}$$

and

$$p = \mathbb{P}_U\left[\max_{\theta\in\Theta}\frac{J(\theta,U)}{J^*(U)} \leq \alpha_p\right] \iff \alpha_p = Q_U\left(\max_{\theta\in\Theta}\frac{J(\theta,U)}{J^*(U)}; p\right) \tag{3.74}$$

Those two formulations Eq. (3.73), and Eq. (3.74) shows two ways of getting to α_p.

Plug-in approach

Again, the plug-in approach is to replace $J(\theta, u)/J^*(u)$ with $m_Y(\theta, u)/m_Y^*(u)$, and to compute the associated estimates.

$$q_p^{\mathrm{PI}}(\theta) = Q_U\left(\frac{m_Y(\theta,U)}{m_Y^*(U)}; p\right) \tag{3.75}$$

the estimation of the relaxation value $\hat{\alpha}_p$ is then the minimal value of the quantiles with respect to θ:

$$\hat{\alpha}_p^{\mathrm{PI}} = \min_{\theta\in\Theta} Q_U\left(\frac{m_Y(\theta,U)}{m_Y^*(U)}; p\right) \tag{3.76}$$

Monte-Carlo approach

Another approach relies on the sample paths of Y. Let y a sample path of Y. We can then compute $y(\theta, u)/y^*(u)$ and get an estimate based on this sample. Using the random nature of Y, we can compute measures of uncertainty on the estimation, and ultimately, reduce this uncertainty by choosing the next point to evaluate accordingly.

Before sampling trajectories however, one should first explore the whole space $\Theta \times \mathbb{U}$ sufficiently. Indeed, in order to get the prediction m_Y and the prediction of the conditional minimum m_Y^* of the GP should be strictly larger than 0. This condition applies to the trajectories too.

Let us say that we sampled N function from Y, namely $y^{(i)}$ for $1 \leq i \leq N$. For each of these samples, we can get $q_p^{(i)}(\theta)$. Using Monte-Carlo, we can get a Monte-Carlo estimation of q_p:.

$$\frac{1}{N}\sum_{i=1}^{N} q_p^{(i)}(\theta) = q_p^{\mathrm{MC}}(\theta) \tag{3.77}$$

and finally, minimizing the value of the estimated quantile leads to $\hat{\alpha}_p^{\mathrm{MC}}$, the MC approximation of α_p:

$$\hat{\alpha}_p^{\mathrm{MC}} = \min_{\theta \in \Theta} q_p^{\mathrm{MC}}(\theta) \tag{3.78}$$

3.3.7.a Enriching the design for the estimation

Reducing the augmented IMSE of the original GP

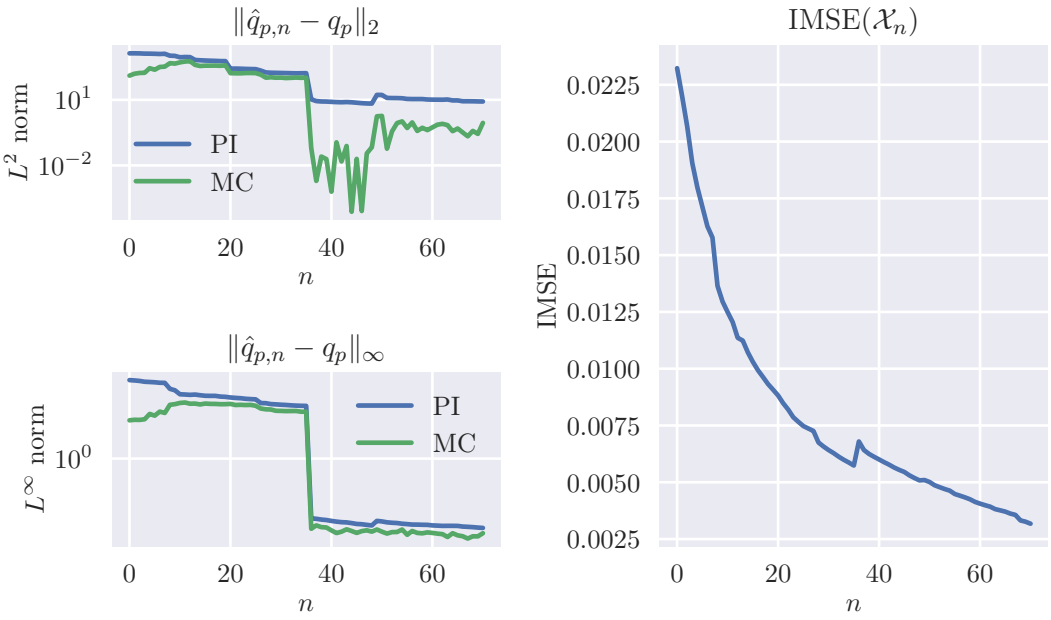

Figure 3.7: Enrichment according to the augmented IMSE of the gp Y

However, so far, we only tried to reduce the uncertainty on the whole space $\Theta \times \mathbb{U}$, in order to have a better approximation of q_p and Γ_α. We can adopt an approach more focused on their optimization. To do so, the principle stays the same: first, we derive a quantity of interest, that lead us to select a certain candidate $\tilde{\theta} \in \Theta$, then we look to reduce the uncertainty for this point:

Two-stages IVPC

We are first going to detail a two-stage approach to improve the maximization of Γ_α using the probabilistic approach based on the probability of coverage. We define first

$$\mathrm{IVPC}(\theta \mid \mathcal{X}) = \int_{\mathbb{U}} \pi_\alpha(\theta, u)\,(1 - \pi_\alpha(\theta, u))\, p_U(u)\,\mathrm{d}u \tag{3.79}$$

that is the IVPC when θ is fixed, or in other words, the uncertainty associated with the estimation of $\hat{\Gamma}_\alpha$ at a point $\theta \in \Theta$. We have also that $\mathrm{IVPC}(\mathcal{X}) = \int_\Theta \mathrm{IVPC}(\theta \mid \mathcal{X})\,\mathrm{d}\theta$ On

a computational note, IVPC($\theta \mid \mathcal{X}$) is easier to compute than the "full" IVPC defined Eq. (3.67), as the integration has to be performed on $\mathbb{U}$ only, instead of the joint space.

Once a candidate $\tilde{\theta}$ which maximizes a specified criterion has been obtained (Eq. (3.80)), we can attempt to reduce the uncertainty associated at this point, by finding the point that reduces at most the expected augmented IVPC (for instance) on $\{\tilde{\theta}\} \times \mathbb{U}$.

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \kappa_n(\theta) \tag{3.80}$$

$$(\theta_{n+1}, u_{n+1}) = \arg\min_{(\theta,u)\in\Theta\times\mathbb{U}} \mathbb{E}_{Y(\theta,u)} \left[\text{IVPC}\left(\tilde{\theta} \mid \mathcal{X}_n \cup \{((\theta,u), Y(\theta,u))\}\right)\right] \tag{3.81}$$

3.3.7.b Sampling based criterion

This technique is described in [DSB11] Let assume that we derived a criterion κ. And let $f(x) = \frac{\kappa(x)}{\int_{\mathbb{X}} \kappa(u)\,\mathrm{d}u}$. f can be seen as a density. Using an appropriate sampler, we can generate N iid samples from this criterion $\{x_i\}_{1\leq i\leq N}$

However, as N should be large, there is no point in evaluating all the samples x_i. This goes by the statistical reduction of the samples: This can be done by KMeans algorithm,

Objective name	Objective to minimize wrt θ	Computational solution
Profile Likelihood	$-\log \max_{u\in\mathbb{U}} p_{Y\mid\theta,U}(y \mid \theta, u)$	
Integrated Likelihood	$-\log \int_{\mathbb{U}} p_{Y\mid\theta,U}(y \mid \theta, u)\,\mathrm{d}u = -\log p_{Y\mid\theta}(y \mid \theta)$	
Marginal maximum a posteriori	$-\log p_{\theta\mid Y}(\theta \mid y)$	MCMC based sampling
Global Optimum	$\min_{u\in\mathbb{U}} J(\theta, u)$	EGO ([JSW98])
Worst-case	$\max_{u\in\mathbb{U}} J(\theta, u)$	
Regret worst-case	$\max_{u\in\mathbb{U}} \{J(\theta, u) - \min_{\theta'\in\Theta} J(\theta', u)\}$	
Mean	$\mathbb{E}_U[J(\theta, U)]$	Projected GP
Mean and variance	$\lambda\mathbb{E}_U[J(\theta, U)] + (1-\lambda)\sqrt{\mathbb{V}\text{ar}_U[J(\theta, U)]}$	Projected GP

Table 3.2: Summary of single objective robust estimators

CHAPTER 4

APPLICATION TO THE NUMERICAL COASTAL MODEL CROCO

Contents

4.1 The CROCO model

CROCO is a new oceanic modeling system built upon ROMS_AGRIF and the non-hydrostatic kernel of SNH (under testing), gradually including algorithms from MARS3D (sediments) and HYCOM (vertical coordinates). An important objective for CROCO is to resolve very fine scales (especially in the coastal area), and their interactions with larger scales. It is the oceanic component of a complex coupled system including various components, e.g., atmosphere, surface waves, marine sediments, biogeochemistry and ecosystems[1].

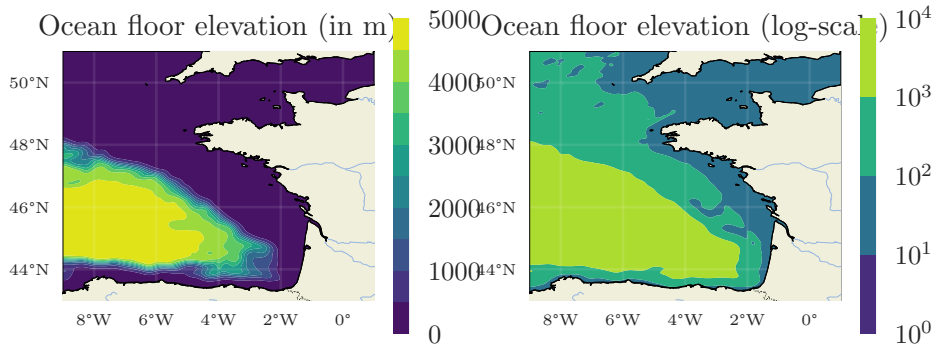

Figure 4.1: Map of the depth in CROCO

4.2 Deterministic calibration of the bottom friction

4.2.1 Physical parametrization of the bottom friction

$$(\tau_b^x, \tau_b^y) = C_d \sqrt{u_b^2 + v_b^2}(u_b, v_b) \tag{4.1}$$

$$C_d = \left(\frac{\kappa}{\log\left(\frac{z_b - H}{z_{0,b}}\right)} \right)^2 \text{ for } C_d \in [C_d^{\min}, C_d^{\max}] \tag{4.2}$$

4.2.2 Twin experiments setup

4.3 Modelling the uncertainties

[EE02] TPX model of tides

[1] taken from http://www.croco-ocean.org/

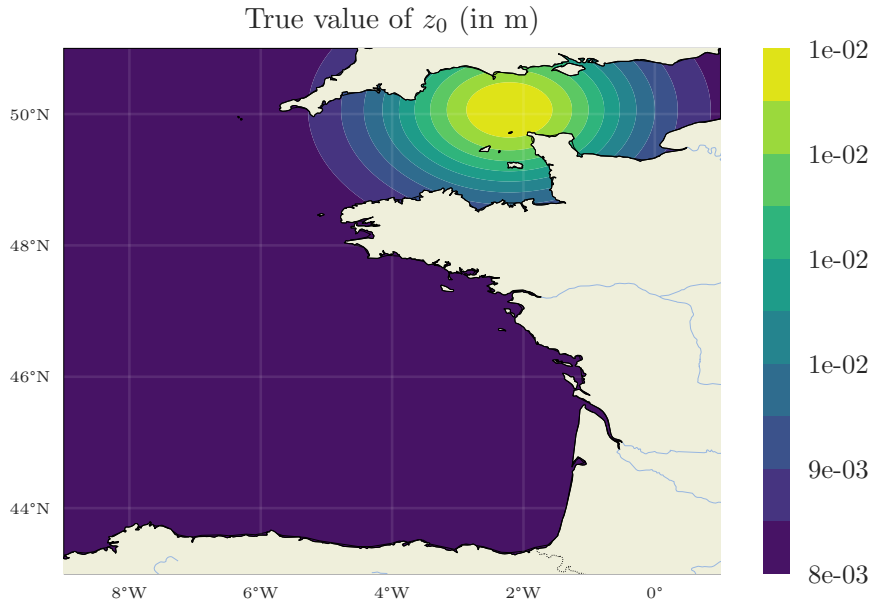

Figure 4.2: Distribution of the true value of the calibration parameter

4.4 Dimension Reduction

4.4.1 Ad-hoc segmentations methods

4.4.1.a Segmentation based on the depth

[Bou15]

4.4.1.b Geographical segmentation

4.5 Sensitivity Analysis

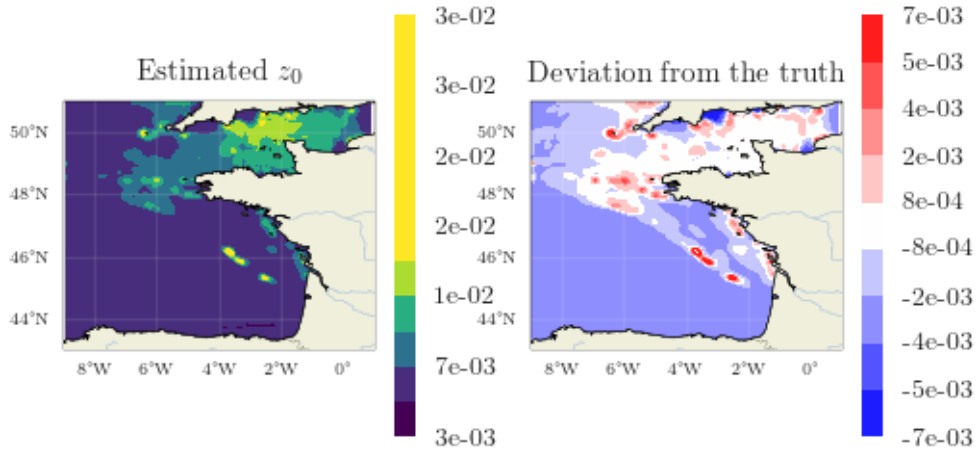

Figure 4.3: Optimization of z_0 on the whole space using gradient obtained via adjoint method, after 126 iterations

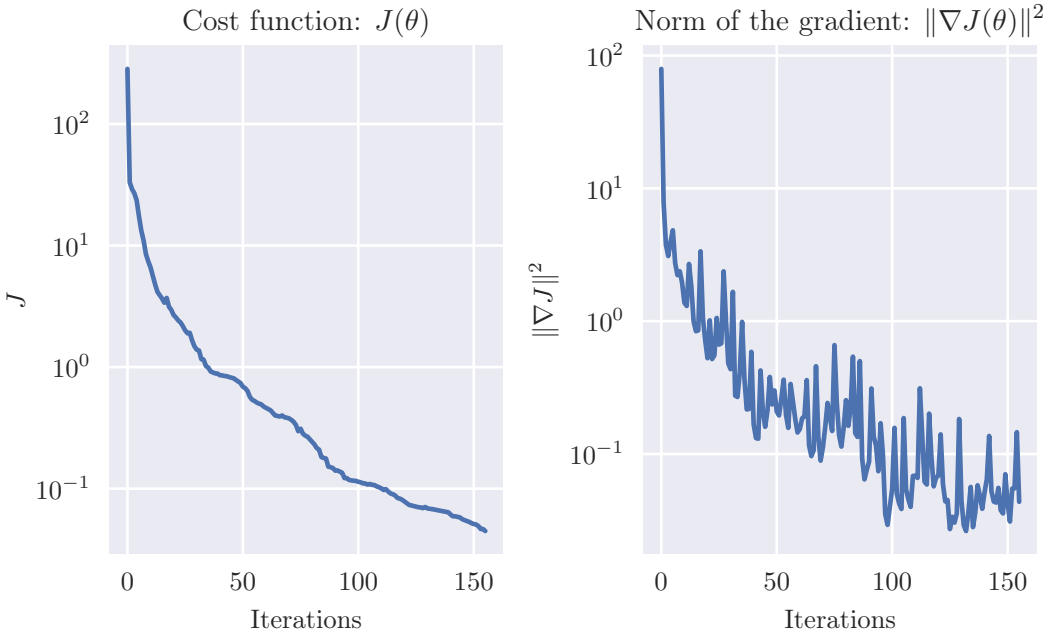

Figure 4.4: Gradient descent procedure

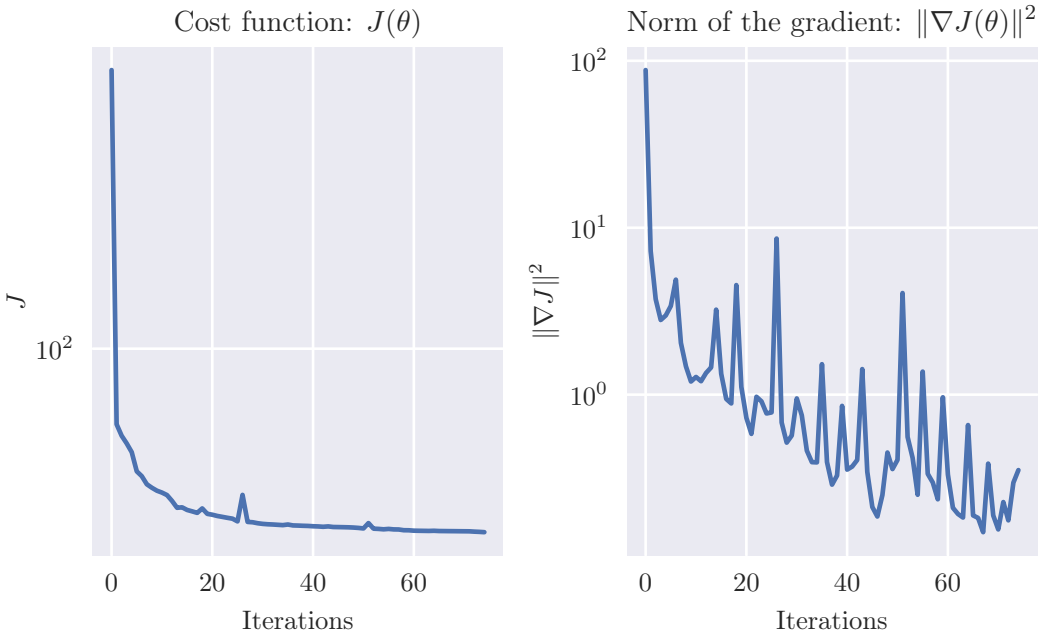

Figure 4.5: Gradient descent procedure in misspecified case

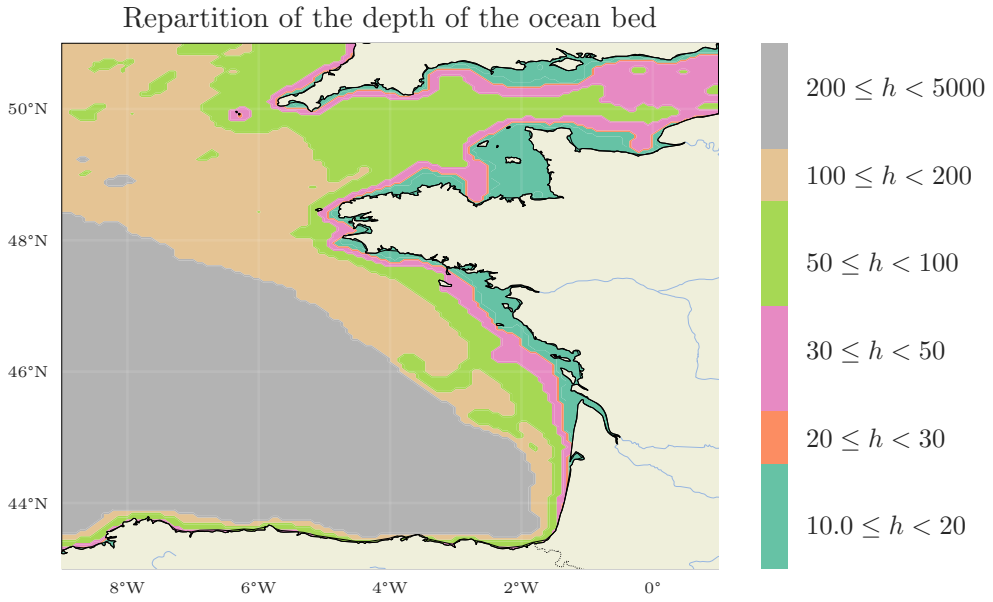

Figure 4.6

BIBLIOGRAPHY

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[ALMP+12] Vazken Andréassian, Nicolas Le Moine, Charles Perrin, Maria-Helena Ramos, Ludovic Oudin, Thibault Mathevet, Julien Lerat, and Lionel Berthet. All that glitters is not gold: The case of calibrating hydrological models: Invited Commentary. *Hydrological Processes*, 26(14):2206–2210, July 2012.

[BA04] Kenneth P. Burnham and David R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, November 2004.

[Bau12] Vincent Baudoui. *Optimisation Robuste Multiobjectifs Par Modèles de Substitution*. PhD thesis, Toulouse, ISAE, 2012.

[BDGV97] Jan Beirlant, Edward J. Dudewicz, László Györfi, and Edward C. Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[Bet17] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, January 2017.

[BGL+12] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, May 2012.

[BHRV17] Christophette Blanchet-Scalliet, Céline Helbert, Mélina Ribaud, and Céline Vial. A specific kriging kernel for dimensionality reduction: Isotropic by group kernel, March 2017.

[Bil08] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. springer, 2006.

[BKJ07] Walter Briec, Kristiaan Kerstens, and Octave Jokung. Mean-Variance-Skewness Portfolio Performance Gauging: A General Shortage Function and Dual Approach. *Management Science*, 53(1):135–149, January 2007.

[BLW99] James O. Berger, Brunero Liseo, and Robert L. Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, February 1999.

[Bou15] Martial Boutet. *Estimation Du Frottement Sur Le Fond Pour La Modélisation de La Marée Barotrope*. PhD thesis, Université d'Aix Marseille, 2015.

[BS07] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization – A comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, 196(33-34):3190–3218, July 2007.

[CJW17] Laurence W. Cook, Jerome P. Jarrett, and Karen E. Willcox. Extending Horsetail Matching for Optimization Under Probabilistic, Interval, and Mixed Uncertainties. *AIAA Journal*, 56(2):849–861, October 2017.

[CMMV13] Frédéric Couderc, Ronan Madec, Jérôme Monnier, and Jean-Paul Vila. *Dassfow-Shallow, Variational Data Assimilation for Shallow-Water Models: Numerical Schemes, User and Developer Guides*. PhD thesis, University of Toulouse, CNRS, IMT, INSA, ANR, 2013.

[Coo18] Laurence William Cook. *Effective Formulations of Optimization Under Uncertainty for Aerospace Design*. Thesis, University of Cambridge, July 2018.

[DGR02] Arnaud Doucet, Simon J. Godsill, and Christian P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1):77–84, January 2002.

[DL91] S. K. Das and R. W. Lardner. On the estimation of parameters of hydraulic models by assimilation of periodic tidal data. *Journal of Geophysical Research*, 96(C8):15187, 1991.

[DL92] S. K. Das and R. W. Lardner. Variational parameter estimation for a two-dimensional numerical tidal model. *International Journal for Numerical Methods in Fluids*, 15(3):313–327, August 1992.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[DSB11] V. Dubourg, B. Sudret, and J.-M. Bourinet. Reliability-based design optimization using kriging surrogates and subset simulation. *Structural and Multidisciplinary Optimization*, 44(5):673–690, November 2011.

[EA19] Mohamed El Amri. *Analyse d'incertitudes et de Robustesse Pour Les Modèles à Entrées et Sorties Fonctionnelles*. Thesis, Grenoble Alpes, April 2019.

[EE02] Gary D. Egbert and Svetlana Y. Erofeeva. Efficient Inverse Modeling of Barotropic Ocean Tides. *Journal of Atmospheric and Oceanic Technology*, 19(2):183–204, February 2002.

[FW11] Nial Friel and Jason Wyse. Estimating the evidence – a review. *arXiv:1111.1957 [stat]*, November 2011.

[GBC+14] David Ginsbourger, Jean Baccou, Clément Chevalier, Frédéric Perales, Nicolas Garland, and Yann Monerie. Bayesian Adaptive Reconstruction of Profile Optima and Optimizers. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):490–510, January 2014.

[GDB+09] David Ginsbourger, Delphine Dupuy, Anca Badea, Laurent Carraro, and Olivier Roustant. A note on the choice and the estimation of Kriging models for the analysis of deterministic computer experiments. *Applied Stochastic Models in Business and Industry*, 25(2):115–131, March 2009.

[Han01] K. Hanson. Markov Chain Monte Carlo posterior sampling with the Hamiltonian Method. Technical Report LA-UR-01-1016, Los Alamos National Lab., NM (US), February 2001.

[HB01] Luc Huyse and Dennis M. Bushnell. Free-form airfoil shape optimization under uncertainty using maximum expected value and second-order second-moment strategies. 2001.

[HKC+04] Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[HMR+10] Marc Honnorat, Jérôme Monnier, Nicolas Rivière, Étienne Huot, and François-Xavier Le Dimet. Identification of equivalent topography in an open channel flow using Lagrangian data assimilation. *Computing and Visualization in Science*, 13(3):111–119, March 2010.

[HP13] Laurent Hascoet and Valérie Pascual. The Tapenade automatic differentiation tool: Principles, model, and specification. *ACM Transactions on Mathematical Software*, 39(3):1–43, April 2013.

[HS11] Philipp Hennig and Christian J. Schuler. Entropy Search for Information-Efficient Global Optimization. December 2011.

[HST12] Philippe Heinrich, Radu S. Stoica, and Viet Chi Tran. Level sets estimation and Vorob'ev expectation of random compact sets. *Spatial Statistics*, 2:47–61, December 2012.

[Hub11] Peter J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[JLR10] Janis Janusevskis and Rodolphe Le Riche. Simultaneous kriging-based sampling for optimization and uncertainty propagation. Technical report, July 2010.

[JNLS09] Anatoli Juditsky, Arkadii S. Nemirovski, Guanghui Lan, and Alexander Shapiro. Stochastic Approximation Approach to Stochastic Programming. In *ISMP 2009 - 20th International Symposium of Mathematical Programming*, August 2009.

[JSW98] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[Kal85] J. G. Kalbfleisch. *Probability and Statistical Inference*. Springer Texts in Statistics. Springer New York, New York, NY, 1985.

[KD09] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, March 2009.

[Ken82] John T. Kent. Robust properties of likelihood ratio tests. *Biometrika*, 69(1):19–27, 1982.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[KO01] Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, January 2001.

[KR95] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[KRTK10] George Kuczera, Benjamin Renard, Mark Thyer, and Dmitri Kavetski. There are no hydrological monsters, just models and observations with large uncertainties! *Hydrological Sciences Journal*, 55(6):980–991, August 2010.

[LBM+16] Nicolas Lelièvre, Pierre Beaurepaire, Cécile Mattrand, Nicolas Gayton, and Abdelkader Otsmane. On the consideration of uncertainty in design: Optimization-reliability-robustness. *Structural and Multidisciplinary Optimization*, 54(6):1423–1437, 2016.

[LC06] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

[Li09] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Science & Business Media, 2009.

[LSN04] Jeffrey S. Lehman, Thomas J. Santner, and William I. Notz. Designing computer experiments to determine robust control variables. *Statistica Sinica*, pages 571–590, 2004.

[LYW06] Kin Keung Lai, Lean Yu, and Shouyang Wang. Mean-Variance-Skewness-Kurtosis-based Portfolio Optimization. In *First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06)*, volume 2, pages 292–297, June 2006.

[MA10] R. Timothy Marler and Jasbir S. Arora. The weighted sum method for multi-objective optimization: New insights. *Structural and Multidisciplinary Optimization*, 41(6):853–862, June 2010.

[Moč74] J. Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, Lecture Notes in Computer Science, pages 400–404. Springer, Berlin, Heidelberg, July 1974.

[OR97] Włodzimierz Ogryczak and Andrzej Ruszczynski. On stochastic dominance and mean-semideviation models. 1997.

[RBGH19] Mélina Ribaud, Christophette Blanchet-Scalliet, Frederic Gillot, and Céline Helbert. Robustness kriging-based optimization. February 2019.

[Rei13] Nancy Reid. Aspects of likelihood inference. *Bernoulli*, 19(4):1404–1418, September 2013.

[Rib18] Mélina Ribaud. *Krigeage Pour La Conception de Turbomachines : Grande Dimension et Optimisation Multi-Objectif Robuste*. Thesis, Lyon, October 2018.

[RSNN15] Vishwas Rao, Adrian Sandu, Michael Ng, and Elias Nino-Ruiz. Robust data assimilation using L_1 and Huber norms. *SciRate*, November 2015.

[RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006.

[Sav51] L. J. Savage. The Theory of Statistical Decision. *Journal of the American Statistical Association*, 46(253):55–67, March 1951.

[Sch78] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, March 1978.

[SCIP14] Pranay Seshadri, Paul Constantine, Gianluca Iaccarino, and Geoffrey Parks. A density-matching approach for optimization under uncertainty. *arXiv:1409.7089 [math, stat]*, September 2014.

[Sco79] David W. Scott. On Optimal and Data-Based Histograms. *Biometrika*, 66(3):605, December 1979.

[SSW89] Jerome Sacks, Susannah B. Schiller, and William J. Welch. Designs for Computer Experiments. *Technometrics*, 31(1):41–47, February 1989.

[SWG14] Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. *arXiv:1402.4306 [cs, stat]*, February 2014.

[TA77] Andrei Tikhonov and Vasily Arsenin. *Solutions of Ill-Posed Problems*, volume 14. 1977.

[Tar05] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pa, 2005.

[TAVD20] Victor Trappler, Élise Arnaud, Arthur Vidard, and Laurent Debreu. Robust calibration of numerical models based on relative regret (Preprint), February 2020.

[Tib11] Robert Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[Vap92] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

[VV03] Oleg Vorobyev and Alexey Vorobyev. On the New Notion of the Set-Expectation for a Random Set of Events. *University Library of Munich, Germany, MPRA Paper*, January 2003.

[VVW06] Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *arXiv:cs/0611143*, November 2006.

[Wal45] Abraham Wald. Statistical Decision Functions Which Minimize the Maximum Risk. *Annals of Mathematics*, 46(2):265–280, 1945.

[Whi82] Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, 1982.

[WHR+03] Warren E. Walker, Poul Harremoës, Jan Rotmans, Jeroen P. van der Sluijs, Marjolein BA van Asselt, Peter Janssen, and Martin P. Krayer von Krauss. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17, 2003.

[Yiz95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, Aug./1995.

ABSTRACT ENGLISH

* * *

RESUMÉ FRANÇAIS